Approach to identify product and process state drivers in manufacturing systems using supervised machine learning

Dem Fachbereich Produktionstechnik

der

UNIVERSITÄT BREMEN

zur Erlangung des Grades eines

Doktors der Ingenieurwissenschaften

– Dr.-Ing. –

genehmigte

DISSERTATION

von

Dipl.-Wi.-Ing. Thorsten Wuest

Hauptreferent: Prof. Dr.-Ing. habil. Klaus-Dieter Thoben

Korreferent: Prof. Christopher Irgens (University of Strathclyde, UK)

Tag der mündlichen Prüfung: 24.11.2014

This dissertation is published in adapted form in the series:

Springer Theses

Recognizing Outstanding Ph.D. Research

http://www.springer.com/series/8790

Cite as follows:

Wuest, T. (2015). *Approach to identify product and process state drivers in manufacturing systems using supervised machine learning*. Springer Theses. Heidelberg, Berlin: Springer.

To my wife Irene and my parents Antonie and Peter.

Abstract

The global manufacturing domain faces major challenges which may be summarized by increasing complexity and dynamics of products and processes as well as increasing requirements towards quality. The research problem of this thesis is set in multi-stage manufacturing programmes and focuses on the holistic handling of information with the goal of improving product and process quality. Existing solutions focus mostly on individual processes instead of the whole manufacturing system and do not incorporate product and process inter- and intra-relations. It was found that these process inter- and intra-relations have a significant and varying impact on the quality outcome of successive processes and thus on the whole manufacturing programme.

In the dissertation, a concept has been developed to describe comprehensively a product by its states along a manufacturing process sequence. For this concept, it is of fundamental importance to identify a set of state characteristics allowing a comprehensive description of the product's state. A major aspect within the work was found to be process intra- and inter-relations between states and state characteristics. Today, most manufacturing programmes lack sufficient knowledge and transparency with regard to process intra- and inter- relations rendering a complete modelling of the system unrealistic. In order to incorporate this crucial element in the analysis, supervised machine learning was employed in the form of SVM based feature ranking to incorporate successfully implicit process intra- and inter-relations of the manufacturing programme.

The evaluation of the research was conducted by using three different scenarios from distinctive manufacturing domains (aviation, chemical and semiconductor) based on 'real world' data sets. The purpose of choosing three different scenarios was to highlight the general applicability of the developed concept. The evaluation confirmed that it is possible to incorporate implicit process intra- and inter-relations on process as well as programme level as required through applying SVM based feature ranking.

The developed concept allows identifying relevant state drivers of complex manufacturing systems holistically. It is able to utilize complex, diverse and high-dimensional data sets which often occur in manufacturing applications. It can be safely said that in the near future, the amount of data derived from manufacturing operations will increase due to these developments. This offers both opportunities and challenges for manufacturing companies and manufacturing research. With the developed concept, the increasing data streams can be analyzed efficiently and applicable results can be derived. The analysis results present a direct benefit in form of the most important process parameters and state characteristics, the state drivers, of the manufacturing system. These can be directly utilized in, e.g., quality monitoring and advanced process control.

Abstract in German

Produzierende Unternehmen finden sich in einem dynamischen Wettbewerbsumfeld wieder, das von hohen Qualitätsanforderungen und kontinuierlich steigender Komplexität von Produkten und Prozessen gekennzeichnet ist. Um den kontinuierlich steigenden Qualitätsanforderungen begegnen zu können, reicht es nicht mehr aus, sich auf einzelne, isolierte Prozesse zu fokussieren. Durch die vielschichtigen Beziehungen zwischen Produkten und Prozessen rückt das komplette Produktionssystem in den Fokus. In diesem Zusammenhang ist es vorteilhaft, alle relevanten Informationen des Produktionssystems transparent und zielgerichtet zu organisieren, zu analysieren und den richtigen Stellen verfügbar zu machen.

In dieser Dissertation wird ein Konzept zum effizienten Management aller relevanten Produkt- und Prozessinformationen entlang einer Fertigungsprozesskette entwickelt und evaluiert. Ein spezieller Fokus liegt dabei auf den bisher vernachlässigten, prozessübergreifenden Wirkzusammenhängen zwischen Produktzuständen und Prozessparametern. Die Integration der Wirkzusammenhänge speziell unter Wissensdisparität stellt dabei vor dem Hintergrund der hohen Dynamik und Komplexität moderner Produktionssysteme eine große, bisher ungelöste Herausforderung dar.

Um das Ziel unter den vorherrschenden Rahmenbedingungen, u.a. hohe Komplexität, Dynamik und Wissensdisparität, zu erreichen, werden Techniken des maschinellen Lernens eingesetzt. Eine Bewertung der verschiedenen Parameter eines dynamischen und komplexen Produktionssystems mit Hilfe eines ‘Feature Rankings’ erlaubt die Identifikation der relevanten Zustandseigenschaften, sogenannter ‘State Drivers’. Die Analyse wird dabei nicht nur für individuelle Prozesse durchgeführt, sondern auch für verschiedene Sequenzvarianten bis hin zum kompletten Produktionssystem. Dadurch werden zusätzlich Wirkzusammenhänge zwischen einzelnen Prozessen berücksichtigt, die bisher unbekannt sind bzw. oft vernachlässigt werden.

Die Evaluation des Konzepts wird anhand von drei Szenarien aus verschiedenen Produktionsbereichen (Luftfahrt-, Chemische- und MEMS-industrie) durchgeführt. Die Ergebnisse der Evaluation zeigen, dass eine holistische, produktionssystemübergreifende Identifikation von relevanten Informationen unter Berücksichtig von impliziten Wirkzusammenhängen möglich ist. Dies ermöglicht nicht nur direkte Qualitätsverbesserungsmaßnahmen und eine bessere Modellierung des Systems, sondern stellt auch einen wichtigen Ausgangspunkt für weitergehende und tiefer greifende Analysen zur nachhaltigen Wissensgenerierung dar.

Acknowledgment

Many people have contributed to my success in completing this dissertation. I was fortunate to have been supported by my advisors, colleagues, friends and family.

First of all, I want to thank Prof. Klaus-Dieter Thoben, my doctoral advisor, for his continuous support over the last years and his genuine interest in my work and development. Without the countless critical discussions and his outstanding mentorship I would not be who I am today. I cannot thank Prof. Christopher Irgens, my second advisor, enough (even though he will insist that it is not necessary). He not only introduced me to a new research field, but more importantly, showed me that looking 'beyond one's own nose' is exciting and stimulating. Together, these two mentors did spark my academic curiosity and encouraged me to develop my own research interest. Furthermore, I want to thank Prof. Dr.-Ing. Hans-Werner Zoch, Dr.-Ing. Jan Ohlendorf, Stefan Wellsandt and Jakub Mak-Dadanski for their willingness to serve in my thesis committee and their constructive feedback.

I would like to thank Prof. Stephen Lu, my CIRP mentor, who invited me to his Lab at USC. He and his team, especially Dr. Ang Liu, made me feel at home from the beginning and opened my world to many fascinating new impressions. In this regard, I would like to thank the DAAD for the support of my research through the generous doctoral scholarship that allowed me to continue my research in the USA.

I want to thank the many inspiring researchers I met over the years and who always took the time to discuss ideas and provide critical feedback. There are too many to mention them all so instead I would like to highlight the great spirit of research communities and events like APMS, IFIP, CIRP. Of course I want to thank my colleagues at BIBA for their continuous support, professionally and privately. They were never too busy for a challenging discussion or a reassuring pep talk. Furthermore I want to thank the many students, whose projects and Bachelor, Master and Diploma theses I had the opportunity to supervise. They all contributed with their work towards the completion of this research. My friends and fellow ruggers, thank you for always being there for me and distracting me from continuously thinking of my research. This allowed the most creative ideas to surface.

Most importantly, I would like to thank my beloved wife Irene. Her support, encouragement, quiet patience and unwavering love despite the hardships of a long distance relationship are the foundation this work was built upon. Many thanks also to her parents and brother, Tere, Gabriel and Javier, who selflessly took me in during my research stay in LA and welcomed me as a son and brother.

Finally, I thank my loving parents, Antonie and Peter, who always encouraged me and taught me not to give up. Since I can remember, I felt safe knowing whatever happens they will always be there for me. Of course, I could always count on my brother Dennis, for needed distraction as well as professional input. Thank you all!

Table of content

1 Introduction

"Manufacturing creates wealth"[1]

The German economy evolved into an engine of growth within the European Union (Thesing, Randow, Kirchfeld, Berberich & Webb, 2010). A large part of this success is built on the industrial base, especially the showcase sector of manufacturing. German engineering products are exported worldwide and have the reputation of advanced technology and premium quality (N.N., 2006). In other parts of the world, even in the United States the manufacturing industry is growing again (Puzzanghera, 2013). The reputation of high quality products is a key factor of success in the fierce global competition (Enderwick, 2005; Levitt, 1993). The backbone of the German engineering success relies on its manufacturing companies, often being Small and Medium sized Enterprises (SME) (Schiersch, 2009). However, being successful does not necessarily mean that there is no potential for further improvements (Schiersch, 2009). At the same time, companies are constantly challenged to improve to meet the steadily increasing customer's requirements towards the quality of products and services (Kovačič & Šarler, 2009) in order to survive in the competitive global business environment (Porter, 2008). This stands especially true for manufacturing companies (Ellram & Krause, 1994). Product quality in this research is understood as the degree of fulfilment of the (quality) requirements by the characteristics of the final product (see section 2.1.4.1) (Yul & Wang, 2009).

Many companies focus on their core competencies and work together in collaborations and production networks to satisfy the increasing customer requirements and gain sustained competitive advantage (Hamel & Prahalad, 1990; Porter, 1998; Porter, 2008; Thomas, Byard & Evans, 2012). All these developments lead to an increasing complexity that companies must deal with in order to remain competitive. Taking into consideration that business success of every company is based on the quality of its business processes (Linß, 2002), it can be said that business success of a collaborative network is based on the quality of business processes of every collaboration partner. Looking at industrial companies, manufacturing processes play an important role through the direct value adding to products and the determination of product quality.

[1] Prof. Ronald G. Askin, Arizona State University (USA) during the INCOM 2012 Keynote speech Wednesday, May 23, 2012 in Bucharest, Romania.

1.1 Motivation[2]

The purpose of every process step during the production process is to add value to the product and therefore, at least in the manufacturing industry, change the products state (Kumar, 2002; Kalpakjian & Schmid, 2009). This change of state typically happens in a progression of successive states until its final state. The final state is generally defined by the requirements towards the product from the customers' side and specific to the individual manufacturing programme. Once the final state is reached, the product is considered ready for delivery to customers. In the end, the quality of a product is directly influenced by the quality of the production processes (Brinksmeier, 1991; Jacob & Petrick, 2007). Finally, the product state has to meet the customers' requirements in terms of product quality. But nowadays, this is not the only customer requirement the companies have to meet. Additionally, customers increasingly insist on transparent manufacturing processes and demand comprehensive information about a purchased product over the different stages of the manufacturing programme as part of the product lifecycle (Terzi, Panetto, Morel & Garetti, 2007; Cassina, Cannata & Taisch, 2009). This sometimes even includes parameter settings of the machines used and source of raw materials (e.g., iron ore). In Figure 1, an example of a manufacturing programme with three processes is illustrated. Ideally, the product increases its value in each process step and the next process gets the product (input) with the expected parameters (internal customer). Practically, the process can never reproduce a process step (output) 100 % due to different factors like: external influence (e.g., temperature, 'Mondays'); process variations (e.g., lubricant, tools); input deviations (e.g., different supplier of steel: even so the steel delivered has the same ISO Number, it can vary due to e.g., tolerance levels within the norm), etc. Therefore, it is also necessary to exchange information about the individual product. In this case, this means not just between stakeholders with a direct interface but along the whole manufacturing process sequence.

The trend of today's products becoming more and more optimized leads to a consistent exploitation of all achievable product properties (Denton, Gupta & Jawahir, 2003; Deja & Siemiatkowski, 2012). Considering the limited availability of resources (Specht & Braunisch, 2008), limited not only by their global presence but also political regulations (Maull, 1988; Frey, 2007; Lange, 2007), and the increasing demand e.g., by emerging economies, like the 'BRIC' countries (Buhr, Graf, Power & Amthauer, 2005; Specht & Braunisch, 2008), the aspect of efficient use of these resources grows in importance. If processes can be optimized in a way to, on the one hand, reduce scrap and rework and, on the other hand, to exploit all possible product properties as said before, the waste of valuable resources may be reduced and the customer requirements towards quality and information can still be

[2] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest & Thoben, 2012; Wuest, Knoke & Thoben, 2014b).

fulfilled. For example, if through an optimized process, e.g., final heat treatment, it is made possible to build a certain product from a widely available resource, e.g., steel, instead of a relatively rare resource, like e.g., titanium alloys, the rare resource of titanium will be preserved and can be put to use where it is absolutely necessary. For a company this practice may be beneficial regarding aspects of resource availability (e.g., widen the range of suppliers).

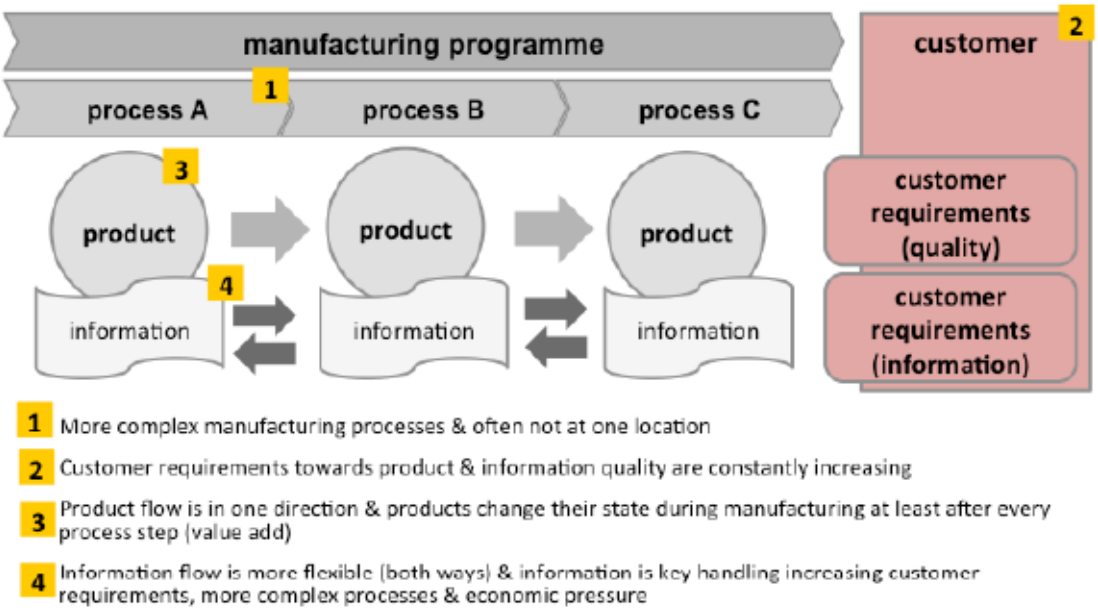

Figure 1: Connection of product, process and information towards customer requirements

To do so, it has to be confirmed that these desired product properties can be achieved through the production process (Mohanty, 2004). Therefore, a detailed understanding of the manufacturing programme, its processes/operations and their influence on products and in the end, their product state, becomes progressively more important. Continuous improvement of the manufacturing programme can be a way to stay ahead of competition. One important factor of the increasing overall complexity in the global business environment has on manufacturing engineering companies is the rapidly increasing need for information and information exchange. This includes many pitfalls, which among other things include: data acquisition, low data and information quality, data analysis, communication problems, security issues, interface problems of Information and Communication Technology (ICT) systems and, of course, the sheer amount of information to process (Choudhary, Harding & Tiwari, 2009).

Today's complex manufacturing programmes, processes and operations have the goal of adding value to the product. A manufacturing programme consists of a process chain with each process having a number of operations. In the chain, processes may be linked one-to-one, disjunctively or conjunctively to preceding or subsequent processes. Each process may have a chain of operations similarly linked within the process (see section 2.1.1 and Figure 6). Adding value to the product

can be done in various ways, e.g., by changing the physical form, hardness or extending the usability by adding services. Each manufacturing processes or operation, their planning and design, depends on information pertaining to the input state of a product. It is very important to distinguish between the planed input state and the real input state. This distinction is explained further in the following paragraphs. Basically, there are two ways how this can be achieved.

On the one side, the information can be provided from the design phase of the product. In this case, the information used is based on the state that the product is expected to inherit at a certain stage, e.g., the final state. This means everything, e.g., every process is going exactly according to plan. This is however unlikely in an industrial environment due to various factors, e.g., wear and tear of tools or quality deviations of raw materials. In this scenario, deviations of the different products will not be taken into account, as the possibility is not included in the system. A slight variation of this way is to include comprehensive quality control of the input states so just products passing the test, as they comply with the assumed "planned state", will be allowed to go through the process step (Garvin, 1984). The downside is that this is expensive (through e.g., extra staff, measurement technology, etc.) and time-consuming.

On the other side, the input information can be generated based on the individual product state at the time of the beginning of the processes or operation. Today, many companies already store operation data from factories and product property data from inspection (Kano & Nakagawa, 2008). The information here is more likely to be accurate as it takes into account the variations that can influence the process quality. Process quality is understood as the ability of the process to comply with certain criteria and to achieve the desired output (see section 2.1.4.2) (Kreutzberg, 2000). In order to provide each step with the individually necessary information, it is essential to be able to describe a product effectively. This can be done in various ways. Most likely the manner of describing an industrial product, e.g., gear made of steel, will be different from the description-style of a product designed to fulfill an aesthetic purpose (in addition to the functional purpose) in mind, e.g., a plastic rear mirror. At the same time, the individual describing a product influences the description based on, among other things, his or her own background, knowledge and experience. Therefore, the approach of describing a product through its product state (Kumar, 2002; Wuest, Klein & Thoben, 2011b) will help to align the descriptions in a commonly understood manner as well as increase transferability and usability of accompanying information by the addressees.

1.2 Problem statement

Along a manufacturing programme, physical products as well as information are exchanged between the partners (Hicks, Culley & McMahon, 2006) (see Figure 1). The availability of information is a precondition to adjust each manufacturing pro-

cess/operation in such a way that the outcome reflects the set quality requirements to a satisfactory degree. Quality, as stated before, constantly gains importance for the customer and for a sustainable use of resources. Sustainable in this sense is understood as not just focused on the environmental and social perspective but mainly focusing on the economical perspective. In addition, distributed production brings forth new challenges for managing quality (Sitek, Seifert & Thoben, 2010).

One way to improve manufacturing processes is to look at the data and information involved and how this information is put to use (Hicks et al., 2006). As stated by Albino, Pontrandolfo & Scozzi (2002), the successful coordination of a manufacturing process is mostly based on a successful handling of information to support process management and other tasks involved. With today's advanced ICT it becomes possible to process, transfer and store large amounts of data and information for a reasonable price (Krcmar, 2005). But too much information can be a threat for improved process quality as it can e.g., distract from the main issues/causalities or lead to delayed or wrong conclusions about appropriate actions (Lang, 2007). Jansen-Vullers, van Drop & Beulens (2003) emphasize the importance of the availability of the right information for quality during manufacturing processes. Hence the question is: What is the right and relevant information in the case of distributed manufacturing process chains and high tech industrial products?

In manufacturing companies, many processes and operations operate automated. Every process needs information about the product at the beginning of treatment to adjust the machine parameters. In an automated manufacturing programme, the information of the product is often not individual for the specific product at a specific time but derived of planning and design as described before. To assure an optimized handling of the individual product at a specific process during the manufacturing programme, information of the current product state (input product state) is necessary. The more precise and the more complete the information relevant for adjusting the process parameters is available, the better the machines can be adjusted and the more the quality of the product will be enhanced.

The complex relationship of information and quality in manufacturing as described in the previous paragraphs represents an important area when it comes to the development of new approaches with the goal of contributing to quality improvements in manufacturing programmes. Next, the chosen approach of this research is detailed, highlighting how the relationship of information and quality in manufacturing is addressed and what more focused and detailed measures to improve the transparency are proposed.

Currently, this dissertation is focused on manufacturing companies producing complex, highly stressed products, e.g., gear wheels or turbine blades. This is reflected within the three evaluation scenario, especially scenario I which is supplied by Rolls-Royce (see section 6.2). Highly stressed products are understood as prod-

ucts which are exposed to higher force and power density than the average product and thus have different requirements, e.g., strength, hardness and wear resistance (Tönshoff & Denkena, 2013). Tönshoff & Denkena (2013) state "in addition to the strength and hardness the quality requirements for highly-stressed components have grown significantly at the same time". This does not however mean the concept cannot be extended or adjusted to other organizations with different product portfolios. The reasons for the current focus on this group of products is, that highly stressed products have high requirements towards product and process quality (Tönshoff & Denkena, 2013), often use expensive materials, and the manufacturing programme itself is rather complex. This situation represents to a large extent the environment for the research problem. These complex manufacturing programmes are characterized by high quality requirements, high-dimensional and multi-variate data, etc. which directly influenced the development of the *product state concept*. Furthermore, manufacturing companies producing highly stressed products will most likely be among the first who consider adopting new methods and concepts to address these issues as their customers expect premium quality and their competitive advantage depends on constant improvements, e.g., reducing scrap and rework as much as possible (e.g., Garvin, 1984; Zoch & Lübben 2011).

Looking at the product information from a holistic manufacturing and quality perspective, the previously introduced relevant information becomes more important. The set of relevant information contains all information that is in one way or another relevant for the manufacturing programme as a whole and occurring transformational activities. In theory, when such a set of relevant information is available for a manufacturing programme and individual product, all information necessary to achieve the desired process and subsequently product quality is available to the stakeholders. However, the question remains how such a set of relevant information may be obtained in theory and practice. In order to determine what information subsets have to be included, in depth knowledge not only of the product itself but of all stages of the manufacturing programme, transformational activities and environmental influences and their inter-relations is required. Even for relatively simple manufacturing programmes and products, the required knowledge is currently not completely available. Given the ongoing trend that manufacturing programmes (e.g., automation) and products (e.g., materials) are becoming more complex, the theoretically required knowledge allowing to identify a relevant set of product information is increasing as well. This lack of case specific knowledge represents a major challenge for an information system and highlights the need for innovative approaches to identify relevant information in a manufacturing system in a comprehensive way even though total transparency cannot be achieved and not all necessary knowledge items are available.

In summary, the problem statement of this dissertation reads as follows: In an increasingly complex manufacturing environment with high quality requirements and an enlarged focus on information, there is a need for a holistic concept that in-

corporates the relevant information describing individual artifacts (products) comprehensively along the whole manufacturing programme and organizes them logically. Such a holistic concept has to incorporate recent developments from various manufacturing related domains such as Product Data Management (PDM), item-level Product Lifecycle Management (PLM) (see section 3.1), quality monitoring (see section 3.2) and basics from intelligent manufacturing systems (see section 2.2) and information and data management (see section 2.3). A major challenge within the development of such a holistic concept is how to identify relevant information given the incomplete knowledge base and high complexity of the task.

1.3 Research goal and research methodology

The research goal of this dissertation is to develop a holistic concept that describes a manufacturing system by utilizing the product's changing state and the relations and information that entails. The focus of this concept is on identification of relevant information, data and knowledge of both, the product and the manufacturing programme (incl. processes and operations), and how this can be utilized.

Within this concept, a methodology is established (Löhr-Richter, 1993) to identify the relevant set of information a manufacturing programme and subsequent sets of relevant information for individual processes (operations) in order to provide a comprehensive basis for a holistic information management (IM) that may contribute to increase process quality and the final product quality. Identification in this aspect meaning to provide users or customers with the knowledge of what information they need and why they need it (context & application) (Devadason & Lingam, 1997; Tilson, 1998). Within this concept and integrated methodology, inter- and intra-relations (incl. hidden ones) between different product states over the whole manufacturing programme are also considered. This dissertation will contribute further to connect the product and the process perspectives in manufacturing systems through the handling of attached information as both have to be considered to reach the quality goal (Brinksmeier, 1991; Jacob & Petrick, 2007; Yul & Wang, 2009; Köksal, Batmaz & Testik, 2011). This will in turn support the transparency of IM in manufacturing systems.

The final result will be a holistic concept that describes a product by its states along a distributed manufacturing programme and organizes the relevant information in a logical way. It will incorporate knowledge, information and data about the product and process (e.g., process intra- and inter-relations and influential state drivers). Through this enhanced content, the stakeholders may gain access to information and knowledge, which may be utilized to increase the overall quality, decrease rework and scrap and thus reduce the waste of resources.

An important aspect of developing the *product state concept* is the identification of relevant information for a manufacturing programme. Within this aspect, the existing correlations between product states and different processes along the manufac-

turing programme are important to consider. As those correlations are just partly known, this presents a specific challenge for the research goal. In order to include hidden and unknown correlations and identify relevant state drivers (relevant information) along the manufacturing programme, an approach based on supervised Machine Learning (ML) is being developed which can indicate hidden cause effect relations by showing unknown correlations. This allows to identify relevant information of complex manufacturing programmes dynamically and to utilize implicit knowledge available on data level. Through a continuous application of this approach, the set of relevant information for the manufacturing programme is continuously becoming more complete and new relations may be discovered.

However, it is not the goal to generate new knowledge about not yet discovered characteristics of products. The concept will represent primarily a framework to organize all available and connected information and help to provide it to the selected addressee in need. The research will use existing knowledge of characteristics and process intra- and inter-relations and support practitioners handling and using the information efficiently.

1.4 Structure of the dissertation

In this subsection, the structure of the dissertation (see Figure 2) and the motivation behind it is elaborated.

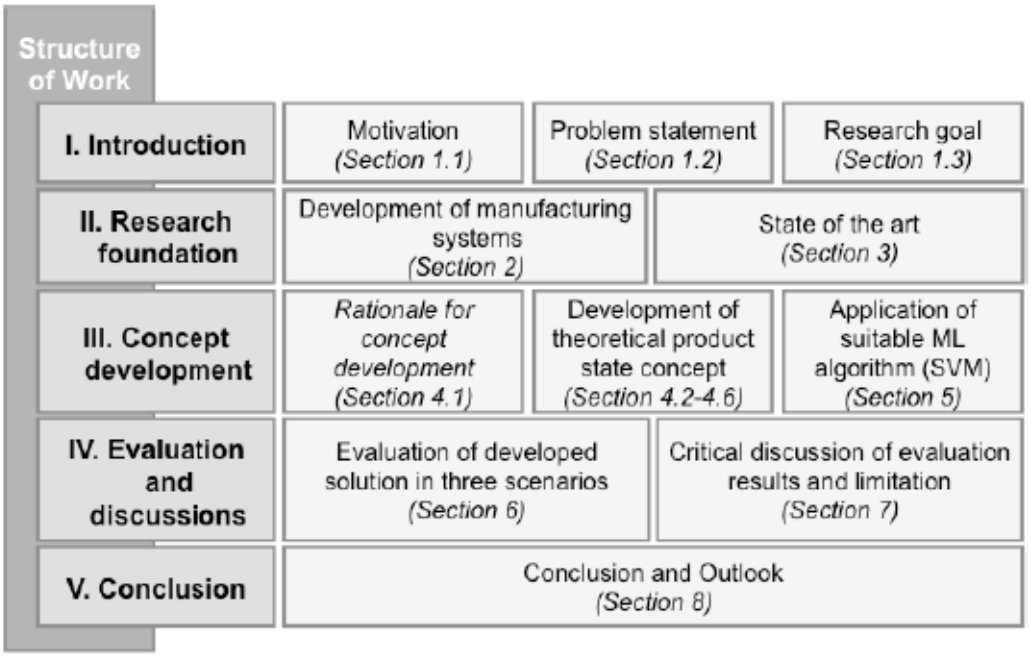

Figure 2: Structure of dissertation

Before the various sections and their content are presented, a general remark concerning the overall structuring of this specific work is necessary. As specified in the previous sections and detailed thereafter, the goal of this dissertation is to de-

velop a holistic concept to describe a product comprehensively along a manufacturing programme through relevant information. During the development of the concept, a major constraint surfaced which is the lack of knowledge concerning the mapping of process inter- and intra-relations between states. This leads directly to it being necessary to investigate an additional field, ML, in order to bring the developed concept to life despite the identified limitations. Therefore, the dissertation contains an additional, brief reflection of the state of the art in the ML domain and a further specified problem analysis and research question. Overall, this necessity to include an additional approach from a different domain yields the slightly deviated structuring of the manuscript compared to average dissertations within the field. Furthermore, this added complexity and additional descriptions lead to an extended page count.

The introduction section (***section one***) of this work illustrates the general motivation behind the conducted research, outlines the research problem statement, research goal and chosen research methodology (see Figure 2).

The ***second section*** the domain and the challenges to be tackled with the dissertation are presented in greater detail. Initially, it offers a general understanding of important background knowledge and definitions on which the developed concept is based upon. This is framed by recent developments in the domain of Manufacturing Systems (MS) from an information perspective. At first a general understanding of the manufacturing domain is presented, focusing in more detail on manufacturing processes, products and quality in manufacturing (section 2.1). Widening the view on manufacturing, manufacturing systems, including holonic and intelligent manufacturing systems are described thereafter (section 2.2). Additionally, being omnipresent throughout the previous sections, the role of information and data management in manufacturing is discussed in more detail in section 2.3. Concluding section two, identified challenges of MS from a product and process information perspective are summarized, highlighting the research problem fundamentals as a basis for the next sections (section 2.4).

The following ***section three*** introduces existing methods and approaches that are dealing with the sketched research problem domain of IM in dynamic and complex MS. The identified approaches are clustered within two main areas. PLM on the one side, including PDM and closed-loop and item-level PLM in manufacturing (section 3.1). These concepts share many overlaps with the later developed *product state concept*. On the other side are approaches from the quality monitoring domain that focus on the previously identified challenges which are described in section 3.2. Based on the analysis of these current methods and concepts, their limitations towards a holistic IM in manufacturing systems are identified. In this concluding sub-section the gaps the product state concept intends to fill are highlighted (section 3.3).

Section four presents first the rationale for the *product state concept* development (section 4.1), highlighting the fit with the identified requirements and challenges of manufacturing systems by picking up the key findings of the previous sections. Describing the structure of section four in greater detail, firstly, the term product state, its origin and definition is described (section 4.2). Next, the topic of relevant state characteristics is discussed on a theoretical level, playing a major role in the following argumentation (section 4.3). Directly related to that, the process intra- and inter-relations of the aforementioned state characteristics are presented in section 4.4. In section 4.4.2, visualization approaches of the *product state concept* are illustrated directly followed by a discussion of the limitations and challenges within this concept (section 4.4.3). In this sub-section an additional research question, which is essential for the successful application of the *product state concept* is identified for the first time. This is expanded on by deriving requirements of state driver identification from the previous findings, describing the NP complete status of the problem at hand and arguing the suitability of applying supervised ML techniques for the identification of state drivers (section 4.5) as a promising way to handle the challenges identified. Concluding, a first basic research hypothesis, specific for the derived research problem is presented (section 4.6)

Based on previous findings, ***section five*** investigates the application of ML algorithms within the product state concept. First, ML in manufacturing is investigated briefly in order to provide a foundation for the selection of a suitable algorithm for the presented research problem (section 5.1). Based on the previous findings, section 5.2 presents Support Vector Machines (SVM) as the ML algorithm of choice and provides details background information on its development, functions and application areas in manufacturing. Furthermore, a solid argumentation for the choice is presented in this sub-section. The final sub-section focuses on the theoretical application of SVM within the *product state concept* by highlighting the application and evaluation approach and giving an outlook on the outcome to be expected.

In order to evaluate the proposed approach, ***section six*** presents the application of the SVM algorithm within the *product state concept* on three scenarios resembling differently structured 'real world' manufacturing programmes and different challenges from the available manufacturing data structure. The first scenario consists of a data set from a manufacturing process of a highly stressed product from the aviation domain provided by Rolls-Royce, whereas the second scenario provides insights in a chemical manufacturing programme. Both scenarios are supplemented by synthetic data adding additional process steps. The third scenario resembles a complex semiconductor manufacturing process. Before applying the proposed approach within the scenarios, the data sets are introduced (section 6.1). The following sub-sections 6.2 (scenario I), 6.3 (scenario II) and 6.4 (scenario III) illustrate the application process in depth documenting all executed steps for each scenario.

The evaluation results of the conducted research are critically discussed in ***section seven***. First the results are elaborated in detail (section 7.1) before the interpretation and critical discussion (section 7.2) structured around the developed research hypotheses is presented. Finally, the limitations of the conducted research and the evaluation results are identified and discussed (section 0).

The last and ***eighth section*** critically questions and reviews the achieved results and knowledge gained of this work and puts it in the greater context. Furthermore, an outlook is presented identifying further research areas related to the findings.

2 Developments of manufacturing systems with a focus on product and process quality

In this section MS as well as recent developments in the area of holistic IM and related topics will be presented. Furthermore, certain basic aspects of manufacturing, MS and related areas are described in detail in order to allow readers to familiarize themselves with the fundamental terms and definitions used throughout this dissertation. In each subsection, concluding paragraphs summarize how the described topic is relevant to the research and putting it in perspective. Main principles and how they are utilized throughout this dissertation is summarized there.

First the manufacturing domain is illustrated, focusing on manufacturing processes, products and manufacturing itself, highlighting process monitoring, process control and process diagnostics. This first subsection is rather descriptive, building a basic understanding of the terms and definitions. As product and process quality and its understanding is used differently in varying contexts, in this section, the definitions of quality related terms and approaches fundamental to the recent developments in manufacturing systems are derived. Presenting holonic and intelligent manufacturing systems in the next subsection, as they are a widely recognized conceptual and holistic view on modern manufacturing. In the previous sections, the connection to the information and data perspective is omnipresent. Therefore, an introduction to information and data management in manufacturing, incl. Big Data and information quality is presented. Concluding, key challenges of the recent developments in MS from a product and process information perspective are discussed.

2.1 Manufacturing terms, definitions and developments

In this section the principle understanding of manufacturing, manufacturing processes and products in this domain is presented. On the highest level, the term manufacturing describes the production of goods using labor and machines, tools, processing, or formulation (see Figure 3) (Steven, 2007; Jehle, 1999). Today, manufacturing is mostly connected to industrial production. Hereby it has to be noted that while the terms production and manufacturing are frequently used interchangeably, their inherent meaning differs to some extent. Whereas it is true that every type of manufacturing is also production, not all production is necessarily manufacturing as it describes converting input to output in a broader term. An example for a production which cannot be described by manufacturing is a book. Whilst the making of the physical book itself can surely be manufactured, the content, the creative work cannot. Despite various researchers argue that manufacturing can also produce non-material products (e.g., Morris & Johnston, 1987), in this research, manufacturing in understood as the making of material goods (see Figure 3) in accordance with Filos (2013).

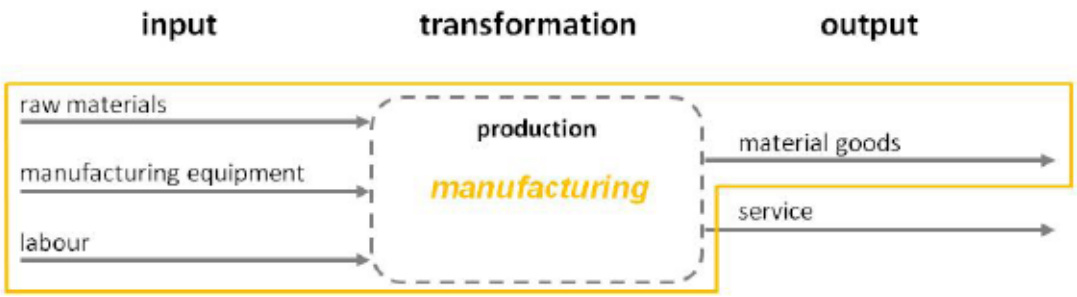

Figure 3: Manufacturing as a transformation process to create material goods as an output

According to Filos (2013), "*manufacturing is the activity to make goods, usually on a large scale, through processes involving raw materials, components, or assemblies with different operations divided among different workers. Manufacturing encompasses equipment for materials handling and quality control and typically includes extensive engineering activity such as product and system design, modeling and simulation, as well as tools for planning, monitoring, control, automation and simulation of processes and factories. It is increasingly seen as a priority area of economic activity especially for economies that have been hit by the recent financial and economic crisis.*"

There are five different manufacturing principles regarding the spatial structure of manufacturing in a facility, the workbench principle, the on-site principle, the function or job-shop principle, the cellular principle and the flow principle (Lödding, 2013). Within this work the focus lies on function or job-shop principle and the flow principle. Within these, products are transported between stations where different transformation processes are conducted to change their state.

Looking at the production types, within this work, the focus lies on mass production with a large number of production runs and continuous production. Also a possible applicable area is a serial production with a large size of production runs. However, a large number of products manufactured are needed as a bases for the developed concept. Next, the basics of manufacturing processes are introduced.

2.1.1 Manufacturing processes[3]

A process is a pattern, designed for a certain purpose. It can describe different variants of combinations of activities or events which are related through causal and/or timed order relations directly to a process confining and activating activity or an activating event and a connected and related result (event or state). This can happen through relations to other activities or events of the process (Hoffmann, Goesmann & Kienle, 2002). This very general definition of a process can be fur-

[3] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest & Thoben, 2012)

ther sharpened and related to the manufacturing domain, looking at the DIN EN ISO 9000:2005 definition of a process. There a process is defined as "set of inter-dependent or interrelated tasks transforming inputs in outputs" (CEN, 2005).

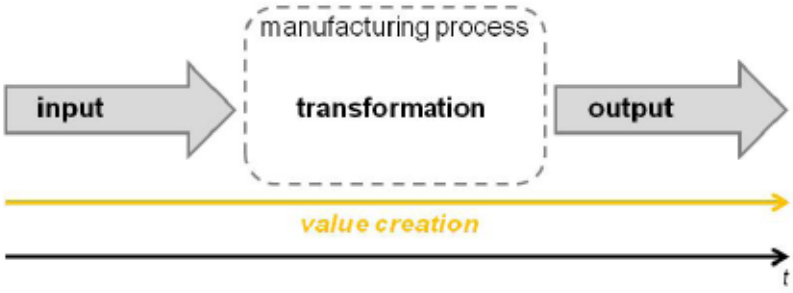

Figure 4: Transformation model in manufacturing and value creation

Manufacturing techniques, used for transforming e.g., products geometry or state can be classified as follows (see Figure 5).

<table>
<tr><td>Create cohesion</td><td>Maintain cohesion</td><td>Reduce cohesion</td><td colspan="2">Enlarge cohesion</td></tr>
<tr><td rowspan="5">1. Primary shaping</td><td colspan="3">Change form</td><td rowspan="5">5. Coating</td></tr>
<tr><td>2. Forming</td><td>3. Cutting</td><td>4. Joining</td></tr>
<tr><td colspan="3">6. Changing material properties</td></tr>
<tr><td>Restoring of elements</td><td>Elimination of elements</td><td>Addition of elements</td></tr>
</table>

Figure 5: Classification of manufacturing techniques according to DIN 8580 (CEN, 2003)

The presented techniques are in general not applied individually but in combination. The six primary techniques are described in more detail in the following list:

- *Primary shaping:* describes the creation of material object out of shapeless matter. By applying certain processes, e.g., casting, cohesion is created. Primary shaping techniques are mostly applied in early parts of a manufacturing programme (Grote & Feldhusen, 2007).
- *Forming:* this technique is changing the form of a product whilst maintaining the cohesion. Through processes like e.g., rolling the elements are restored without changing the mass or cohesion (Fritz & Schulze, 2006).

- *Cutting:* describes the production through changing the form of a product by reducing the cohesion and elimination of elements. Cutting represents an important area of manufacturing (König & Klocke, 2008).
- *Joining:* summarized processes to join two or more parts or products. Examples for processes are adhesive bonding or welding (Westkämper & Warnecke, 2010).
- *Coating:* is realized by permanently adding a shapeless material as an outer layer on a physical body. The added layer can e.g., improve the friction behaviour (Grote & Feldhusen, 2007).
- *Changing material properties:* whereas the above stated manufacturing techniques change the outer form of a product, this one changes the material properties within the product itself. The changing of properties can be done by applying physical processing, chemical processing or biological processing (Steven, 2007) e.g., heat treatment.

The transformation within a manufacturing process can either be base on actions of humans or machines (Zingel, 2009). This definition is already very closely related to the manufacturing definition presented above describing manufacturing as a transformation process (see Figure 4). The transformation within a manufacturing process needs time; a direct production of outputs is not possible (see Figure 4). This implies that a manufacturing process mostly involves more than one stage or sub-processes (Gutenberg, 1970). As the result of a manufacturing process is a product, which represents the customer needs, the manufacturing process is necessarily part of a business process or a business process (Körndorfer, 2003) with the goal of adding value to the product (Porter, 2008; Hutton & Denham, 2008).

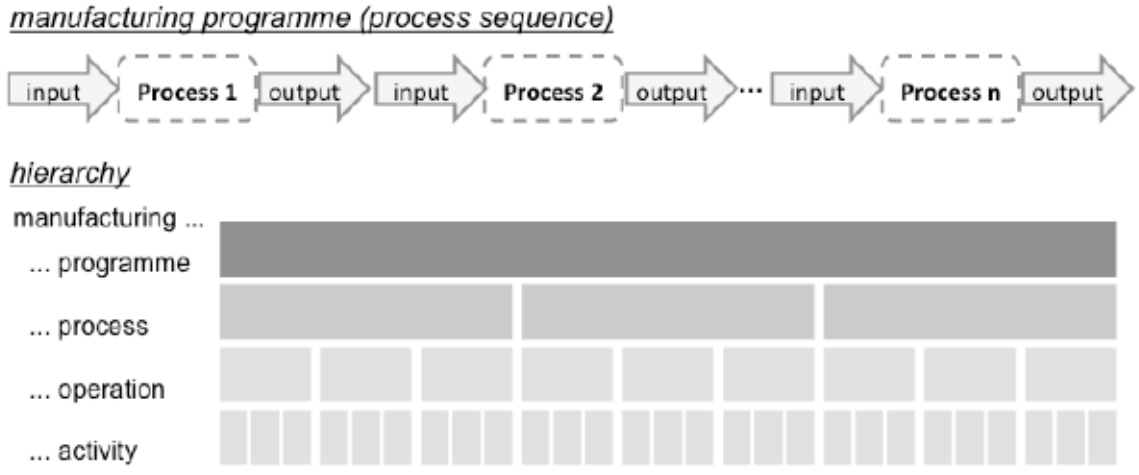

Figure 6: Process sequence and hierarchy (adapted from Becker, 2008)

Figure 6 illustrates a manufacturing programme (sequence of manufacturing processes) connecting input of process n with output of process n-1 through interfaces. These interfaces can either be internal or external. The terminology of the process hierarchy used in this dissertation is presented in the figure as well. A manufactur-

ing programme represents the manufacturing system with all manufacturing processes, -operations down to the individual manufacturing activity involved.

After having clarified what a manufacturing programme, -process, etc. stands for in general and having introduced the main techniques, next, a more detailed discussion will focus on implications of a manufacturing process and its relation to quality. When looking at improving manufacturing processes, as a first step towards efficient manufacturing, it has to be ensured that the manufacturing processes, the entire manufacturing programme for that matter, design is capable to produce the desired product properties (Mohanty, 2004). After this overarching requirement, a functional process design, is given, the process quality plays a major role, as it is directly connected to product quality (Brinksmeier, 1991; Jacob & Petrick, 2007). In every process, a certain degree of variation of the input parameters of individual products can be found even in state of the art manufacturing which can influence the product quality (Yu & Wang, 2009).

The product quality can be influenced at the end of the manufacturing programme (final product quality) or during the different processes or operations. It is important to consider, that the processes and operations are often linked via process intra- and inter-relations to each other and thus, the variations can, even being tolerable from an individual (isolated) process perspective, lead to an unacceptable accumulation causing failure of the final product to meet the customer requirements (Wuest, Irgens & Thoben, 2013b). Taking a closer look, some of these influences are not or just partly known today and in most cases hard or impossible to quantify (with monetary and technical restrictions) as it is mostly very specific to product and process. In this context, the system view gains importance as new research indicates that an isolated focus on single processes during monitoring or improvement initiatives may lead to an incomplete understanding of relations (Zantek, Wright & Plante, 2006; Jiang, Jia, Wang & Zheng, 2012). This is further illustrated in section 2.2.3. Furthermore, Viharos & Monostori (1999) state that having reliable process models is extremely important, as they are required e.g., for selecting optimal parameters during process planning, for designing and implementing adaptive control systems or model based monitoring algorithms.

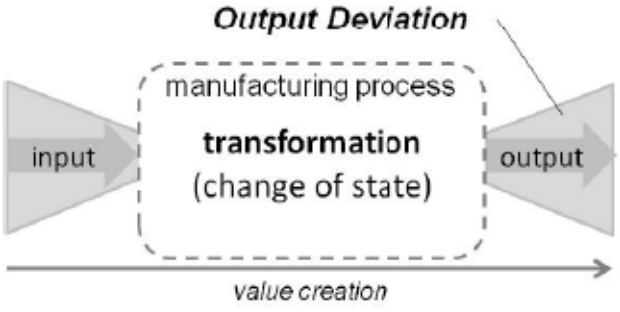

Figure 7: Input and output deviation of manufacturing process

Looking again at the input-output model of a manufacturing process as presented in Figure 4, transformation in this case can be described as a change of the product state, and thus of one (or multiple) relevant state characteristics, from input (product state) to output (product state). Every manufacturing process has an input product state, which deviates to a certain extent from the originally planned input (see Figure 7) (Ding, Shi & Ceglarek, 2002). The term product state used here describes a product at a certain point during a manufacturing programme. This will be presented in greater detail in section 4.2. This deviation is always there and due to some degree to process 'noise' such as machine/material variability, environmental factors, thermal effects, operator error, etc. (Kaiser, 1998). The level of the deviation and with it, the impact on the product quality, however varies a lot. These deviations of the input product state have an influence on the output product state after the state change (transformation) if transformation parameters can be considered unchanged (see Figure 8) (Jansen-Vullers et al., 2003).

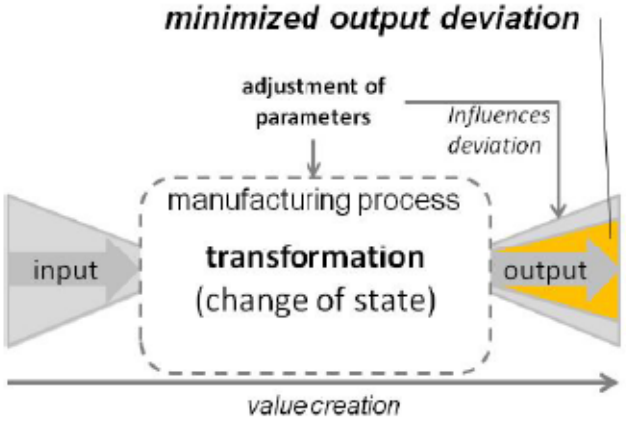

Figure 8: Output deviation based on adjustment of parameters of manufacturing process

One option is to base the adjustment of the process parameters on information of the product, for example the input product state (see Figure 9). Taking known cause-effect relations and the process view into account, this information can be described by the *product state concept*, introduced in this research (see section 4.2 ff.). The comprehensive approach can contain all necessary information needed by processes involved.

However, identifying this set of relevant information is not trivial. A special focus within such a concept has to be laid on the question, how can the relevant information which provides the basis for the adjustment be identified. Once this relevant information, among it being the drivers of product state, is identified, experts can apply it to adjust the process on an informed basis accordingly.

Manufacturing is an area with a constant need for efficiency and product and process quality improvement. There are many different areas in manufacturing tackling this issue. In order to structure the following findings, the areas are grouped in

different domains. These domains constitute of areas with similar requirements and challenges towards supporting techniques and technologies. However, there are various forms of semantics out there and as the areas are not sharply distinguishable in their focus, overlaps between the domains will occur. The three overarching domains are monitoring, diagnostics and control. They are complementary in as much as it is necessary to monitor in order to control and without diagnostics control is unfocused/undefined. The additional domain of scheduling stands out as it is not directly related to the above. As this work is focusing on monitoring, the domains of control, diagnostics and scheduling are briefly introduced in the following paragraph, before process monitoring is detailed in the next subsection.

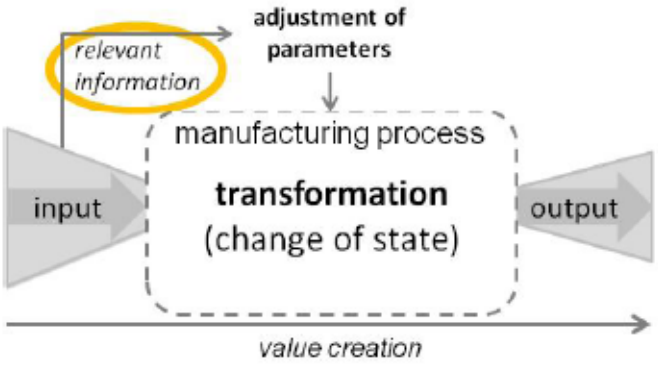

Figure 9: Importance of relevant process information process parameter adjustment

The domain of ***control*** includes a wide variation of areas and is closely related to monitoring (Kang, Choe & Park, 1999). Control is the action of bringing a process back into a desirable state. Harding, Shahbaz, Srinivas & Kusiak (2006) state that "[ML] and computational intelligence tools provide excellent potential for better control of manufacturing systems". The areas represented by this domain include but are not limited to (intelligent) manufacturing control (e.g., Bowden & Bullington, 1996; McFarlane, Sarma, Chirn, Wong & Ashton, 2003), (statistical or automated) process control (e.g., Qin, Cherry, Good, Wang, & Harrison, 2006; Jenab & Ahi, 2010) and simulation (e.g., Baker, 1988; Fowler, 2004). The domain of ***diagnostics*** (e.g., Chinnam & Baruah, 2009) compromises the areas of process analysis (e.g., Arbor, 2000) and fault diagnosis (e.g., Widodo & Yang, 2007). Additionally there is the domain of ***scheduling***, which is required to ensure the control and/or process actions happen in the right order. However, scheduling, as part of internal and external logistics will not be in the focus of this work. In order to present a rather complete picture, the different areas summarized under scheduling are: scheduling (e.g., Aytug, Bhattacharyya, Koehler & Snowdon, 1994), sequencing (e.g., Lödding, 2013) and capacity planning (e.g., Lutz, Boucher & Roustant, 2012).

2.1.2 Process monitoring

Using product and/or process data to monitor and/or forecast certain events, chains of events and/or outcomes is a topic, widely discussed among scholars for more than the last 20 years. Du, Elbestawi & Wu (1995) describe monitoring as an act of identification of characteristic changes of a process by evaluating process data without interfering running operations. Stavropoulos, Chantzis, Doukas, Papacharalampopoulos & Chryssolouris (2013) describe monitoring as the manipulation of sensor measurements (e.g., force, vision, temperature) in determining the state of the processes. Ge, Song & Gao (2013) define process monitoring simply termed as fault detection and diagnosis, and as a tool for process safety and quality enhancement. These definitions already highlight again the connection to process control and process diagnostics as described before. The task of monitoring is to separate the normal process data samples from the faulty ones (Ge, Gao & Song, 2011). The extraction of useful information from the recorded process dataset enables the monitoring and prediction of the process operation condition and the product quality (Ge et al., 2011).

Due to increased number of variables measured and monitored and the improved controllability of these variables a method of analyzing the data is required. Without an appropriate method only limited data about the processes can be extracted (Lee, Yoo & Lee, 2004). Du et al. (1995) find in their research that monitoring based on learning from examples turns out to be more effective in manufacturing programmes than learning from instructions.

Monitoring in manufacturing includes the areas of machine performance monitoring (e.g., Spoerre, & Ben Wang, 1995), (machine) condition monitoring (e.g., Peng, 2004; Widodo & Yang, 2007), quality monitoring (e.g., Ribeiro, 2005; Wuest et al., 2013b) and process monitoring (e.g., Skitt, Javed, Sanders & Higginson, 1993; Qin et al., 2006). More detailed application areas include the analysis of high-dimensional and correlated process data, e.g., in chemical and biological plants and products (Ge et al., 2011), wastewater treatment processes (Lee et al., 2004), model-based monitoring for fault detection and diagnosis in aerospace, engine and power systems (Ge et al., 2013), tool wear and tool breakage (Stavropoulos et al. 2013). The challenges in the domains control and monitoring are very similar, reflecting the large overlap and connection of the two domains. For example, in order to identify a faulty process, the cause-effect relations play an important role. When control kicks in to get the process back on track based on the monitoring information, cause-effect relations are essential in order to take the right measures. Within the monitoring domain the challenges can be stated as follows: unclear/unknown cause-effect relations, high-dimensionality, incomplete (product & process) data. The relevant sub-domain of process monitoring, quality monitoring will be elaborated in a later section (see section 3.2.2). Next, the term product and its understanding within the manufacturing domain is introduced.

2.1.3 Product in manufacturing

In the definition of manufacturing introduced before, the purpose of manufacturing is the production of material goods. In industrial production, these goods can be referred to as products. As the term product is a central aspect of the developed concept, first commonly accepted definitions are presented, before the agreed upon understanding of the term is presented.

A general definition describes a product as representing an output offered on the marketplace which satisfies the customer needs through specific functions and characteristics in a beneficial way. The output can be material goods, services, information or experiences (Kotler, Armstrong, Saunders & Wong, 2011).

According to the Quality Management (QM) standard (DIN ISO EN 9000:2005), representing an engineering perspective, a product is defined as the final result of a process. The results of the previously discussed manufacturing processes can therefore be defined as products. During the manufacturing process, the to-be-transformed material is referred to as "work piece" (CEN, 2005). Finalizing the manufacturing process, the work piece becomes a product (see Figure 10). It has to be noted, that two or more work pieces can be combined to a single product (Schmachtenberg, 2000).

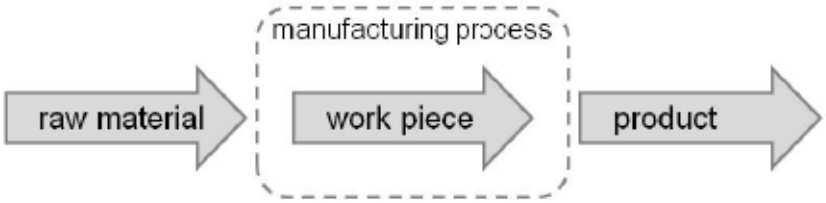

Figure 10: Raw material, work piece and product in relation to a manufacturing process

In manufacturing, the PLM perspective is increasingly gaining attention. In closed-loop, item-level PLM, an object over all phases of its lifecycle, beginning from raw material over work piece and final result of manufacturing (product) to the to-be-recycled materials after usage are considered and referred to as product (Jun, Kiritsis & Xirouchakis, 2007; Terzi et al., 2007; Taisch, Cammarino & Cassina, 2011) (see section 3.1).

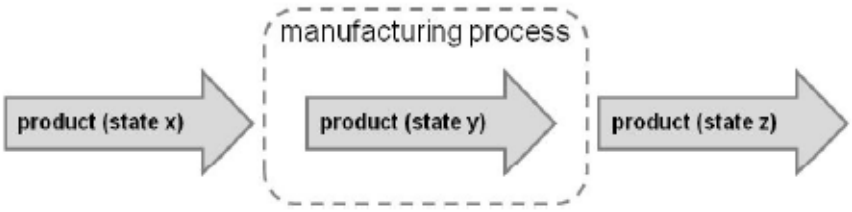

Figure 11: Product with changing state during a manufacturing process

In this research, the term product is used comprehensively to describe an artifact over various stages of its product-life-cycle replacing the more technically accurate terms e.g., raw material before and work-piece during manufacturing. Reasons include the focus on individual products (item-level) and the reduction of complexity. Based on this understanding of the term product, the product state describing a product at different stages of a manufacturing programme, will be defined later on (section 4.2) (see Figure 11). Next, basic quality terms and definitions are presented in the following subsections.

2.1.4 Quality in manufacturing

Quality has been a focus area of manufacturing for several decades and the market success of companies successful in utilizing their understanding of quality and customer requirements highlight the importance of quality. De Weck, Ross & Rhodes (2012) found in their recent study on system lifecycle properties ('Ilities') that quality is and was the most dominant 'ility' of engineering systems for over a century, rated higher than e.g., reliability and safety (De Weck et al., 2012).

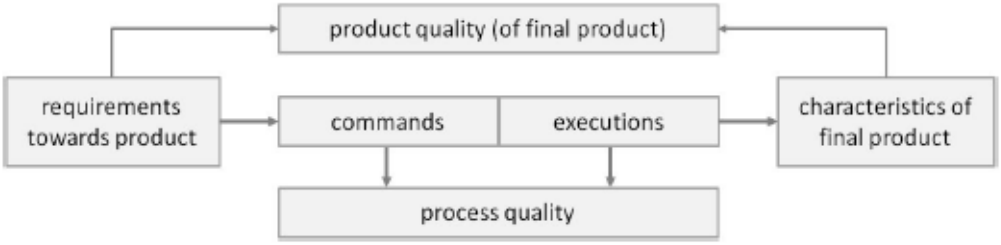

Figure 12: Elements of quality (adapted from Masing, 2007; Sitek, 2012)

In this research the term quality is understood as "the degree to which a set of inherent characteristics fulfils requirements" (DIN EN ISO 9001:2008 – CEN, 2008). Requirement within this context is defined as the "need or expectation that is stated, generally implied or obligatory" (DIN EN ISO 9001:2008 – CEN, 2008). According to this definition, quality depends on the fulfillment of requirements. The fulfillment of these requirements depends on the planning of processes (commands) and the execution of processes (executions) (see Figure 12) (Masing 2007). Quality of the final product is regarded as achieved to a higher degree when more of the original customer requirements match with the achieved characteristics of the final product (Sitek, 2012). The *product state concept* corresponds with this definition as it defines so called state characteristics by which the state of a product can be described at all times during its lifecycle. It has to be considered that a product can inherit different qualities, the sum of these (sub-)qualities like e.g., security, workmanship or durability finally represent the final product quality (Kamiske & Brauer, 2008).

There are different definitions of quality in manufacturing available. Some researchers have a very technical view on quality in manufacturing. An example for

such a technical definition is presented by Kaiser (1998), who defines quality in manufacturing as "primarily a factor of machining tolerances" This implies that quality can be achieved when the machining tolerances are controlled. This view does not reflect common problems like input deviations or environmental influence on the processes. Other researchers define quality in manufacturing more generally, as "confirming with requirements", thus focusing on the customers (Garvin, 1984). However, researchers agree, that in most cases "products with small variations in shape and size are considered high quality, while products with large variations are considered poor quality" (Kaiser, 1998). This corresponds with the former definition, as "a product that deviates from specifications is likely to be poorly made and unreliable" (Garvin, 1984). However, these variations have to be viewed form a customer requirement perspective. Some variations of parameters not important for the customer with not impact on other important parameters have no influence on quality. As manufacturing companies constantly try to improve the quality of their products and processes, it has to be noted that quality improvement generally requires collection and analyses of data to solve quality related manufacturing problems (Köksal et al., 2011).

According to the quality definition above, the final product quality depends on the fulfillment of the customer expectations and thus the customer requirements. Overall, there are many possible reasons for a discrepancy from these requirements, e.g., the requirements of customers where not correctly retrieved or the designers interpreted and transformed the requirements differently than the customer fancies. However, within this research it is assumed that the requirements were correctly retrieved and the product will fulfill the customer expectations if it meets the specifications set by the designers and process planers. Using the terminology of Figure 12, the commands are considered correct and the execution is the focus area. The reason behind this is that this research is focusing on supporting the manufacturing process and does not directly support phases like e.g., the design or product planning. Following Taguchi's (1989) six stages of activities of manufacturing companies, the focus lies on stage (4) manufacturing and partly (3) manufacturing process design, whereas the stages (1) product planning, (2) product design, (5) marketing and (6) sales will not be looked upon.

2.1.4.1 Product quality

The term product quality has been introduced partly in the previous section. As can be seen in Figure 12 product quality is determined by the fulfillment of the (quality) requirements by the characteristics of the final product. To adapt this definition to the process and system view, product quality can also be determined for processes and/or operations within a manufacturing programme. The final product is to be understood as the outcome of a process or operation instead of the overall manufacturing programme. However, the requirements are not as easily determinable because of existing cause-effect relations between different processes and/or

operation during the manufacturing programme. In manufacturing programmes, a wide variety of potential errors can influence the quality characteristics of a product. The product end quality is finally determined by all stages of the manufacturing program (Zantek et al., 2006; Jiang et al., 2012). This challenge is addressed by the *product state concept,* as it is one of the pillars towards the identification of a set of relevant information (see section 4.4).

Some quality characteristics can be easily measureable, for example length, depth or weight, some are hard to measure, like functions or aesthetic. Easily measurable characteristics have the advantage of being easier to monitor and control. The quality characteristics being hard to measure are mostly hindering the checking of the fulfillment of requirements. Additionally, quality characteristics are an element for control of the impact of quality management processes (Eversheim, 1997). It is however a challenge to determine the actual real life requirements according to which the product quality is finally determined (Olbertz & Otto, 2001). As stated above, this question is not in the focus of this research.

2.1.4.2 Process quality

Quality principles cannot just be applied to product but also to processes. The process quality definition depends to a large extent on the understanding of process itself. A process, e.g., a manufacturing process, inherits a specific order of transformation activities alongside temporal and spatial dimensions with a defined input and output. The quality of a manufacturing process is determined by the compliance with criteria for order, time, place, input and output (Kreutzberg, 2000).

Process quality determines the product quality, given that the entire manufacturing programme and product/process design is capable of meeting the requirements, (Brinksmeier, 1991; Jacob & Petrick, 2007) (see Figure 12). Even if a process is executed with the exact same parameters, a certain degree of variation of the input parameters of individual products can be found even in state of the art manufacturing which can influence the process quality and thus the product quality (Taguchi, 1989; Yu & Wang, 2009).

It is a major task of QM to ensure a high process quality in manufacturing. Continuous improvement is widely employed in order to reduce failure and to optimize manufacturing processes and the quality of the output (Eversheim, 1997). This QM tasks, involving a lot of information and data and efficient handling of such, are introduced in the following subsections.

2.1.5 Example of a manufacturing programme[4]

A manufacturing programme consists of different processes and operations, each with a certain very specific task and goal. To transform a raw material to a final product, all processes are necessary and have to be executed in a certain order. To make the theory introduced in the previous sections more feasible, an exemplary description of the manufacturing programme of a highly stressed steel product will be presented. This example is based on an adapted manufacturing programme following (Klein, Thoben & Nowak, 2005) which consists of three process steps: forging, machining and heat treatment (Figure 13).

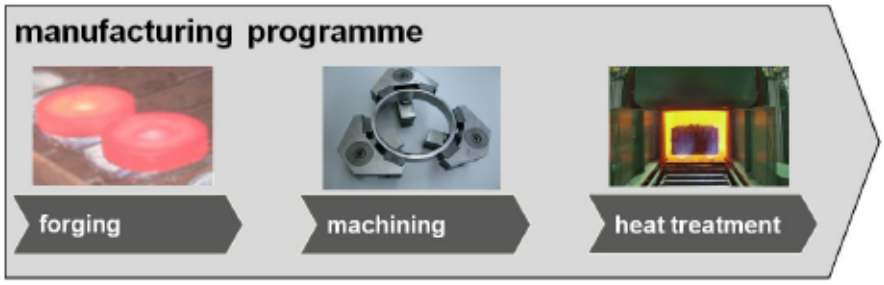

Figure 13: Exemplary manufacturing programme with three processes

In industrial practice, a manufacturing programme involves generally more processes and/or some have to be executed multiple times at different stages of the whole manufacturing programme. To build a foundation for the following concept, the author chose to use a simplified example in order to focus on the main ideas behind the concept instead of getting lost in details.

Today, it has to be taken into consideration that manufacturing programmes are not executed by a single company at a single location any longer but rather in collaboration with other companies (Seifert, 2007). This includes extra interfaces and interdependencies between stakeholders as well as manufacturing and business processes. For example, could the forging (process 1) in the exemplary manufacturing programme (see Figure 13) be done by company A in country X, whereas the processes machining (2) and heat treatment (3) are executed by company B's department C (country Y) and D (country Z). As this adds further complexity to the manufacturing itself by involving additional logistics and information exchange, there is an indispensible need for a clear structure to identify, share/distribute and use product and process information (Merali & Bennett, 2011).

This section presented the basic terminology, e.g., manufacturing, product and process used in this research. It described how manufacturing processes transform

[4] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2013b)

products by adding value. This value adding can be done in various ways, e.g., machining or heat treatment. Furthermore, the importance of the availability of the right information for a manufacturing process is introduced. Together with process monitoring, which can be understood as the capturing of information in a manufacturing process this is the basic principle the *product state concept* is build upon.

2.2 Developments of manufacturing system

A manufacturing system describes the method of manufacturing in a generalist way. A manufacturing system sub-summarizes all means necessary for the production of a certain product, including the manufacturing programme and processes, machines, production method, etc. It represents the overarching layer connecting the different stakeholders involved and is mostly complex and of large-scale (Höpf & Schaeffer, 1997). Koren, Hu & Weber (1998) have shown that the configuration of the manufacturing system affects the performance of the system. The effects identified include productivity, capacity scalability, and part quality and thus, influence the lifecycle cost of the manufacturing system.

Over time there have been many different methods and concepts concerning manufacturing systems. From flexible and integrated manufacturing systems (e.g., Collins, 1980; Kimemia & Gershwin, 1981) towards today's Holonic (HMS) and Intelligent Manufacturing Systems (IMS) (e.g., Höpf & Schaeffer, 1997; McFarlane & Bussmann, 2003) there are many different definitions available, each focusing on certain aspects with smaller or larger overlaps with each other. Even so some concepts were adopted in the 1980ies, they are still to some extent valid and applied in their original or adapted/updated form today (ElMaraghy, 2006).

Flexibility is still an important factor for today's manufacturing systems especially given the trend towards mass customization (He, Zhang & Li, 2013). However, the focus is increasingly shifting towards a combination of reconfigurability, flexibility and even adaptability (ElMaraghy, 2006; Almeida, 2011). Reconfigurability, flexibility and adaptability reflect the customer demand driven production of today's business environment. These concepts focus to a large extent on scheduling and production planning and control activities. IMS, in detail explained in section 2.2.2, try expanding that view by expanding the focus on further characteristics like e.g., autonomy, learning and efficiency (Kumar, 2002; Oztemel, 2010; Almeida, 2011). HMS, while based on the IMS concept, focus on the self organization aspects of large complex systems and how this integrates in and influences the performance of the system (McFarlane & Bussmann, 2003) (see section 2.2.3).

2.2.1 System view on manufacturing

The general systems theory (Von Bertalanffy, 1972) and the derived systems perspective has had an effect on various disciplines and has partly been adapted to the needs of various disciplines like operations, information systems and also engi-

neering (Maddern, Smart, Maull & Childe, 2013). A System represents "a set of interacting components having well-defined (although possibly poorly understood) behavior or purpose; the concept is subjective in that what is a system to one person may not appear to be a system to another" (Magee & de Weck, 2004). A complex system expands on the above system definition by being "a system with numerous components and interconnections, interactions or interdependencies that are difficult to describe, understand, predict, manage, design, and/or change" (Magee & de Weck, 2004). Engineering (and thus manufacturing) systems) are "systems designed by humans having some purpose" (Magee & de Weck, 2004).

However, in the manufacturing domain often the focus is on individual processes or operations, disregarding the previous or following ones, which can have an impact on the products final quality. Hoffmann, Keßler, Lübben & Mayr (2002) found that there are cause effect relations across process borders which have a significant influence on the behavior of a product during manufacturing (Sölter, 2010). Such often complex process intra- and inter-relations are common in engineering systems (Giffin, de Weck, Bounova, Keller, Eckert & Clarkson, 2009). In line with the principles of systems theory, the environment of the system also has an influence of the behavior of a system (Maddern et al., 2013). In manufacturing programmes, a wide variety of potential errors can influence the quality characteristics of a product. The product end quality is finally determined by all stages of the manufacturing programme (Zantek et al., 2006; Jiang et al., 2012). Therefore, taking the whole system into account instead of individual, isolated processes can help to accomplish sustainable product and process quality improvements (Zoch, 2009). Supply Chain Management (SCM) represents a very common variant of a system view, focusing mostly on logistics and collaboration efforts (Christopher, 2005), whereas the research focus in this manuscript lays on product and process quality improvements in manufacturing.

2.2.2 Intelligent manufacturing systems

Increasing market pressure towards quality, efficiency and flexibility together with new developments in ICT, Artificial Intelligence (AI) and optimization techniques lead to the concept of intelligent manufacturing. Intelligent manufacturing is also known as smart manufacturing, being used almost interchangeable. A comprehensive definition of smart/intelligent manufacturing is presented by Wallace & Riddick (2013) as follows: "Smart [or intelligent] manufacturing is a data intensive application of information technology at the shop floor level and above to enable intelligent, efficient and responsive operations" (Wallace & Riddick, 2013).

Another definition of intelligent manufacturing describes the concept as "an intelligent manufacturing process [that] has the ability to self-regulate and/or self-control to manufacture the product within the design specifications" (Kumar, 2002). In this definition the autonomous aspect of intelligent manufacturing is

highlighted. What is commonly accepted among researchers is the importance of product and process information and data, technology and (human or machine inherent) knowledge (Chand & Davis, 2013). This understanding already implies that in order to make a manufacturing process intelligent, various functions of a manufacturing company have to work together, e.g., design, process planning, production planning, operations and process control. Looking at the final product, individual quality control and based on that, corrective measures are required. During manufacturing itself, monitoring, diagnostics and measures like predictive maintenance play an important role (Mazumder, 2008). Overall, continuous improvement is crucial to make the system intelligent.

However, the degree of autonomous behavior is not specifically defined. Kumar (2002) defines three ways to achieve the above-defined intelligent manufacturing:

- "Existing manufacturing processes can become intelligent by monitoring and controlling the state of the manufacturing machine
- Existing processes can be made intelligent by adding sensors to monitor and control the state of product being processed.
- New processes can be intelligently designed to produce parts of desired quality without the need of sensing and control of the process."

According to these findings, existing manufacturing processes can be made "intelligent" by monitoring and control the state of the products via sensor technology and the application of ICT. This is highly relevant to the conducted research as the here stated "state of a product" is in line with the basic understanding of products and processes of the developed concept.

The intellectual father of IMS, Yoshikawa, defines them as follows: "The IMS takes intellectual activities in manufacturing and uses them to better harmonize human beings and intelligent machines. Integrating the entire corporation, from marketing through design, production and distribution, in a flexible manner which improves productivity" (Yoshikawa according to Piddington & Pegram, 1993).

The global, IMS program compromises a R&D program established to develop the next generation of manufacturing and processing technologies, led by industry (Nagy, Jering, Strasser, Martel, Garello & Filios, 2005). The first idea for IMS came up by the end of the 1970ies (Hatvany & Nemes, 1978) shortly after followed by early IMS definitions (Hatvany, 1983). Hatvany (1983) gave the next generation of manufacturing systems a perspective combining findings of AI research "to solve, within certain limits, unprecedented, unforeseen problems on the basis of even incomplete and imprecise information". (Monostori, 2002) Being widely discussed, a worldwide IMS initiative, initiated by Japan 1989 (EC, 2009), was formally started in the mid 1990ies with the kick off of six test cases. One of the cases were Holonic Manufacturing Systems (HMS) (TC5), which looked into the ability of companies to react to rapidly changing market conditions (see section

2.2.3), others looked into knowledge systemization in product and process design (TC7), whereas others focused on clean manufacturing (TC2), concurrent engineering (TC3) and rapid product development (TC6), etc. (Kopacek, 1999). Even so the impact of conducted project within the first phase of IMS was positive, there is still potential for future development in IMS especially given the rapid development in ICT (Zobel & Filos, 2006).

According to Kumar (2002), "IMS

1. uses technology which can minimize the use of human brain.
2. regulation for product mix and priority production, self regulated.
3. self controlled operations with automatic feedback mechanism.
4. monitoring and control of the manufacturing machine.
5. monitoring and controlling the state of product being processed.
6. new processes with intelligence can be made to produce parts of desired quality without the need of sensing and control of process" (Kumar, 2002).

One has to bear in mind that the points stated above are rather idealistic goals as a realization in the near future is unlikely due to e.g., the high dimensionality and complexity involved in modern manufacturing and PLM approaches.

IMS, being based on the intelligent manufacturing paradigm, are supposed to support various characteristics, starting with flexibility and reconfigurability combining them with ideas from the ICT domain like autonomy, decentralization, flexibility, reliability, efficiency, learning, and self- regeneration (Liu, Zhang & Venuvinod, 1997; Revilla & Cadena, 2008; Mekid, Pruschek & Hernandez, 2009; Shen, Hao, Yoon & Norrie, 2006; Almeida, 2011).

Looking at the above, the importance of state monitoring of both, processes and products within IMS is evident. This research is contributing to support state monitoring issues in complex manufacturing programmes to support the IMS goals.

2.2.3 Holonic manufacturing systems

The word "holon" is an artificially created term based on the Greek word "holos" meaning whole and the Greek suffix "on" meaning particle or part as in proton or neutron (Höpf & Schaeffer, 1997; McFarlane & Bussmann, 2003). A holon is understood as "an identifiable part of a system which has a unique identity, yet is made of subordinate parts and in turn is part of a larger whole" (Kopacek, 1999). McFarlane & Bussmann (2003) define a holon in manufacturing as an "autonomous and cooperative building block of a manufacturing system for transforming, transporting, storing physical and information objects". Given the above definition, a holon itself can contain a unlimited amount of holons as subsystems, providing the necessary processing, information, and human interfaces to the outside world (McFarlane & Bussmann, 2003).

HMS were originally established as part of the global IMS initiative as TC5 "Holonic Manufacturing Systems" in 1989 to create "companies able to react promptly and efficiently to changes in environmental and marketing conditions" (Kopacek, 1999). Especially SMEs require flexibility in their manufacturing systems to survive in the future global market environment. Holons offer allow those companies to create flexible manufacturing systems based on principles known from ICT. HMSs are supposed to be intelligent, flexible and modular (Kopacek, 1999).

As a basis for this research HMS present an interesting foundation as it combines the detailed view on an "excerpt" (holon) of an overarching system and the implications of its performance/changes and inherent information/data representation. This is strongly related to the approach taken when looking at the manufacturing programme by the different product and process states and the identification of state drivers based on data from different defined sub-systems (see section 5.3). The interpretation of the results strongly depends on how the findings of the analysis of sub-systems affect the manufacturing programme as the overall systems.

In conclusion, the previous subsections highlighted different approaches to describe manufacturing systems. Instead of looking at operations or processes individually, the importance of considering all elements of the manufacturing system, as there are correlations across process borders is described. The *product state concept*, describing a product holistically by its state over a manufacturing programme is a reflection of the system view on manufacturing.

2.3 Developments in information and data management in manufacturing[5]

This section presents a closer look on information and data and its handling and management in manufacturing. Most advanced manufacturing approaches, e.g., the above discussed IMS and HMS initiatives, rely strongly on information and data. The developed *product state concept,* as a holistic product focused information system is dependent on a functional information and data management as well.

Along a manufacturing programme, physical products as well as information are exchanged between the partners (Hicks et al., 2006). The availability of information is a precondition to adjust each manufacturing process in such a way that the outcome reflects the set quality requirements to a high degree. Quality, as stated before, constantly gains importance for customers and for a sustainable use of resources. At the same time, distributed production brings forth new challenges for managing quality (Sitek et al., 2010). Looking at quality improvements of manu-

[5] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest & Thoben, 2012)

facturing products and processes, the collection and analysis of data/information is essential to solve quality related manufacturing problems (Köksal et al., 2011).

In order to present a solid foundation and highlight the current challenges in the domain, first information and data management are presented before looking more closely into information quality and their understanding within the manufacturing domain. Based on this general introduction, selected standards and tools used in practice are presented. The widely discussed Big Data domain is briefly discussed at the end of this section; mainly to distinguish the differences and similarities of the developed concept with regard to the Big Data perspective. Two specific topics related to the domain of information and data management, namely PLM and PDM, are discussed in the next section due to the available practical applications and their close relation to the theoretical foundation of the *product state concept*.

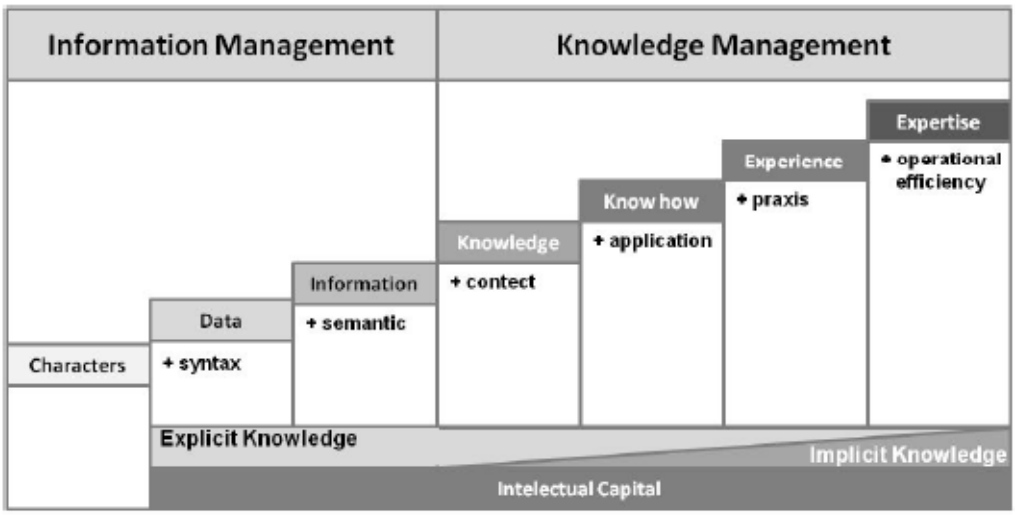

Figure 14: Differentiation of knowledge and information management (Wuest & Thoben, 2012, inspired by (North & Güldenberg, 2008; Auer, 2010))

Before focusing more closely into information and data management, Figure 14 distinguishes the difference between IM (incl. data management) and Knowledge Management (KM). Overall, it can be stated, that information management is considered more technical than knowledge management and that knowledge is backed by information and data. An important differentiation, is that data and information can be stored relatively easy compared to knowledge which is always connected to a person (Probst, Raub & Romhardt, 2006). The differentiation in explicit and implicit knowledge is crucial for the transferability and applicability. They are based on Polanyis findings concerning the personalized nature of knowledge (Polanyi, 1962). In comparison, information is relatively easy to transfer and data even easier. However, information itself does not foster realizations or new findings; it has to be connected to a context in order to become knowledge (Haun, 2002). This is important for the developed concept as the process of connecting data and information with context represents a major challenge during the application.

However, the above illustrated distinctions between the different areas are not as clear as Figure 14 indicates. One of the reasons is, that solid measures are missing resulting in large gray zones and overlaps between the different terms which make a clear distinction impossible at times. Within this manuscript, the focus is on information and data as a source of product state and process knowledge.

KM is the systematic and explicit control of knowledge based activities, programs and governance within the enterprise with the goal to make effective and profitable use of the intellectual capital (Wiig, 1998). (Davenport, De Long & Beers, 1998) emphasize that KM does not only imply successful utilization of knowledge but also creation and allocation. The KM research field is a very broad one and there are various research areas involved, from social science over psychology and business to engineering. Therefore, the number of publications and available information is vast. Setting the focus on identifying knowledge, (Probst et al., 2006) with their model of knowledge building blocks defined one of them as "knowledge identification" (Probst et al., 2006). Taking a closer look, this block describes the need to increase transparency of internal and external sources of knowledge. It also is supposed to ease the way the own employees have access to knowledge needed. The pioneers in the field of KM, (Nonaka & Takeuchi, 1997) created the well-known model of the "knowledge spiral", an illustration of the knowledge creating process focusing on transforming implicit to explicit knowledge. Other concepts, like e.g., process oriented KM (Mertins & Seidel, 2009), are variations or combine the models of Probst et al. or Nonaka & Tekeuchi and combine it with other theories like Porter's value chain (Porter, 2008). None of these approaches and models offers a defined and accepted concept clearly to identify very specific sources of information or data about an individual product or process. But they all emphasize the importance of having the right knowledge or information available at the right place for all business processes. This can be seen as the overarching argumentation for the *product state concept*, as it is supposed to provide the right information/data to experts who can apply their knowledge to improve the process and thus product quality on that basis. In the future expert systems could support knowledge creation based on the *product state concept*.

Another interesting distinction of knowledge, information and data, this time including the relations among each other in both directions is presented by the information pyramid (Fink, Schneidereit & Voß, 2005) (see Figure 15). The highlighted relations in Figure 15 between knowledge, information and data can be seen throughout this research. In order to identify a relevant set of product state information, knowledge of the manufacturing programme, the individual processes and their process intra- and inter-relations has to be applied as well as available process and product data has to be analyzed when there is a knowledge gap.

The question if process and product quality can be improved through transparent IM based on identification of relevant product state characteristics along a manu-

facturing programme through modeling process intra- and inter-relations between these characteristics has not yet been addressed sufficiently in literature or practice. Areas related to this question were identified as follows: knowledge, information and data management; SCM (incl. process management and related areas); research on collaborative production and quality management.

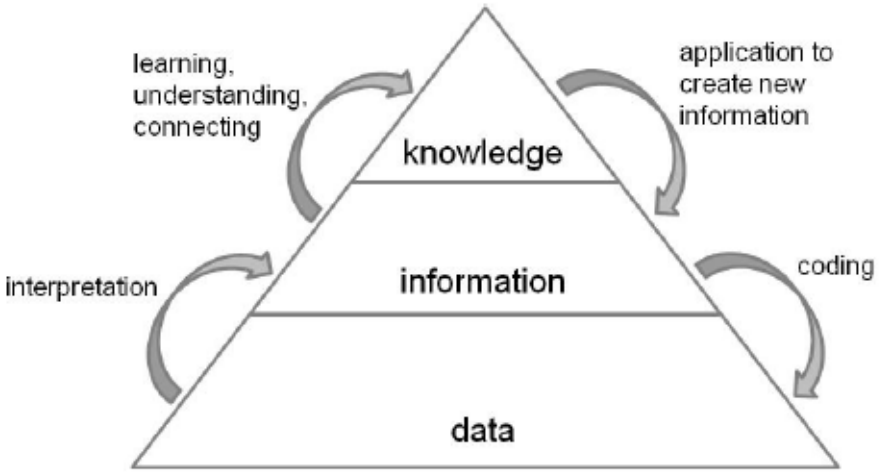

Figure 15: Information pyramid (Fink et al., 2005)

Next, the areas of IM will be explained in greater detail to clarify the domain in focus of this research and provide a solid background before including a brief explanation of the Big Data domain.

2.3.1 Information management (systems) in manufacturing

IM and the more technical term IM systems are strongly linked to ICT and related technical solution in the manufacturing domain. In this area, a lot of progress has been made over the last years. As stated above the understanding of IM and IM systems in manufacturing does not distinguish itself sharply from e.g., PDM and PLM systems, also with strong links to ICT, which will be discussed in next section 3.1. However, IM is closely connected to the quality of the data and information to be managed (Storey, Dewan & Freimer, 2012). Garetti and Terzi (2004) highlight that the "product" information and data management is representing a key aspect of product centric and product driven approaches, also emphasizing a strong overlap between the two areas.

Nevertheless, the research focus of IM in manufacturing is mostly focused on how already existing information has to be managed (e.g., Choe, 2004; Hicks, 2007) or what existing IM system should be chosen (e.g., Beach, Muhlemann, Price, Paterson & Sharp, 2000; Gunasekaran & Ngai, 2004). The general principles of IM (e.g., Augustin, 1990; Jehle, 1999; Hoke, 2011), the right information at the right time in the right granularity at the right place in the right quality can be seen as the

general vision this research builds on without providing a problem definition for the domain or a proposed solution.

The relatively new but widely discussed topic of Big Data plays an important role in the current developments within the information and data management domain in manufacturing. The economically reasonable retrieval and usage of crucial insights from qualitatively diverse and versatile structured information, which are subject to constant change and which accumulate in large scale is defined as Big Data. The Big Data development is seen as a paradigm shift, as the importance of hard- and software diminishes, the importance of data as a value adding factor rises. The industrial domain is seen as one of the main benefactors of Big Data developments. In a digital world, Big Data is seen as a fourth production factor besides capital, labor and raw materials. The rapid increase in the amount of data is partly based on new developments in e.g., sensor technology, improved (mobile) communication and social media content. Big Data applications tackle an area where traditional approaches reach their limitations, basically to handle the sheer amount of information for decision making support.

Even so Big Data is widely used in recent times, this reflects a contrast to the fact that there is no commonly accepted general definition. One can argue that due to the rapid developments in data processing technologies, concrete numbers might not be useful within a definition. So the amount of data needed for an application to be considered Big Data is vaguely considered too big for traditional approaches to handle with acceptable effort. This is not the only defining factor of Big Data applications. The complexity of the to-be-analyzed data and the velocity of the processing are crucial (Küll, 2013).

Today a lot of sensor data is lost due to missing commonly accepted standards for data communication, processing and handling. Challenges are e.g., the large data volumes accrued by continuously recording sensor solutions. It not only the large amounts that propose challenges but also the rapid development in sensors and thus continued emergence of new data types which have to be handled (Lohr, 2012). Especially wireless sensor networks are prone to outliers due to various factors. As there are many different sensors active in these networks, failures can accumulate fast (Branch, Giannella, Szymanski, Wolff & Kargupta, 2013). This 'contaminated' data streams are a big challenge also for Big Data applications. Researchers look into various methods to identify and eliminate negative effects in sensor data, ranging from ML to Hopfield nets (Aggrawal, 2013).

In contrast to traditional data analysis methods, where the solution space is at least sketched, Big Data principles look at large amounts of data and try to identify new findings hidden in the data in real time. The approach used within the *product state concept* can be seen in between, however leaning towards traditional analysis paradigms. The goal of the *product state concept* is stated beforehand, as of identify-

ing a relevant set of state characteristics to support quality monitoring in manufacturing. However, as there are large knowledge gaps in regard to cause effect relations across manufacturing processes/operations. In order to define a set of relevant information for the manufacturing programme, all possible information artifacts have to be considered initially and the identification of cause effect relations, in this case applying pattern recognition shares similarities to Big Data principles.

Technically, the amount of information artifacts will most likely not be considered Big Data due to the comparable small amount. Thinking ahead, considering improvements in sensor technologies this can change in the near future. Furthermore, the interpretation of real time is different in Big Data applications, closer to milliseconds, than it is in the developed concept where real time is understood as 'available when needed'.

The importance of knowledge, information and data was already introduced as early as in the introduction of this dissertation and further detailed throughout this section by presenting existing domains and definitions. The different approach of IM, trying to gather relevant data for pre-defined problems and big data, looking at all available data in real time, trying to identify patterns in order to create new knowledge, is explained. Both approaches have an influence on the *product state concept* development. The goal of the *product state concept* is to identify a comprehensive set of relevant information to describe a product along the manufacturing programme. However, there are many unsolved issues and discrepancies between the available knowledge about the manufacturing processes and the needed knowledge. Therefore, Big Data principles of looking at available manufacturing data in order to identify patterns, which help in return to identify relevant information of the product and process, are included in the concept.

In the following sub-section, the topic of data and information quality is introduced as it plays an important role in all information based applications in manufacturing.

2.3.2 Data and information quality[6]

Data and information quality is a topic of great interest for many domains, be it social sciences, natural sciences or engineering. In manufacturing, especially in the area of process monitoring and control, data and information quality can play a decisive role in weather an analysis and the subsequent action is successful or not. As was stated previously within this section, information and data is not sharply distinguished in literature. From now on, to simplify the understanding, the term data quality will be used comprehensively, integrating both, information and data. In

[6] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest, Tinscher, Porzel & Thoben, 2014).

the following subsection, the current state of the art in data quality is presented from a research point of view.

Data quality is a multi-dimensional concept (Pipino, Lee & Wang, 2002). Data and information quality is usually defined in terms of contribution to the objectives of the end-user (Helfert, 2002). It can be additionally described as the adequacy for the relevant data processing application (Naumann, 2007). Poor data quality can be a major cause for damages and losses on organizational processes (Storey et al., 2012). To avoid the damages and losses data quality problems and solutions should be considered as early as possible, best at the design stage of the information system (Storey et al., 2012).

2.3.2.1 Data and information quality dimensions

Pipino et al. (2002) list under the data quality dimensions the following attributes: accessibility, appropriate amount of data, believability, completeness, concise representation, consistent representation, ease of manipulation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability, value-added. Data quality can thus be also described as a set of quality characteristics (Naumann, 2007). Many of the listed attributes contribute a lot to the overall data quality, as tested by a third-peer.

The starting point for consideration of data quality is the user-oriented quality concept. Helfert divides data quality in design and execution quality. The fulfillment of end-user requirements and specifications can be met through a choice of properties in the data design. Design quality refers as such to the collection of specific quality requirements from the user's perspective. Execution quality includes compliance with the specifications (Helfert, 2002). Helfert's basic data quality criteria are: correctness, completeness, consistency and timeliness (Helfert, 2002). Data quality criteria developed by English are: data standards, data definitions and information architecture (English, 1999). These criteria can be understood as the access capability, timeliness and interpretability of the data and the data system. Data quality as understood by Wang and Strong can be divided into internal data quality, contextual data quality, presentation, and access quality (Wang & Strong, 1996). Wang and Strong focus on user-related data quality - interpretability, usefulness, credibility, time reference, and availability have been rated as important criteria (Wang & Strong, 1996; Helfert, 2002). Jarke, Jeusfeld, Quix & Vassilidis's data quality criteria are: completeness, credibility, accuracy, consistency, and interpretability (Jarke et al., 1999). A poll conducted by Helfert delivers additional quality criteria important to organizations: clearly defined data descriptions, formal data syntax, delivery times (for data), and specific information about selected properties of the data, e.g., number of errors (Helfert, 2002).

Rohweder et al. describe data quality as the degree the characteristics suffice the requirements on the data product. The requirements for the data are determined

through particular decisions and goals set on data quality. Rohweder et al. define data quality with the help of 15 IQ (Information Quality) dimensions (Rohweder, Kasten, Malzahn, Piro & Schmid, 2011). These can be applied to e.g., master data, to assess if the data is useful or not acceptable. The IQ dimensions have been divided into four categories and form a regulatory concept for data quality (Rohweder et al., 2011). The 15 IQ dimensions are as follows:

System support (e.g., user interface)

- Accessibility: accessible & easy to access.
- Ease of manipulation: easy to use & to change.

Inherent (content examination)

- Reputation: data source & processing highly trustworthy.
- Free of error: error free and consistent with reality.
- Objectivity: strictly objective and value-free.
- Believability: reinforced with quality standards, etc.

Representation (overall presentation, e.g., the form of statistics)

- Understandability: ability of users to directly understand & use information.
- Concise representation: clear, saved in appropriate & understandable format.
- Consistent representation: uniform & held in consecutive & equal manner.
- Interpretability: understandable in same, technically correct manner.

Purpose-dependant (data use in the processes)

- Timeliness: actual properties of data (described) accurately & up-to-date.
- Value-added: usage leads to quantifiable increase in monetary cost function.
- Completeness: no missing information contained.
- Appropriate amount of data: amount meets requirements set on data.
- Relevancy: provides all necessary information for user.

Overall, it is accepted, that all IQ dimensions should exhibit a high or at least sufficient quality for an information system to be functional (Rohweder et al., 2011). Looking at the IQ dimensions, it can be stated that the *product state concept* can contribute to several of those. Especially the purpose-dependant dimensions, focusing on the data usage in processes, are reflected in the development. In section 4.1, features of the *product state concept* are mapped to these information quality dimensions (Table 1) according to how they address the issues.

2.3.2.2 Avoiding data errors

Data errors can be avoided most effectively and most sustainably at the moment of their emergence, e.g., throughout the manual data entry or automatic data collection (Naumann, 2007). A direct capture of the data from the source to an electronic device without human interference is the best way to minimize data input errors.

When human interface is unavoidable input errors may occur and consequently degrade data quality (Verma, 2012). To prevent quality degradation a quality check can be performed in the moment of data delivery (here data transformation to the target system). The data can be further checked by end-users through the use of complaints-forms or other rating-systems, e.g., statistical methods (Helfert, 2002). In case of data sets from external sources, it is essential for the researcher/information-manager to deal consciously with the data and the data quality; it is crucial to mark the problematic data to be able to deal consciously with it (Naumann, 2007). The external party and the person responsible for integrating the data into the target system should be clear about the purpose of why the data are being collected, and it should be clearly stated (Verma, 2012).

The most common data quality issues are incorrect or missing values, duplicates, and errors in the recording process (Helfert, 2002; Winkler, 2004; Naumann, 2007; Verma, 2012). Errors in data cause errors in reports generated from the data, thus reinforcing the "garbage-in-garbage-out-effect". Errors can be found within the schema and/or the data level. The schema level describes the errors in the structural, semantic and schematic heterogeneity of the data characteristics. The data level includes value-, unit-, accuracy-, and duplicates errors (Naumann, 2007).

Duplicates, one of the most costly data errors (Naumann, 2007) can arise, e.g., due to typographical errors in the unique identifiers (e.g., the name of the researcher). Missing identifiers and contradictions in data indicate low quality (Winkler, 2004). These issues can be prevented with data quality ensuring practices, e.g., marking of problematic data, auto correction of format errors, manual correction of the data values, troubleshooting and coordination with the data suppliers, and organizational rules (Helfert, 2002). Furthermore file-linkage can be used to create "more complete" data (Winkler, 2004). The traceability of data origin and documentation of discrepancies is also relevant (Helfert, 2002). Semantics and identifiability, as well as the precision of the value ranges, the granularity of data models, and the technical aspects of the data are less critical for the overall data quality (Helfert, 2002).

Data quality can be assessed by a third-peer. The assessment can either be task-independent, where no contextual knowledge is required, or task-dependent, with specific application context (Pipino et al., 2002). The data quality methodologies can be classified according to various criteria (Batini & Scannapieco, 2006):

- Data-driven vs. process-driven
- Measurement vs. improvement (assessment or improvement of data quality)
- General-purpose vs. specific-purpose
- Intra-organizational vs. inter-organizational

The previously presented basics and subsequently described relation to the conducted research and developed concept are underpinned by an elaboration on challenges of MS from an information and system perspective in the next subsection.

2.4 Challenges of MS from a product and process information perspective

In this section the challenges in the manufacturing domain with regard to the increasing importance of product and process information are derived. This provides a broad understanding of the research area and research problem in a wider sense this dissertation is based upon. In the following section 3, current concepts and approaches tackling these challenges to a certain extent will be presented. The resulting gaps between the challenges and how the current approaches tackle them further specifies the research problem.

The European Commission (EC) predicted the development of manufacturing along three paths[7]: (1) On-demand manufacturing; (2) Optimal (and sustainable) manufacturing and (3) Human-centric manufacturing (Filos, 2013). Especially the second path highlights that manufacturing has to be prepared to produce high quality products with high security and durability, competitively priced without avoidable waste and scrap (Filios, 2013). This focus on quality of individual products and efficient processes supports the arguments brought forth within this research.

There are several studies available proposing key challenges of manufacturing on a global level. The following key challenges most of researchers agree upon (Gordon & Sohal, 2001; Shiang & Nagaraj, 2011; Dingli, 2012; Thomas et al., 2012):

- Adoption of advanced manufacturing technologies
- Growing importance of manufacturing of high value added products
- Utilizing advanced knowledge, information management and AI systems
- Sustainable manufacturing (processes) and products
- Agile and flexible enterprise capabilities and supply chains
- Innovation in products, services and processes
- Close collaboration between industry & research to adopt new technologies
- New manufacturing management paradigms.

However, these key challenges highlight the ongoing trend of manufacturing operations growing complexity. This complexity is inherited not only in the manufacturing programmes but increasingly in the to-be-manufactured product itself as well as in the (business) processes of the companies (Wiendahl & Scholtissek, 1994). Adding to the challenge is the fact that the business environment of today's manufacturing companies is affected by uncertainty (Monostori, 2002).

Focusing from the global challenges towards the challenges of monitoring in manufacturing systems, the inherent complexity in manufacturing systems brings several challenges to the table when it comes to modeling and/or monitoring and con-

[7] www.actionplant-project.eu/public/documents/vision.pdf (retrieved Feb. 12, 2014)

trol approaches of manufacturing programmes. Some of the challenges new concepts have to be able to deal with are:

- the great number of different machining operations,
- multidimensional, non-linear, stochastic nature of machining,
- partially understood (cause-effect) relations between parameters,
- lack of reliable data,
- missing parts of data sets,
- high-dimensionality and multi-variate nature of data.

(Derived from: Tönshoff, Wulsberg, Kals, König & Van Luttervelt, 1988; Van Luttervelt, Childs, Jawahir, Klocke & Venuvinod, 1998; Monostori, 2002; Viharos, Monostori & Vincze, 2002; Kano & Nakagawa, 2008; Wuest et al., 2012b)

When trying to increase quality through a monitoring of manufacturing processes, it is tough to tackle the challenge of identifying problematic states throughout manufacturing processes by modeling cause-effect relations between product states as of these process intra- and inter-relations along the process chain due to this and other factors. The problem at hand has an inherent high complexity and high dimensionality (in this context high-dimensionality is understood as a multidimensional system with a large number of dimensions) (Suh, 2005; Lu & Suh, 2009; Elmaraghy, Elmaraghy, Tomiyama & Monostori, 2012). Optimization tools in this field need to be able to handle a large number of dimensions and variables in order to be useful in practice. Even so it would be desirable to use precise first-principle models, the development and application of such models is hindered by the complex nature and the above stated challenges of manufacturing programmes, especially when it comes to new manufacturing programmes, processes or operations (Kano & Nkagawa, 2008). The NP complete nature of the problem of identifying process intra- and inter-relations is described in more detail in section 0.

Kano & Nakagawa (2008) identified three functions that systems intended to improve product quality in manufacturing need to fulfill in order to be considered useful: "(1) to predict product quality from operating conditions, (2) to derive better operating conditions that can improve the product quality, and (3) to detect faults or malfunctions for preventing undesirable operation". Tönshoff et al. (1988), outlined already in the late 1980s the necessity of sensor integration, sophisticated models, multi-model systems and learning ability in monitoring and control of manufacturing programmes, especially machining processes. A possible clustering of concepts based on the kind of knowledge applied, leaves fundamental, heuristic and empirical models that can be distinguished (Viharos et al., 2002).

In the next section, existing approaches which focus on the identified challenges are discussed. The gaps between the successful tackling of the raised issues by these approaches provide a further basis for the developed *product state concept*.

3 Current approaches with a focus on holistic information management in manufacturing

The developments within the domain of manufacturing, intelligent and holonic manufacturing systems from an information and data perspective were presented in the previous section. This was concluded by a brief elaboration of the key challenges in that area as a basis for this section and the later development of the *product state concept*. In this section, the focus is laid on existing approaches and concepts that try to address some of the identified challenges of MS when it comes to transparent and product specific information and data management. The main focal methods and concepts are PDM, PLM and quality monitoring in manufacturing. The presented domain specific knowledge is discussed within this section as it has strong relations with the later concept development. In order to allow the reader to easily identify the relation of the individual method to the *product state concept*, a short conclusion after each section highlights the relevancy and connection to the topic. The final sub-section of this third section will furthermore briefly summarize the complete section and help the reader with the transition towards the next section where the *product state concept* is presented.

3.1 Product lifecycle management in manufacturing

PLM as mentioned in the previous chapters is focusing on the whole product lifecycle and promises to manage all data and information involved. This promise overlaps with the set goal of the developed concept, as it is based on information and data of an individual product over a whole manufacturing programme and beyond. PLM research has a long tradition not only in the engineering domain but also in management science. Strongly connected to PLM is the area of PDM, which is briefly discussed in the following subsection. After that, first the product lifecycle in manufacturing is investigated before taking a closer look at PLM in general and closed-loop, item level PLM in particular.

3.1.1 Product data management

Today there are many technologies available and widely used in industry. One of the first, Computer Aided Design (CAD) has developed greatly and has been supported by many other specialized tools like Computer Integrated Manufacturing (CIM), Computer-aided Engineering (CAE), Computer-aided Process Planning (CAPP), PDM and PLM systems (Chryssolouris, Mavrikios, Papakostas, Mourtzis, Michalos & Georgoulias, 2009). Other IM systems focus more on operations like Manufacturing Resource Planning (MRP), Manufacturing Execution Systems (MES), Advanced Production Systems (APS) and Enterprise Resource Planning (ERP) (Wiers, 2002). Taking a closer look, MES, widely districubted in industry, are software packages designed to manage factory floor material control and labor and machine capacity (Helo et al., 2014) in real time. They are usually located at

the factory level (Brecher et al., 2013), and are an integration of the management system and systems nearer of the shop floor operations (Simao et al., 2006). MES can be described as control systems with the goal to fill the gap between the upper planning level and the lower shop-floor execution. MES aim to control the production, maximize the workload of equipment, release unneeded machine tools, track and trace components and orders, manage inventory, and optimize production activities from order launch to finished goods. They are usually linked with ERP systems which issue, e.g., production orders to the MES system, linking quality control, scheduling and material information (Helo et al., 2014). The latest developments in MES include building flexible workflows and supporting distributed manufacturing (Helo et al., 2014). A MES can as such be understood as the operational arm of the ERP system. It implements the ERP's production plan and reports the current processing status back to the ERP level. The MES system monitors the local production lines and gathers data regarding the logistics and the technical parameters in the production process. In addition to monitoring the production, and materials status, it also provides the execution and construction plans of the production orders. The overall goal is the improvement of productivity and reduction of cycle-time. Overall, MES and related systems do not focus on individual product and process information.

However, this research is looking mainly on product and process data and information. Therefore, as focusing mainly on the product and a product-centric perspective, PDM stands out as it focuses on data and information directly connected to the products. PDM systems were developed to support CAD systems through management of the CAD-data and drawings. Since then the domain continuously developed further towards today's integrated solution for the management of product and process data between various systems. However, over time a large amount of terms were created (e.g., Digital Product Definition (DPD), collaborative Product Definition Management (cPDm)), enhancing the original meaning of PDM but also creating some confusion among practitioners (Abramovici & Sieg, 2001).

Literature on PDM promises integration and management of all information that defines a product (Liu & Xu, 2001). However, it is mostly seen as a tool to store, administrate and share product data, not to decide what information should be stored or determine how the stored information is connected to each other. Other researchers extent the view further. Saaksvuori & Immonen (2004) describe product data as "information broadly related to a product". PDM is one of the major focus areas of engineering and manufacturing companies (Fasoli, Terzi, Jantunen, Kortelainen, Sääski & Salonen, 2011). Fasoli et al. (2011) claim, that especially within today's distributed production processes, it is most important that data is first correct and second correctly distributed. It is essential that data is correct and in a format that it can be transferred to all addresses in need electronically (Saaksvuori & Immonen, 2004; Gimenez, Vegetti, Leone & Henning, 2008). A system realizing this in an applicable way for industry practitioners has not yet

been developed (Abramovici, 2007) and the "requirements for efficient management of product data have been steadily increasing" (Leong, Yu & Lee, 2002). Other researchers looked into how product data manipulated by a manufacturing process can be integrated into a PDM system (Peltonen, Pitkänen & Sulonen, 1996).

Within the PDM field, certain standards have evolved, tackling the interface issue and communication issue between different systems. Communicating advanced information about a product through current PDM standards like STEP (Standard for the Exchange of Product Model Data) are mainly focused on geometric information and does not explicitly support information like chemical composition of material. Combined with the fact that products become more complicated (increasing number of parts and variations) (Leong et al., 2002) this highlights the need for innovative concepts for structuring and handling product related data efficiently.

Looking at the claim of PDM systems that they can integrate and manage all application, information and processes that define a product (Chryssolouris et al., 2009), there are many challenges in data management in the manufacturing domain still to be faced (Fasoli et al., 2011). However, PDM systems today are used by most manufacturing companies for e.g., "controlling information, files, documents, and work processes and are required to design, build, support, distribute, and maintain products" (Chryssolouris et al., 2009). Chryssolouris et al. (2009) define typical product related information managed by PDM data as: "geometry, engineering drawings, project plans, part files, assembly diagrams, product specifications, numerical control machine-tool programs, analysis results, correspondence, bill of material, and engineering change orders among others".

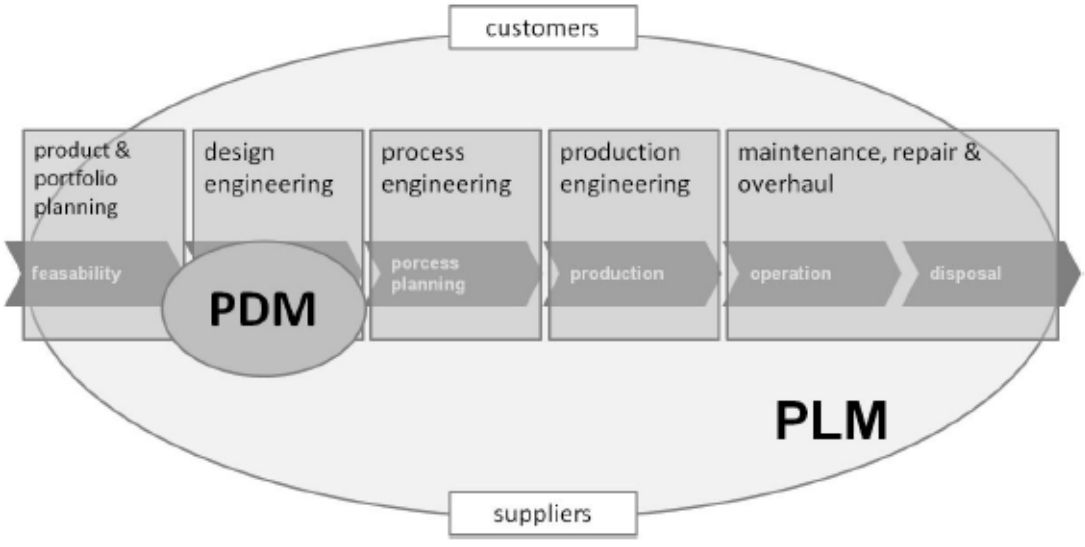

Figure 16: Distinction of PDM and PLM along the value chain (Paul & Paul, 2008)

In Figure 16 the focus area of PDM is highlighted along the value chain. In contrast to the very narrow focus of PDM on design engineering, PLM is shown as a

more overarching and thus more applicable from a systemic perspective on manufacturing. Furthermore, as PDM is looking mainly on a product class (e.g., tire model abc) the resulting extension of PDM to PLM and from that to closed-loop and item-level PLM (see section 3.1.3) can be seen as a step towards a holistic IM of individual products (e.g., No. xyz of tire model abc). This is presented in the following subsection, starting with the product lifecycle itself.

3.1.2 Product lifecycle management[8]

Product lifecycle literature generally differentiates organizational/marketing and production engineering/ICT perspectives (Sundin, 2009). In marketing, practitioners and academics tend to adopt a sales-oriented view, dividing the lifecycle into five phases: introduction, growth, maturity, saturation and degeneration of a product. Here, the economic success of a product is the main concern of classification (Meffert, Burmann & Kirchgeorg, 2008). The scope a product refers to may be a model, type or category.

The engineering and ICT perspective used here follows (Kiritsis, Bufardi & Xirouchakis, 2003) (see Figure 17). The basic product lifecycle framework in production engineering differentiates three main phases (Jun et al., 2007; Cao, & Folan, 2012), describing the product from the "cradle to grave" (Stark, 2011):

- ***Beginning-of-Life (BOL):*** processes related to development, production & distribution
- ***Middle-of-Life (MOL):*** processes related to a product's use, service & repair
- ***End-of-Life (EOL):*** processes related to reverse logistics like reuse, recycle & disposal

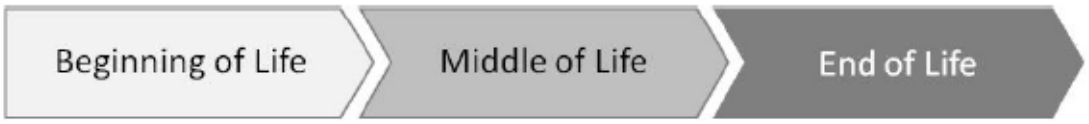

Figure 17: Phases of the product lifecycle

Being shortly introduced in the PDM section before, PLM extends the concept of PDM beyond the usage in product design and partly manufacturing (Paul & Paul, 2008). Some researchers use the terms almost interchangeable whereas others clearly state that PLM is a central approach of an integrated management data related to products but also manufacturing processes and beyond (Fasoli et al., 2011). Classic PDM functionality encompasses object, component and document management, classification and search functionality, change management and tools for system administration and configuration (Abramovici & Sieg, 2001). The production engineering and ICT perspective towards PLM also differs from the organiza-

[8] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest, Hribernik & Thoben, 2014a)

tional and marketing perspective, as does the view on the product lifecycle itself (see above). In production engineering and ICT, PLM is commonly understood as a concept which *"seeks to extent the reach of PDM [...] beyond design and manufacturing into other areas like marketing, sale and after sale service, and at the same time addresses all the stakeholders of the product throughout its lifecycle"* (Golovatchev & Budde, 2007). PLM consequently includes strategically modeling, capturing, exchanging and using information in all decision-making processes throughout the product lifecycle (Stark, 2011; Moorthy & Vivekanand, 2007). It implements an integrated, cooperative and collaborative management of product data along the entire product lifecycle (Terzi et al., 2007).

By definition, every product has a lifecycle. Manufacturers are becoming aware of the benefits inherent in managing those lifecycles (Sendler, 2009). At the same time today's products are becoming increasingly complicated. For example, the amount of component parts is increasing. Simultaneously, development, manufacturing and usage cycles are accelerating (Sendler, 2009) and production is being distributed geographically (Seifert, 2007). These trends highlight the need for innovative concepts for structuring and handling product related information efficiently throughout the entire lifecycle. On top that, customer demand for more customization and variation stresses the need for a PLM at item, not merely type-level.

Besides merely handling product and process related data, PLM also has to take into account the interdependencies of information and communication between all of the stakeholders involved in the product lifecycle. Common graphical representations of the product lifecycle encompass three phases, beginning of life, middle of life and end of life (see Figure 17). Recent research clusters available PLM methods/tools in three major groups (Gimenez et al., 2008; Fasoli et al., 2011): information (e.g., focus on identification methods) and process management (e.g., operational activities) as well as application integration (e.g., definition of interfaces).

3.1.3 Closed-loop and item-level PLM[9]

Conventional views of PLM tend to stress the first phase of the product lifecycle, due to its beginnings in PDM and CAD. Processes highlighted here are product design, development, production and sales. Emerging approaches such as closed-loop PLM (Jun et al., 2007) take a holistic view upon the entire product lifecycle, from product ideation to end-of-life processes, ideally also the end of one lifecycle into the beginning of the next. It thus puts forward a paradigm shift from 'cradle to grave' to 'cradle to cradle' (Pokharel & Mutha, 2009) (see Figure 18).

[9] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest, Klein, Seifert & Thoben, 2011a; Wuest, Werthmann & Thoben, 2013c; Wuest et al., 2014a; Wuest, Liu, Lu & Thoben, 2014c)

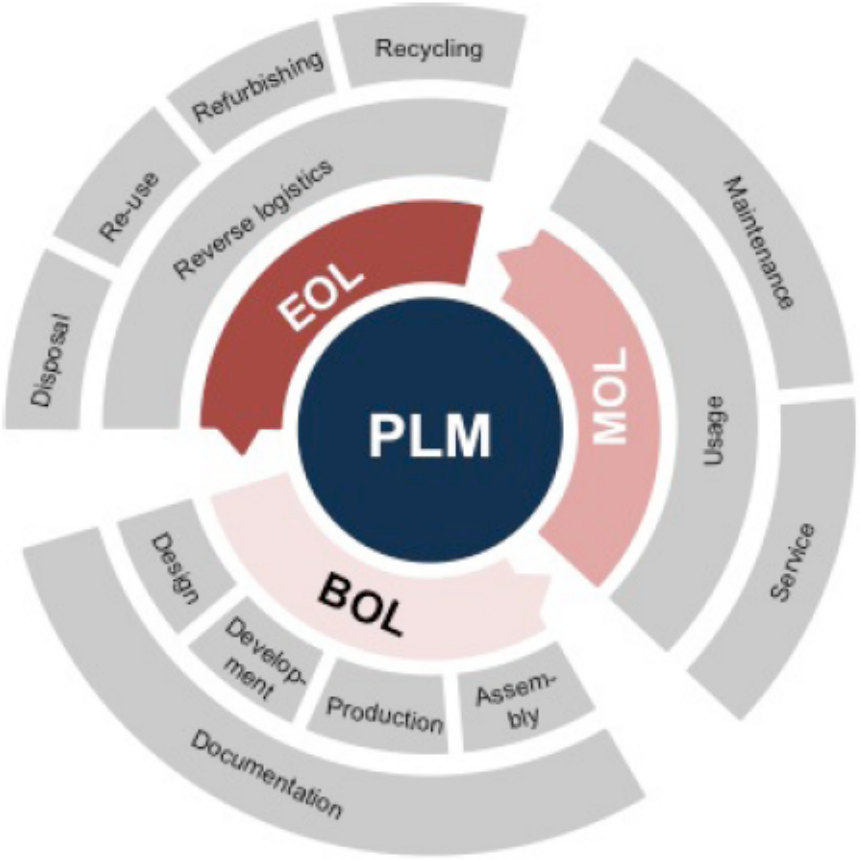

Figure 18: Closed-loop, item-level product lifecycle phases (Wuest et al., 2014a)

An example is the refurbishment of components from decommissioned products for use in new ones. The aim of closed-loop PLM is to close information gaps between the different phases and processes of the product lifecycle of individual products. This can be backwards, for example providing usage data to design processes, or forwards, for example providing production and assembly information to recycling processes. It deals with products not as classes or variants, but as individual items ("item-level"). Additionally, item-level PLM allows focusing more on the individual product and therefore creates the basis to e.g., monitor product quality on an individual level rather than on a batch level.

Being in the focus of researchers and practitioners for over a decade, data management in PLM still has to face the challenge of a gap between the existing reality and the expected features in data management (Fasoli et al., 2011). Especially in the area of item-level product data management along complex manufacturing chains are still many issues, like e.g., a generally accepted and interchangeable format that contains all relevant information, to be solved. The developed concept contributes to this development in the abstract way of systemizing product information around the product state.

The *product state concept* may be argued to be a sub-domain of item-level PLM or an extension of existing approaches. Both concepts are looking at individual prod-

ucts over different phases or processes. The *product state concept* incorporates many principles of item-level PLM and faces similar challenges, however the *product state concept* focuses mainly on the manufacturing phase as of today. For future application it is important to understand the similarities as it might present an interesting option to include the *product state concept* as a module in existing PLM tools. Furthermore, the *product state concept* highlights the need of understanding the process intra and inter relations between states whereas current PLM solutions do not look into this issue in detail.

Within the perspective of item-level PLM, the focus on the individual product, e.g., for optimization purposes is evident. It can be said that traceability is the basis for item-level PLM (Terzi et al., 2007). The *product state concept*, based on the principles of item-level PLM in manufacturing, has a basic requirement of traceability of individual products throughout the manufacturing programme. This is essential in order to derive state information and data at the checkpoints to take appropriate actions, e.g., adjust parameters accordingly. Tracking and tracing of individual products and batches of products is a well-established research field. In SCM research, there are already very advanced solutions available and applied in industry (Hribernik, Pille, Jeken, Thoben, Windt & Busse, 2010; Musa & Gunasekaran & Yusuf, 2013). The logistic chain in the food industry is an example where advanced tracking and tracing solutions are already applied (Van Dorp, 2002; Jansen-Vullers et al., 2003; Stark, 2011). For the before mentioned item-level PLM it is essential to trace individual products throughout the lifecycle.

The topic of tracking and tracing increasingly gains attention in the manufacturing domain highlighting also the importance of processes (e.g., Terzi et al., 2007; Brinkheinrich, 2008; Zhang, Jiang, Huang, Qu, Zhou & Hong, 2010). The importance of continuous tracking and tracing of a product throughout the whole manufacturing process for manufacturing companies as well as participating stakeholders is widely accepted. This creates the basis for advanced information management and quality improvement and assurance (Van Dorp, 2002). Being well established in the logistics domain, tracking and tracing in manufacturing faces different challenges which will be illustrated in the following paragraphs.

Jansen-Vullers et al. (2003) distinguish different types of traceability depending on their usage, whereas "tracking" was described as "method of following an object through the supply chain and registering any data considered of any historic or monitoring relevance." Looking at their reference model for traceability Jansen-Vullers et al. (2003) highlight the importance of the ability to distinguish between instances of products as a prerequisite of traceability.

Within this research the definition of Terzi et al. (2007) builds the basis for the understanding of product traceability in manufacturing. The definition is as follows: "Generally, product traceability is the ability of a user (manufacturer, supplier,

vendor, etc.) to trace a product through its processing procedures. Concretely, the product traceability deals with maintaining information records of all materials and parts along a defined lifecycle (e.g., from raw material purchasing to finished goods selling) using a coding identification. Product traceability is by definition a PLM topic, since it is related to a product centric approach, where product data and information might be retrieved and managed along the whole lifecycle."

There are numerous benefits that result from implementing a functional tracking and tracing system. A general benefit is that companies are able to combine important information with individual products. In the production of small batches or even single products this offers the advantage of always being capable of checking what manufacturing processes the product already passed and what parameters were used. This can be the basis for an in-process adjustment of parameters based on the product state before each process step to increase quality in terms of reducing scrap and rework. Other benefits include, for example, proof for demanding customers, efficient PLM and feedback in case of product failure.

Overall tracking and tracing in manufacturing companies is different from other industries like food or areas that utilize tracking and tracing for mainly logistics purposes including regulatory purposes. There are more things to consider as the product itself goes through extreme conditions and can even changes consistency and shape during the manufacturing processes. This is not only a challenge for the physical marking of products but also for capturing information. The marking goes as far as possible at certain situations to mark the product directly with a tag or a code but it still must mark it indirectly through accompanying documentation. The information layer must also be capable of handling the information and processing it to make it practical and finally beneficial.

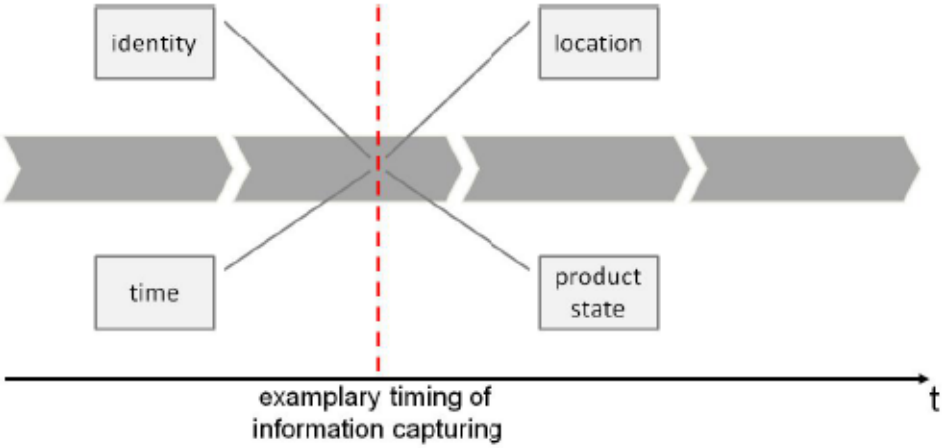

Figure 19: Elements of information captured (Wuest et al., 2013c)

As mentioned above, an important aspect of tracking and tracing in manufacturing is that not only time and location have to be considered but also product state in-

formation has to be captured. Figure 19 depicts the critical elements of information capturing in manufacturing.

It is always necessary to link the captured information to a specific object (***identify***). Therefore, the object has to be identified precisely and uniquely. The identification can take place automatically by e.g., scanning a barcode or a RFID transponder or by entering the information manually into an IT system. Another critical element of information captured is ***time***. A time stamp integrated into every event captured is necessary for having unique information. Moreover, the time stamp is necessary to have a precise history of every object being tracked within the supply chain. Knowing about the ***location*** of an object is also very important when generating an event as information of the current process may be derived based on location/time. Last but not least, the ***product state***, which incorporates various characteristics of a product e.g., quality and/or dimensions of an object, is considered relevant information. Based on the product state's characteristics, the following process steps and their parameters within supply chains can be planned. An example for a state characteristic is the diameter after machining, but also residual stress allocation within a steel disc (Wuest et al., 2013c) (see section 4.2). In this context the question of the time horizon of information capturing comes up. As stated before, the information and data has to be captured in real time, which is understood within this work as available when needed.

3.2 Quality monitoring in manufacturing

Quality, as discussed before is of major influence in the manufacturing domain. The term itself, for both product and processes was introduced before. In this subsection, the existing applications in quality monitoring in manufacturing are presented. The *product state concept*, developed in the following section, may be understood as part of or extension of a quality monitoring system.

3.2.1 Quality management in the manufacturing domain

QM is widely used and the term is understood slightly different depending on the domain (Steffelbauer-Meuche, 2004). Manufacturing companies have been focusing on improving the quality of their products and processes in a structured way for the last few decades (Robinson & Malhotra, 2005). Research on this topic can be summarized within the term QM. In this research, suitable for manufacturing domain, QM is understood as the entity of all quality related actions and goals. Within the framework of QM, certain measures are taken which are supposed to improve products and processes of a manufacturing company. In order to achieve these goals, quality standards are defined, which are organized in a QM handbook and which have to be met. In order to control the efforts, specific persons in charge have to be defined which are responsible for documenting and communicating possible deviations, develop improvements and monitor the implementation (Corsten & Gössinger, 2008).

	verification of results	process excellence	quality management
	quality control →	quality assurance →	total quality management
goal	• zero defect production	• zero defect production • process excellence	• zero defect production • process excellence
action	• measuring technology • statistical methods	• zero defect philosophy • certification according to DIN ISO 9000 ff.	• customer satisfaction • employee involvement • organisational structure
	1950	1970	1990

Figure 20: Development of quality management (Wannenwetsch, 2010)

In the manufacturing domain, QM principles are applied in industry since the early 1970s (Robinson & Malhotra, 2005). In academia, many well-known researchers are continuously working on this topic for the last decades (Hoyer & Hoyer, 2001). The previous Figure 20 illustrates an overview of the development of QM in the manufacturing domain. The development describes the way QM took, from a very technical perspective towards a more customer oriented approach (Wannenwetsch, 2010). According to DIN EN ISO 8402, QM includes all management activities which define within the QM process the quality politics, goals and responsibilities as well as the measures like quality planning, control, assurance and improvement necessary to realize the former (Wannenwetsch, 2010).

QM has a long history in both industrial application and academic research as elaborated above. That has led to a wide variation of available concepts, methods and tools for companies to increase process, product and documentation quality like Failure Mode and Effects Analysis (FMEA) to identify potential failure within a system (Tietjen, Decker & Müller, 2011) and Total Quality Management (TQM) to continuously improve products and processes (Forza & Filippini, 1998) are just two out of many available. Köksal et al. (2011) e.g., list "inspection (100%), Statistical Quality Control (SQC), Total Quality Control (TQC), zero defects, […] kaizen, ISO 9000 quality standards, quality award programs (Malcolm Baldrige, European Quality Award and so on), 6r, DFSS, lean six sigma have been among the most recognized ones (Fasser & Brettner, 2002; Montgomery, 2005)" (Köksal et al., 2011) as some of the most frequently used QM tools. Overall, most QM methods and tools are adjustable to various environments within the manufacturing industry and can be combined in order to realize the wanted outcome. Overarching approaches like Computer Aided Quality (CAQ) combine several of the quality philosophies with software tools in order to enhance the impact, support the practitioners without confusing them and create a companywide standard.

However, due to the diversified nature of manufacturing, manufacturing programmes, requirements and quality problems, etc. there is an almost endless variation of methods, tools and techniques available which can confuse practitioners. Partly due to the confusion and the lack of clear communication what quality means and which method, tool or technique is suitable, many of those quality initiatives produced mixed results often failing to reach the quality goal (Samson & Terziovski, 1999; Kaynak, 2003; Robinson & Malhotra, 2005; Sitek, 2012).

3.2.2 Quality monitoring in manufacturing programmes[10]

Quality monitoring has strong ties and overlaps with process monitoring (section 2.1.2) and QM in general. Certain tools like TQC and zero defects have an incremental need for quality monitoring in order to be employed effectively (Nebl, 2007). However, quality monitoring can focus on different areas, e.g., product quality and process quality. Again the ties to process monitoring are showing as e.g., machine health monitoring has an impact on quality monitoring of the process and thus of the product itself. Quality monitoring checks if the quality of a product or process is within the accepted range at certain checkpoints (Nebl, 2007), which is the basis for successful quality control in manufacturing. Köksal et al. (2011) state that the "process industries and discrete parts manufacturing industries have had a long history of these [quality monitoring] activities that aim to reduce variability. While quality monitoring tries to reduce variability by detection and removal of assignable causes, process control is based on the idea of process compensation and regulation to reduce variability."

Whereas, quality and condition monitoring is already well established and to some part successfully implemented for monitoring only one manufacturing process/ operation at a time (e.g., Silva, 2009; Jenab & Ahi, 2010), concepts taking the importance of the system view, monitoring of the whole manufacturing programme, into account are still rare. Additionally, some like Ding et al. (2002) recognize the importance of the system view for monitoring but focus on a specific characteristic, in that case diagnosing fixture faults. Other research, also taking the whole manufacturing programme into account, focuses on the identification of the critical manufacturing process causing a deviation from the planed characteristics (e.g., Zantek et al., 2006). Jiang et al. (2012) and Sukchotrat, Kim & Tsung (2009) are both presenting novel approaches tackling similar issues as this research to monitor multistage manufacturing programmes using either error propagation networks (Jiang et al., 2012) or multivariate control charts (Sukchotrat et al., 2009) on a still conceptual level with further research ongoing. Quality monitoring is increasingly using modern AI and PR methods to improve the results. Among the used AI and PR tools and methods are: Artificial Neural Networks (ANN), Principal Compo-

[10] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2013b).

nent Analysis (PCA), Partial Least Squares (PLS), SVM, and Decision Trees (DTs), etc. (e.g., Zorriassatine & Tannock, 1998; Hussain, 1999; Ganesan, Das & Venkataraman, 2004; Köksal et al., 2011).

Common quality issues occurring in manufacturing programmes and a selection of solutions proposed by literature (see Annex Table 11) show that the more detailed the manufacturing issues become, the more general are the proposed solutions. For specific and detailed issues there are many targeted solutions available, which are described in several case studies or implemented in industrial manufacturing programmes. However, there is an overall emphasis on information and data when it comes to (technical) manufacturing quality problems and how to tackle the issues.

In this section the other constant besides information and data is highlighted: quality. After the terms and definitions were introduced in the previous section, the important domain of quality monitoring is presented. Quality monitoring is expanding on the already introduced process monitoring as a framework on the *product state concept* and how successive states' inherited set of relevant information can be utilized. In order to understand industrial needs and challenges concerning quality in manufacturing, a selection of quality issues is presented in a table (see Annex Table 11). It can be concluded, that information and data are considered important for tackling manufacturing quality issues. Today, the focus area of most QM and quality monitoring tools is still very specific in contrast to the holistic nature of many manufacturing programmes. This is reflected directly in the *product state concept*'s holistic design as it is an information-based system with the goal of improving manufacturing quality.

In the next section the limitations of existing approaches for holistic information management in manufacturing systems based on individual products are facing today is presented, highlighting aspects to be tackled by the developed *product state concept* which is described in later sections.

3.3 Limitations of current approaches for holistic information management in manufacturing systems

In this section existing approaches to tackle the key challenges of modern MS are presented. The main concepts and methods introduced that focus on information management in manufacturing are PDM, PLM, QM and quality monitoring. After each subsection a short conclusion was presented that discussed the gaps of the specific approach when it comes to address the previous challenges and goals. Furthermore, the relation to the later *product state concept* and overlaps are shown.

PDM and PLM are both concepts that gained increasing attention over the last years. Especially PLM moved slowly away from being a mainly design focused tool towards a more holistic approach considering other phases of the lifecycle. However, being in the focus of researchers and practitioners for over a decade, data

management in PLM still has to face the challenge of a gap between the existing reality and the expected features in data management (Fasoli et al., 2011). Especially in the area of item-level product data management along complex manufacturing chains are still many issues, such as a generally accepted and interchangeable format that contains all relevant information, to be solved. The developed concept contributes to this development in the abstract way of systemizing product information around the product state.

The later developed *product state concept* may be argued to be a sub-domain of item-level PLM or an extension of existing approaches as both concepts are looking at individual products over different phases or processes. The *product state concept* incorporates many principles of item-level PLM and faces similar challenges, however the *product state concept* highlights the need of understanding the process intra and inter relations between states and actively includes the means to identify those. Current PLM solutions do not take this issue into account.

Recalling how QM and quality monitoring addresses the key challenges of holistic information management identified before, the overall more specific nature of the methods and concepts surfaces. In QM, due to the diversified nature of manufacturing, manufacturing programmes, requirements and quality problems, etc. there is an almost endless variation of methods, tools and techniques available which can confuse practitioners. However, the various tools and methods can be successfully used in combination, but still present no conclusive approach in a sense of a holistic concept for the whole manufacturing programme as of now. Quality monitoring on the other hand, even so specific approaches exist in the dozens has to be seen more as a philosophy than a method or approach. In this sense, the developed *product state concept* may be seen as a way to incorporate a holistic quality monitoring approach in manufacturing.

Concluding, the diversified challenges modern complex manufacturing operations have to face in order to improve their process and product quality is just partly addressed today by the above presented approaches. Most tools target a very specific area and have to be used in combination to effectively tackle the key challenges identified before. Especially the increasing complexity of large scale high-dimensional and multivariate product and process data involved in high-tech manufacturing and unknown cause-effect relations along the manufacturing programme highlight the need for supporting concepts helping the companies to cope with their product information needs. Such a concept has to be able to reduce the inherent complexity of today's manufacturing operations and provide relevant information about the individual product and process along the (manufacturing) lifecycle. In this case the term relevant means relevant to all stakeholders involved in the manufacturing programme and not just focusing on independent single manufacturing process or operation. The *product state concept* which will be intro-

duced in the next chapter is based on a set of relevant information about an individual product and process which can be used as a basis for newly developed systems for product quality improvement according to the principles by Kano & Nakagawa (2008). The main challenge within this concept lies in the identification of such a set of relevant product state information.

In the next section, first the key findings of section 2 & 3 will be picked up in a detailed argumentation and to set the boundaries for the developed *product state concept*. A key point will be the question how to identify process intra- and interrelations between states especially under the existing knowledge gap which will also be shown in detail in the next section 4.

4 Development of the product state concept

In this section, the *product state concept* and its development will be illustrated from a theoretical perspective. The main intension is to provide a general understanding of the goals and basic pillars of the concept and its argumentation. Another major goal of this section is to discuss and present the challenges and limitations to the application of the presented theoretical approach in practice. This outcome is crucial for the selection of appropriate methods and the following approach to identify state drivers despite the knowledge gap concerning process intra- and inter-relations using ML which will bring the *product state concept* to life.

In this section, first, the rationale for developing the *product state concept* is presented based on the previously presented state of the art of intelligent manufacturing systems. The argumentation incorporates the identified challenges and requirements of modern manufacturing programmes towards information and data based approaches. The term product state is then defined for the manufacturing domain before the focus shifts onto product state characteristics and the challenge of how to distinguish which ones are relevant for a specific product and process.

In order to create a comprehensive concept, a complementary approach of applying a method to identify quality checkpoints in manufacturing, adapted from the stage gate model (Cooper, 2010) briefly introduced (for details refer to Wuest et al., 2014c). This is essential as determining the checkpoints' influences on the complexity of the system to be analyzed and monitored. Determining more than the necessary checkpoints means that more process intra- and inter-relations and data points have to be taken into consideration and this consequently may influence the ease of applicability of the *product state concept*. Process intra- and inter-relations between states and state characteristics along the manufacturing programme are discussed in detail, as they are considered fundamental within the concept for the identification of a set of relevant state characteristics. In this section process intra- and inter-relations of state characteristics in manufacturing programmes are not only described but also the limitations towards describing and illustrating them are derived. Based on these important principles and definition, a possible visualization of the *product state concept* is introduced, based on bipartite graphs as well as Unified Modeling Language (UML) and Business Process Modeling Notation (BPMN) principles. In this section the challenges arising with the so far presented approach and the to a large part theoretical discussion are already visible and it seems questionable how the originally stated goal of this work can be reached.

Following, the challenges and limitations of an approach based largely on process intra- and inter-relations are described, building a basis for the next step, the presentation of requirements towards supporting tools and methods for the identification of product state drivers. A brief analysis of the NP complete nature of the problem at hand paves the way towards ML techniques, which have a proven rec-

ord of handling such issues rather well (e.g., Yang & Trewn, 2004; Harding et al., 2006). By identifying the requirements towards state drivers identification, the suitability of modern ML approaches is discussed further. This leads over to the derived specific research hypotheses and approach to identify product state drivers of a manufacturing programme using SVM algorithm based feature selection presented in the following section 5.

4.1 Rationale for the product state concept[11]

In the previous sections different domains' importance for today's MSs and their basic paradigms were discussed. In this section the rationale for the *product state concept* development is presented. The basic paradigms the argumentation follows are summarized in Figure 21.

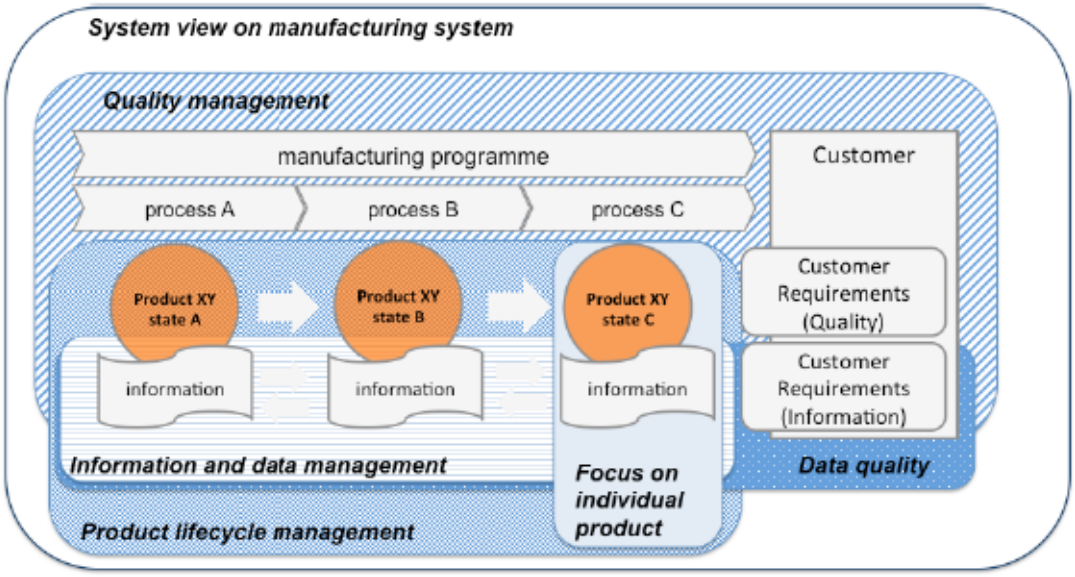

Figure 21: Basic paradigms of the development of the product state concept

It has been established that a successive incremental improvement approach in manufacturing can focus on an individual processes or operations or on a manufacturing programme as a whole (see section 2.2). It has to be noted that this is not a 'black or white' differentiation and there are overlapping areas, e.g., approaches looking into the whole manufacturing programme whilst at the same time focusing on selected processes. Both improvement approaches have their justification. A focused approach on individual manufacturing processes can bring forth significant improvements in different areas like e.g., machining performance. New research results indicate the importance of cross-process relations and their influence on product quality (Zoch & Lübben, 2010). Looking at a manufacturing programme

[11] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2014b)

as a system with different components, e.g., processes and operations, not being independent form each other and thus on the contrary, influence each other's outcomes (quality) can contribute to improvements. For example is it possible to induce reasons for quality variations during an early process of a manufacturing programme, which are triggered during a later stage (Zoch, 2012; Surm & Rath, 2013). An example is internal stress allocations, induced during the operation 'clamping' in the machining process of a manufacturing programme, which may have no significant effect until the final heat treatment process, where they may influence the process and lead to distortion of the product (Sölter & Brinksmeier, 2008; Sölter, 2010; Surm, 2011). Increasing the transparency and ability to support the understanding of such relations may support quality initiatives. However, the complexity of a manufacturing programme increases rapidly with its number of processes and operations, and this represents a challenge newly developed concepts and approaches have to deal with.

The development of the *product state concept* is based on the 'system view' and tries to contribute to generating increased understanding of co-relations within a manufacturing programme. It furthermore may contribute to the further development by reducing the complexity by supporting the identification of relevant information within this system. This way information not regarded as relevant does not have to be processed and the dimensionality can be reduced. The relations between the different processes or product states, which contribute to the complexity, are actively taken into regard within the development of the *product state concept*. In this scenario, all available product and process data is used to identify relations, which accordingly may indicate relevance to certain state characteristics. This tactic utilizes basic Big Data principles combined with PR technologies to support the IM principle of just focusing on relevant information.

The second principle of the *product state concept* development is the focus on individual products instead of looking on product groups or families. This focus on individual products takes into account the high quality requirements towards highly stressed products. Variations between similar products can be the reason for quality problems during the manufacturing programme, and a resultant failure to comply with the final customers' quality requirements. In order to respond to individual variations, which cannot be entirely avoided (Kaiser, 1998), clearly predicates that a focus on individual products is necessary. This complies with the view taken in item-level PLM where the focus is on each individual product and its inherited information and data. The development of the *product state concept* adopts the item-level PLM perspective of looking at an individual product over the whole lifecycle, with the slight adaption of exchanging the focus domain from the whole lifecycle to the manufacturing programme. What makes the *product state concept* differ from existing approaches is that the individual item is not an identity attached to a raw material, work pieces and components, but from the start of the

value adding processes of a manufacturing programme until the final stage, the individual item is considered being the same just changing its state.

Incorporating this focus on individual products throughout the manufacturing programme leads to questions about tracking and tracing. This question, even though crucial for industrial application will not be the focus of this research as there are already existing solutions available to track and trace individual products throughout a manufacturing programme. However, in this context the question of determining the checkpoints for information capturing needs to be addressed. Choosing the right time during the manufacturing programme is just as important as the question what information needs to be captured and often interrelated. In order to create a comprehensive concept, an approach of how to determine suitable checkpoints has been developed and is integrated within the *product state concept*. This approach, based on adapting the stage gate model and its principles from product development to manufacturing is described in detail in (Wuest et al., 2014c).

In the previous sections the importance of information and data for quality improvements in manufacturing was highlighted. The development of the *product state concept* focuses directly on the information and data layer and the individual product's data and information connection to corresponding processes. Additionally, the concept requires that information and data will be presented in a universally accepted manner (e.g., standardized formatting) to address upcoming interface issues. As the information is mainly of descriptive nature, describing product state characteristics and process parameters, exiting standards (e.g., STEP) may be supported and information may be stored in Comma-Separated Values (CSV) format.

Quality is one of the main focal points of improvement initiatives in manufacturing. According to the definition of quality the product has to fulfill the customer requirements. This in turn is the basis for fitness for purpose as a measure of product quality. This is reflected throughout the *product state concept* development as the identification of the set of relevant information is based on quality considerations. Furthermore, the concept supports the quality idea for all types of customers, internal or external, e.g., in a collaborative network or internal, e.g., another business unit or another process. In summary, by contributing to the understanding of the mechanisms within a manufacturing programme and the increase of transparency, the *product state concept* may support product and process quality improvement of manufacturing programmes.

Related to the information and data management issues that have to be considered during the development of the *product state concept*, section 2.3.2.1 presented information quality dimensions. The focus lies on the purpose-dependent dimensions as they represent the main contribution of the *product state concept* towards information quality in manufacturing. These dimensions are related to customer requirements combined with a system support perspective. The five purpose-

dependent IQ dimensions (Rohweder et al., 2011) and how they are addressed by the *product state concept* development, is summarized in Table 1. A table with all 15 IQ dimensions and the connection to the product state concept is illustrated in Table 10 (Annex). Overall it can be stated that a product focused concept needs an integrated management of all information connected to the product (Garetti & Terzi, 2004; Taisch et al., 2011).

Table 1: Purpose-dependent IQ dimensions and their influence on the product state concept

Purpose-dependent IQ dimension	Addressed by product state concept
Timeliness	The product state information and data properties are accurately stored and uniquely identifiable to an individual product through the checkpoint system and mapping.
Value-added	The goal of the *product state concept* is to derive new knowledge and support the increase of transparency through the manufacturing chain in order to support process and product quality improvements.
Completeness	The product state information and data should be stored as complete as possible within the set of relevant information. However, this depends also on the external circumstances like sensors, etc.
Appropriate amount of data	The *product state concept*s main objective is to identify a set of relevant information in order to reduce the amount of information and data to be handled.
Relevancy	The *product state concept*s main objective is to identify the set of relevant information. This contributes to ensure that the data and information captured is relevant for the chosen purpose.

In this section the rationale behind the development of the *product state concept* was introduced. The *product state concept* aims at reducing the complexity of a manufacturing programme by supporting the identification of relevant information within the system. In order to achieve this goal, the *product state concept* is actively analysing exiting co-relations within the manufacturing programme, which consequential may contribute to transparency and support knowledge acquisition. This increase in transparency and applicable knowledge may contribute to product and process quality improvements of manufacturing programmes. The *product state concept* is looking at individual items. This differs from most existing approaches by considering the individual item as 'one product' from the first value adding process to the last only by subsequent change(s) of its state. Each individual product migrates through the set of states along the manufacturing programme and thus the population of products relevant to each specific manufacturing programme grows. This growing population of products represents product state knowledge, which can be used for manufacturing monitoring and analysis. In order to achieve the goal of supporting transparency and reduce complexity, a method of determining checkpoints for data capturing within the manufacturing programme has been developed as part of the *product state concept* (Wuest et al., 2014c).

In the next subsections, different elements of the *product state concept* are developed, described and defined, starting with a definition of the product state itself.

4.2 Product state[12]

A manufacturing programme transforms raw material to final products through different value adding manufacturing processes in order to deliver to the customer the desired product. Consequently, the goal of every manufacturing programme is to add value to a product (Kalpakjian & Schmid, 2009). Adding value in manufacturing implies physical transformation of the product (e.g., transformation of form, hardness, chemical composition, etc.) over the course of the manufacturing programme. The specific purpose of every manufacturing process and operation is to execute a part of the physical transformation of the product. Thus, the state of the product is changed at least with every (value adding) process or operation. Looking at a product by its state has the advantage to being able to describe and/or monitor this transformation. Therefore, looking at the product state along the whole manufacturing programme accumulates a complete picture of realized measures and transforming processes. The product state is the core of the *product state concept*. In this section the term itself and its background will be described. For a comprehensive understanding the usage of the term product state and similar notions in other domains are briefly introduced.

The term product state itself is sporadically used in different fields. In physics for example, the term product state is used in the research field of quantum systems (Verstraete, Wolf & Cirac, 2007; Chen, Su, Wu & Oh, 2012; Haegeman, Cirac, Osborne & Verstraete, 2012). In the engineering domain, the term product state is generally accepted and adopted. However, certain researchers used the term in different domains, describing different circumstances. All of which have an overlap to a certain extent with the definition of product state used in this work. Musa et al. (2013) are looking at product visibility from a supply chain perspective and mention the physical state of a product, which is important to be tracked during the product lifecycle. In principle in line with the understanding of product state in this work, they focus mostly on logistics operations. However, they mostly look at perishable goods like food items and not at manufacturing products going through value adding manufacturing processes. Anderl, Picard & Albrecht (2013), when looking into designing smart products, use the term to derive the description of functional behavior. Toenshoff & Denkena (2013) present an interesting view on states connected to a product in the production engineering domain. Though they do not talk directly about the product state by itself, but connect the term state to a work piece in the research domain of CAD and simulation. They connect certain state variables to the state of a work piece, describing it under different circum-

[12] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2011b; Wuest, Klein, Seifert & Thoben, 2012c; Wuest et al., 2013b).

stances, like e.g., after hardening. Together with Kumar (2002), who states "existing processes can be made intelligent by adding sensors to monitor and control the state of product being processed" (Kumar, 2002) these usages of the term are in principle in line with the sense as it is understood within the *product state concept* presented here. However, both do not go into detail how they define the "state of product" (Kumar, 2002) or the state of a work piece (Toenshoff & Denkena, 2013).

Next, a definition of the term product state is described. It has been noted before, that the manner of describing a manufacturing product, e.g., gear made of steel, will be different from the description-style of a product designed to fulfill an aesthetic purpose (in addition to the functional purpose) in mind, e.g., a plastic rear mirror. At the same time, the individual describing a product influences the description based on, among other things, its own background, knowledge and experience. Therefore, the approach of describing it through its product state will help to align the descriptions in a commonly understood manner as well as increase transferability and usability of accompanying information by the addressees.

At the moment, the term product state itself is not sufficiently defined for the use in manufacturing programmes as planed in the *product state concept*. Looking at literature, the term *'state'* is frequently used in various areas like physics, chemistry, medicine or even philosophy. However, the transfer and usability of these existing definitions to the context of industrial production processes is not easily achievable. Nevertheless, understanding what stands behind different definitions and what the major differences are is important for the definition of product state in industrial production processes. One of the oldest and most common used definitions is the state of aggregation, which describes the simplified classification of material as solid, fluid and gaseous (Hüttig, 1943). The example of the state of aggregation of water can help to understand two aspects which are universal over most state definition and will be important for the definition of product state at a later point. First, the state is time-dependent. Water can be at checkpoint A (t=0) in a liquid state and at checkpoint B (t=1) in a solid one (see Figure 22).

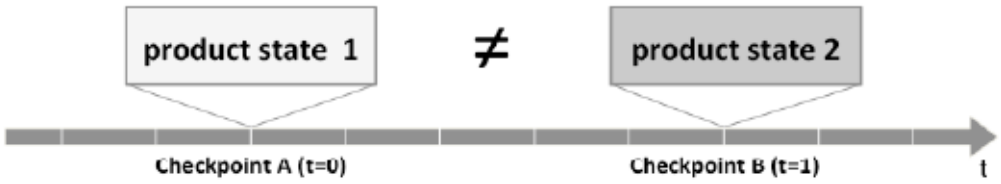

Figure 22: Product state is time-dependent

Another aspect is the descriptive character of state. Within the research field of ICT the term state is used, for example, in UML to describe a constraint of an object during the lifecycle (Gogolla & Parisi-Presicce, 1998; Schöning, 2001). The state is then active when the constraint becomes true. Another way of using state in

ICT is with finite-state machines. Within this theory, a new state (t=1) depends always on an original state (t=0) and an input. Again, the time-dependency of the state occurs, plus the dependency of state on external input or change.

A closely related concept is the so-called 'state transition model', describing an existing state and actions which transform the original state in a new one (Chander, Dean & Mitchell, 2001). The state transition model is applied in different domains, e.g., in medicine to describe e.g., changes in tumor growth by its state (Sonnenberg & Beck, 1993; Mei, Xie & Zhang, 2004) or in defense application like intrusion control (Goseva-Popstojanova, Wang, Wang, Gong, Vaidyanathan, Trivedi & Muthusamy, 2001). The descriptive character of common state definitions is also pointed out within the field of thermodynamics in which the term state is clearly defined. In thermodynamics, the state of a system describes a situation in which all variables of the system can be allocated with a clear numerical value. These variables are called state variables. The number of state variables, which is necessary for a definite determination of state, depends on the inner structure and complexity of the system (Geller, 2006).

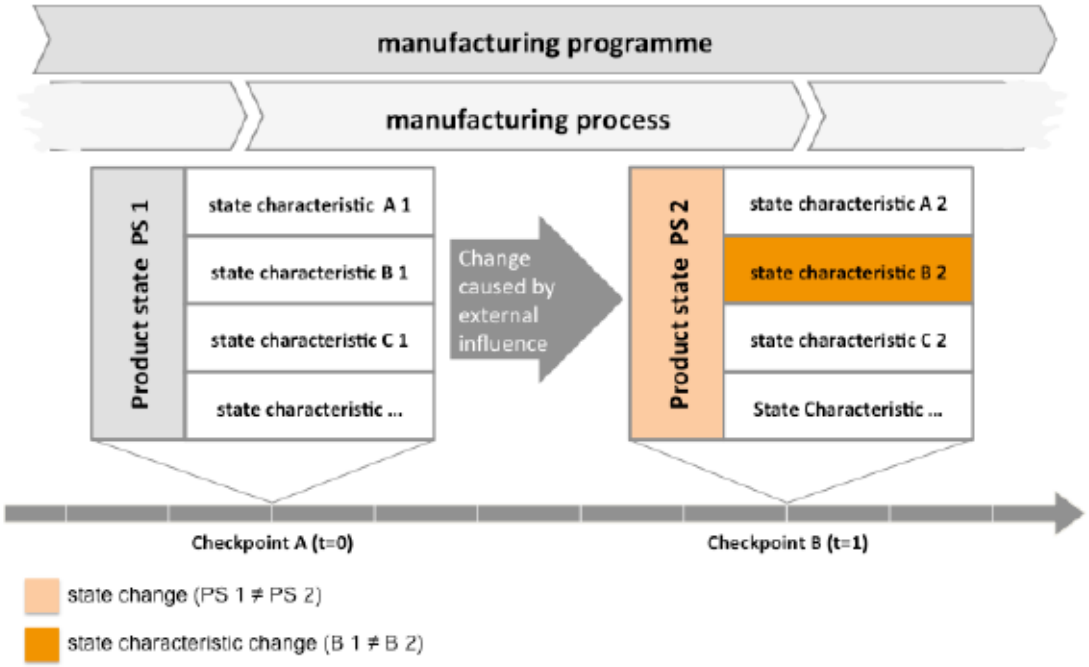

Figure 23: Schematic product state change due to external influence

This definition introduces variables with attributed certain numerical values. It also implies that the number of variables needed depends on the situation. In other definitions or descriptions of state, these variables are also known as properties, parameters, attributes, factors or characteristics. Within this dissertation, the term characteristics will be used from now on as a descriptive element of state. Looking at the thermodynamic literature on how characteristics can be defined, one can find them described as qualitatively definable and quantitatively measureable physical quantities (Geller, 2006). Other fields take a more uncommitted approach and view

characteristics as a qualitative describable value or appearance without the means to quantify it (Mayer-Bachmann, 2007). For the product state in manufacturing, a combination of quantitative and qualitative characteristics will be used.

Based on elements of the above definitions combined with the requirements of the manufacturing domain, the product state can be defined as follows:

The product state describes a product at a certain time during the manufacturing programme or after through a combination of state characteristics. State characteristics are definable and ascertainable measures, which can be described in a quantitative or qualitative way, e.g., weight or chemical composition of the material. The product state changes due to external influence, for example machining or corrosion from checkpoint A (t=0) to checkpoint B (t=1) when at least one descriptive state characteristic changes (see Figure 23).

The external influence causing the change of product state along a manufacturing programme can be linked to manufacturing process parameters (Chryssolouris & Guillot, 1988; Monostori, 2002) and other factors like environmental parameters, e.g., humidity or vibrations. The change or transformation of product state can be categorized in different categories. To illustrate, during a manufacturing programme the state changes to a certain degree with every manufacturing process as value is added. This change is intentional, but it also involves the repercussion of unintentional changes. For example, when cutting a steel pipe the intension is to reduce the length to a certain degree. However, wanted or not, at the same time the weight of the pipe will be reduced due to the cutting of material. This is a simplified example. In reality many more characteristics change with every process in a manufacturing programme. This issue and the different categories and their relations to one another are investigated in the following subsection 4.3.3.

A more complex example also emphasizing the importance of looking at a manufacturing programme as a whole is the straightening of steel bars after heat treatment in the steel mill. The changed geometry of steel bars due to distortion is often straightened out before delivery to customers. There are many reasons for this, such as ease of transportation, continued processing or simply the look and feel of the product. The main purpose is to clear the bending of the steel bars. Nevertheless, at the same time residual stress is caused which can lead to problems at a later stage of the manufacturing programme, e.g., after the final heat treatment. For this reason, it is important to think of what characteristics have to be known to describe the product state. This will be highlighted in the next section.

The *product state concept* allows for a description of the actual current state of a product during a manufacturing programme. The final goal of every manufacturing programme is to produce products of the desired quality and thus meet the customer requirements and expectations. The customer requirements for the final product may be seen as an ideal product state, the final product of the manufacturing pro-

gramme has to reach. All final product states meeting this goal are 'good states' and all product states that do not inherit the expected ideal product state may be considered of 'bad state'. As this is not a black or white matter, there are several 'shades of grey' in between, depending on the degree of fulfillment of the customer requirements. In this example this is solely focused on the final product state after the product went through all stages of the manufacturing programme.

This may be overly simplified, as in reality the issue is more complex. First of all, there may be also "good states" and "bad states" including the gray areas in between, throughout the manufacturing programme, which represent not final product states, e.g., the product state a product inherits after operation XX before operation XY. Those and their implication on the final product state may be more challenging to distinguish. With the categorization as 'good' or 'bad' of the intermediate states depends additionally on other factors as on whether or not the state has the potential to be transformed towards 'good' by the end of the manufacturing programme. This can be achieved e.g., by adjustment of the process parameters of following manufacturing processes. This of course adds further to the complexity.

The second issue is, that customer requirements may include factors of which the fulfillment by the product shows only over the usage phase and may not directly be measured after manufacturing at the current state of knowledge. However, applying customer feedback and item-level PLM principles, these customer requirements may be included in the analysis of relations between states, intermediate states and state characteristics during the manufacturing programme. An approach how this information may be utilized and the newly generated knowledge be integrated in the *product state concept* is presented in section 5 by applying supervised ML on product state data.

In this context the issue of selecting appropriate checkpoints for information capturing in manufacturing programmes represents a crucial prerequisite for the determination of states. It has been established that manufacturing systems are becoming more complex and with this development the challenges towards information management increase. With each step along the manufacturing programme, with each value adding process or operation, the number of subsequent product states increases. The *product state concept* relies on the availability of relevant information and data (of the product state) along the manufacturing programme at the right time (use of data). There is a lot of research available on when the relevant information must be available during a manufacturing programme. In certain areas of production with a specific purpose of monitoring, e.g., concerning cycle times, structured methods exist to determine checkpoints (Heinecke, Lamparter & Kunz, 2011). However, determining the right time (capturing of data) within the manufacturing programme to capture the relevant information (product state) is more complex and is not yet sufficiently discussed by industry and academia. This

theoretical background of information capturing timing in manufacturing programmes has been published in Wuest et al. (2013c).

Based on those findings, a more practical approach has been developed transferring the well-established stage gate model (Cooper, 2008; Cooper, 2010) from the product development domain to the manufacturing domain (for further details Wuest et al., 2014c). Concluding, the findings indicate that the determination and positioning of checkpoints in a manufacturing programme is a very individual task. Despite many practitioners placing checkpoints in between processes and operations, this can be appropriate but is not the solution for all cases (Shetwan, Vitanov & Tjahjono, 2011). The findings indicate that common sense, in-depth knowledge of the manufacturing programme assisted by following certain rules, as presented in (Wuest et al., 2014c) may help to choose relevant checkpoints.

However, as in practice it is a continuous struggle to get enough data from manufacturing processes and given the computing power and Big Data developments, the question may have to be adapted from "how can we choose checkpoints in order to not get too much data" towards "how do we have to choose checkpoints in order to get as much data as possible". This issue will also be addressed briefly in the later section 5.

4.3 Relevant state characteristics

In the definition of product state in the previous section, state characteristics were introduced as the determining factor of product state in combination. In this section, product state characteristics are described in greater detail. Furthermore, a distinct focus is laid on state transformation, which occurs when at least one state characteristic changes due to external influence. Therefore, different categories are developed in order to categorize and further illustrate such state transformation in manufacturing programmes. Questions like 'what are relevant state characteristics' and 'how can they be identified' are discussed within this section, before the next section discusses the important occurrence of co-relations between state characteristics and their influence on state transformation.

4.3.1 Product state characteristics[13]

The product state represents a combination of different state characteristics describing a product at a certain stage during the manufacturing programme. Product state characteristics are definable and ascertainable measures, which can be described in a quantitative or qualitative way, e.g., weight or chemical composition of the material. In this section, state characteristics are described in more detail. A

[13] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2011b; Knoke, Wuest & Thoben, 2012)

categorization of state characteristics is presented for a technical product and an example including a selection of state characteristics is provided.

Theoretically all state characteristics describing a product one can think of could be included for a complete description of a products state at a certain time during a manufacturing programme. However, this is neither reasonable nor practical. There are various reasons for a selection, important ones being summarized under 'technical', 'economical' and 'knowledge' reasons. Therefore a selection is sensible under these circumstances. However, this does not mean, just as a state characteristic and/or its influence of others or the manufacturing programme is not known for example it is not important. The identification of relevant state characteristics, which describe a product as complete as possible at the time of description is one goal of this research and described in detail in the following section 4.3.5. The identification is not a static process but a dynamic one, which implies that the set can change at all times. A trigger for such a change can for example be when new knowledge concerning the state characteristics, their relation, the manufacturing programme or the customer requirements is available. To create such new knowledge may be a result of the application of the *product state concept*, making it a continuous process.

State characteristics change, or transform, due to external influence. This can be through a manufacturing process or operation or environmental influence, like e.g., corrosion due to high humidity. When state characteristics are subjected to transformation through a process, the process parameters have a major influence. Whereas the process parameters are supposed to be planned and controlled, the environment is often not taken into consideration to the same extent. However, the environmental parameters can have a big influence on the output state as well, e.g., vibration during machining. In some manufacturing processes the environmental parameters are actively controlled, e.g., dust particles in chip manufacturing.

The process and environmental parameters are not the only influential factors when it comes to the transformation of state characteristics. Very important factors represent the relations between the state characteristics themselves. For one, the input state characteristic has a major influence on the output state characteristic: directly, as it determines what can be achieved and indirectly, as it influences the impact of other parameters like process and environment on the output state. Another factor are process intra- and inter-relations between different state characteristics which influence the output directly and indirectly through other parameters. This aspect gains in importance when looking at the manufacturing programme with different processes and operations. The process intra- and inter-relations of state characteristics and their influence on the *product state concept* are described in detail in the later section 4.4.

Following, an example for state characteristics of a technical product is presented. The intention of this example is to present what state characteristics in an industrial environment and introduce the complexity inherit within just this single manufacturing process. Additionally, the example shall provide a perspective of the possible high number of possible state characteristics. A steel cylinder during the manufacturing process of machining is chosen to represent a simple example for such a technical product. It has to be noted that the more complex a product is, the more state characteristics and categories of state characteristics it may inherit. For the chosen example the product state characteristics can be clustered in three major categories (Brinksmeier, 1991) (see Figure 24):

- surface state characteristics: describe e.g., the geometry of the product
- peripheral-zone state characteristics: describe e.g., the structure of the peripheral layer which often differs form the internal structure
- internal state characteristics: describe e.g., the material and material related properties of the product

Figure 24: Categories of state characteristics applied to a steel cylinder (peripheral-zone enlarged compared to reality for illustrative reasons)

Surface state characteristics are defined by their form, their location and/or their dimension/measurement (in a coordinate system). Form and location elements are only required when the defined dimensions/measurements are not sufficient. This primarily applies to co-axiality, symmetry and running deviations (DIN EN ISO 1101 according to Keferstein, 2011).

A product can be described by form elements which are based on standard geometries, e.g., cylinder. Location elements are described by the positioning of elements towards each other, e.g., direction, location and running. In Figure 25 a summary of form and location elements is summarized (DIN EN ISO 1101 according to Keferstein, 2011).

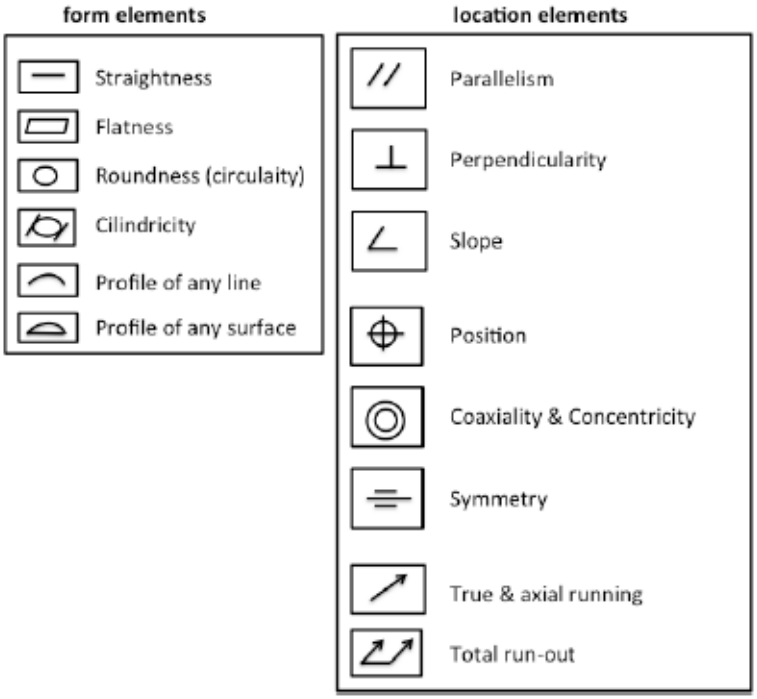

Figure 25: Summary of form and location elements (DIN EN ISO 1101 acc. to Keferstein, 2011)

Dimensions and measurements of surface state characteristics can be linked to one or more form and location elements. In general there is a distinction between internal and external dimension. External dimensions e.g., describe the distance between two parallel plains or tangent planes, which represent external boundaries of the product body. Internal dimensions describe e.g., the distance between two parallel plains or tangent planes, which represent internal boundaries of the product body. Figure 26 illustrates the difference between external and internal dimensions using a cylinder with a drill hole as an example (Westkämper & Warnecke, 2010).

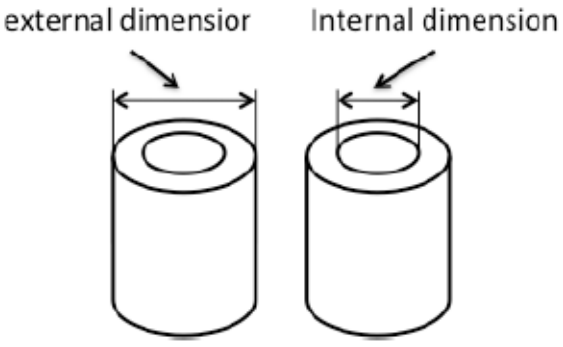

Figure 26: External & internal dimension (exemplary) (Westkämper & Warnecke, 2010)

The surface state characteristics form, location and dimension may be understood as a rough description of a product and the micro geometry of the product surface as a more deliberate description. This is not a judgment on the importance but

based on the dimensions of tolerances. The micro geometry focuses mainly on two characteristics: (surface) roughness and waviness.

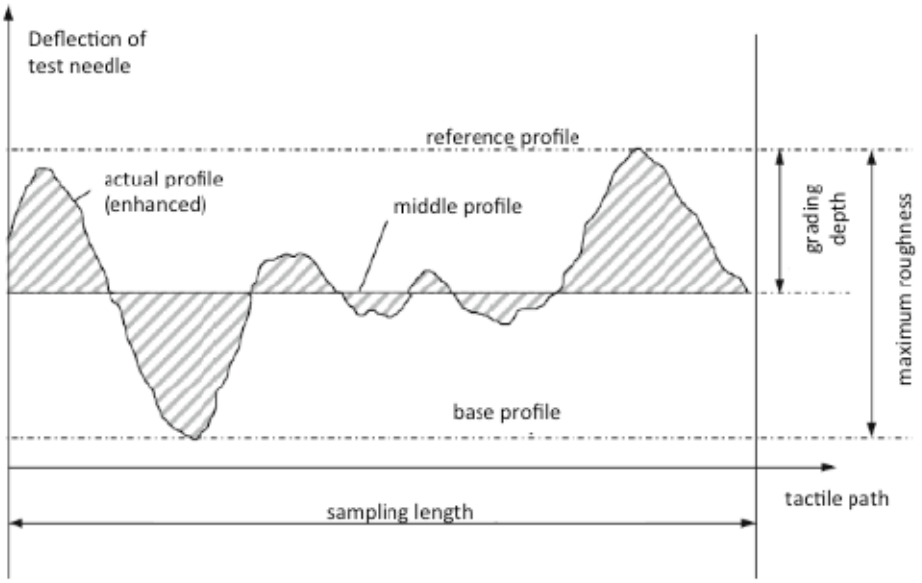

Figure 27: Profile illustration of a product surface (ad. from Westkämper & Warnecke, 2010)

The (surface) roughness derives mostly from a regular or irregular, short-wave deviation in shape whereas waviness derives in most cases from periodical long-wave deviations in shape. In case of the exemplary machining process, (surface) roughness is triggered by process parameters like geometry and kinematic of the cutting tool and type of chipping. Waviness in this scenario is a consequence of e.g., disturbances and oscillations, which may occur during the machining process (König & Klocke, 2008). The (surface) roughness and its parameters can be illustrated through a profile, which can be compared to a lateral cut of the surface region (see Figure 27) (Westkämper & Warnecke, 2010). In Figure 28, an exemplary selection of surface state characteristics is illustrated with no claim for completeness presenting three groups: 'coarse' characteristics, micro geometry characteristics and optical characteristics. There may be various other possible ways of grouping, additional groups and/or characteristics possible.

The peripheral-zone state characteristics and the internal state characteristics of a technical product influence various parameters, which are often relevant in accordance with customer requirements like endurance and reliability. In the following paragraphs, these categories will be presented within a similar scenario of a cylinder during the manufacturing process of machining.

The ***peripheral-zone state characteristics*** of a technical product are determined by the entity of the physical and chemical characteristics of the peripheral-zone. Among those characteristics are the microstructure, hardness, cohesiveness and residual stress (Brinksmeier, 1991).

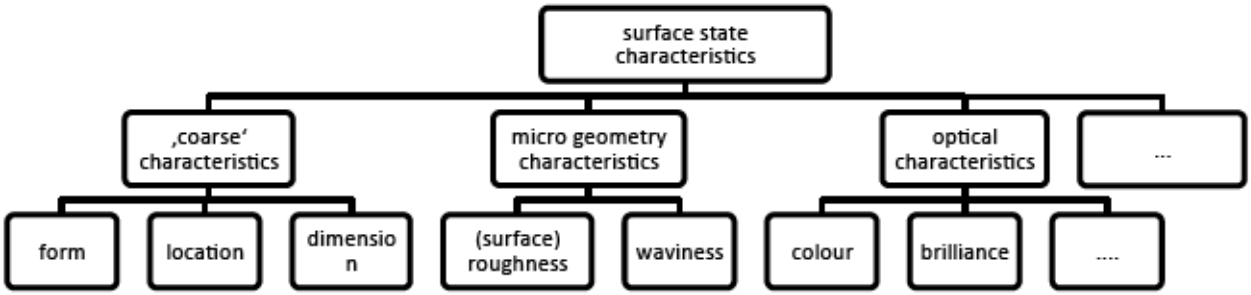

Figure 28: Selection of surface state characteristics with focus on machining processes

The microstructure is a grain structure, which may have an influence on transformations of mechanical strength and hardness. The hardness is e.g., being measured by a micro hardness test (DIN 55676) in the peripheral-zone (ca. 3mm distance to the surface) using a test load of 1N, 0.1N or 0.05N (Vickers test). With increasing distance from the surface towards the core, the hardness is decreasing. This is one of the reasons for the differentiation between peripheral-zone state characteristics and internal state characteristics. Residual stress is stress that comes into effect without influence of external force(s) and thus loading stress. The degree of oxidation resistance and/or corrosion resistance is to a large extent depending on the peripheral-zone of a product and describes the resistance against influence of external factors like e.g., air, water or chemicals. Among peripheral-zone state characteristics is one that may also be part of the surface and internal state characteristics or at least has an overlap depending on its size: the crack. A crack is a locally distributed separation with limited width but often of considerable length. It can be caused by e.g., internal and/or external stress (Söhner, 2003).

Summarizing, peripheral-zone state characteristics may have an influence on the functional properties and therefore the functions and fulfillment of customer requirements of a product (König & Klocke, 2008).

Internal state characteristics are state characteristics located within a product. Products are composed of basic materials, being elements of the periodic table. The smallest parts of those elements are atoms, which consists of a certain amount of elementary particle, thus distinguishing the different elements. The atomic arrangement in a solid-state body can be amorphous or crystalline. In a crystalline state the atoms (molecules) are arranged in a periodical, spatial mesh (Seidel & Hahn, 2010). Internal state characteristics reflect the characteristics of the basic material, which are comprised in a product. In a technical product like the chosen steel cylinder, these characteristics can be categorized in chemical, mechanical and physical characteristics (Brinksmeier, 1991). In Figure 29 the categories and selected internal state characteristics are illustrated.

Chemical state characteristics of a material can be described by the type, size, arrangement and orientation of the atoms or metallographic constituent (Schatt & Worch, 2003). The chemical composition and structure of a solid-state body have a strong influence on other internal state characteristics. The mechanical and physical characteristics are determined by the base grid as well as by type, number and location of grid imperfection/defects and grid contaminants (Seidel & Hahn, 2010).

Mechanical state characteristics are determined by the behavior of the material towards strain and stress by (external) forces and/or momentums, e.g., cohesiveness and viscosity. They are defined by specific values, which are established through e.g., a tensile test. Among the mechanical characteristics are e.g., stiffness, wear resistance and fatigue strength (Schatt & Worch, 2003).

Physical state characteristics represent substance-specific values, which are derived through measurements and experiments. The characteristics of the material are not changed through the measurement or experiment. Among physical state characteristics are e.g., mass density [Kg/m³] and thermal conductivity [S/m] (Seidel & Hahn, 2010).

Additional to the above-presented categories of internal state characteristics there are certain characteristics, which may be important for the manufacturing programme and are based on the internal structure but do not fit the above categories. Among those are e.g., sinter ability, weldability or castability (Reuter, 2007).

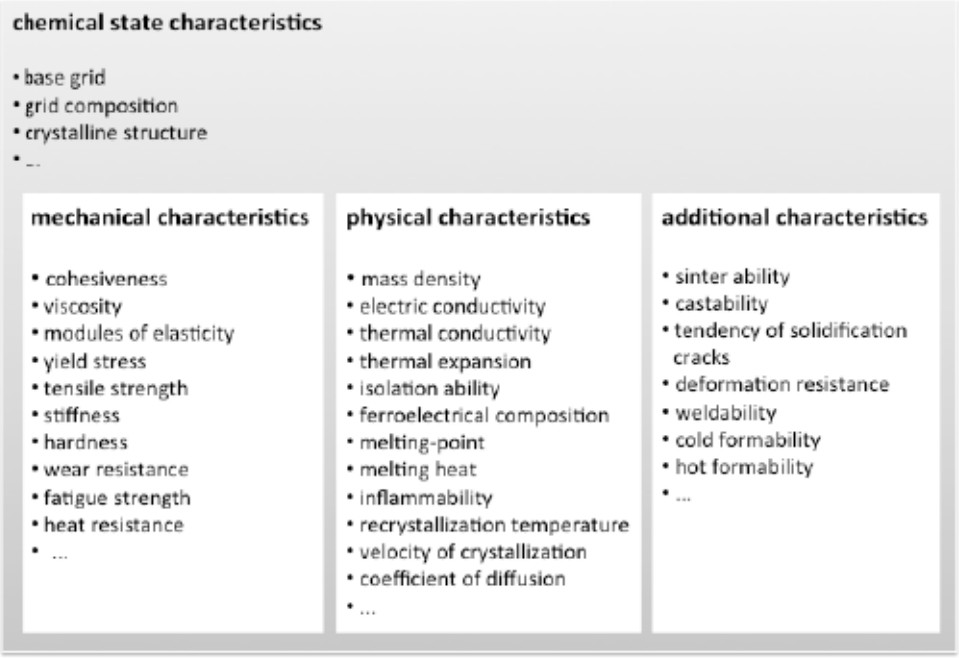

Figure 29: Selection of internal state characteristics and categories

It is important to acknowledge that there is a very large number of state characteristics available to describe the state of a product. However, it can be assumed that

certain 'higher level' product characteristics exist among them. Their 'importance' or 'relevance' is based on their relation towards the final product state ('good' or 'bad'). Various parameters, e.g., the aforementioned reliability, may influence the determination of relevance for these state characteristics. These relevant state characteristics are discussed in further detail in section 4.3.4. However, looking at the possible process intra- and inter-relations, it is important to distinguish between correlation, which is a statistical relationship between variables, and causation (causality), referring to an event (e.g., change of variable) being the consequence of another (e.g., process parameter). This is further detailed in section 4.4. The influence of these relevant state characteristics for the application of subsequent application of ML techniques and the question of correlation or causation is further investigated in section 5.3.

In this section product state characteristics were introduced as being part of a product state description. Next, the transformation of a product's state along a manufacturing programme is illustrated. In order to make the theoretical construct more comprehensible, the example based on a manufacturing process 'machining' introduced above is being continued with a focus on state transformation.

4.3.2 Product state transformation

The previous sections introduced the product state itself, state characteristics and briefly mentioned the change of state/state characteristics due to external influence during manufacturing programmes. It has been established before that the goal of every manufacturing programme is to add value to the product (Kalpakjian & Schmid, 2009) with each process or operation by transforming its product state (see Figure 4). In this section, this change of state, from now on referred to as transformation of state, will be described in detail.

At first the difference of the product state transformation and the transformation of individual state characteristics has to be discussed. Whereas the product state transforms as soon as a single, individual state characteristic changes according to the definition of product state, not every individual state characteristic changes when the product state transforms. The extent of how many state characteristics transform can vary theoretically from one to all of them. Most of the time the number of changing state characteristics will be in between these two extremes.

The state transformation is influenced by different factors. These factors can be internal or external. So can e.g., the environmental conditions of the manufacturing process have an influence on the transformation of state. An important factor are the process parameters. They determine to a large extent the outcome of the process. The influence can be directly or indirectly, which is detailed in section 4.4.

Next, the influencing factors and their direct and indirect impact on state transformation are presented based on an example. This is due to the otherwise sheer end-

less number of possible factors. A concrete example provides boundaries within the descriptions and allows the reader to connect the description to previous elaborated information. The example continues based on the previously introduced example around the manufacturing process 'machining'. A selection of manufacturing process parameters of 'machining' with an influence on the state characteristics and their transformation are presented in the following paragraphs. Even though the presented selection has no claim for completeness, it again provides an indication and highlights the complexity already inherited by a single manufacturing process. Projecting this example on a whole manufacturing programme with multiple processes and operations and additional cross-process relations, the overall complexity and thus high-dimensionality of information needed can be envisioned.

Machinability is defined as the characteristic of a product to be machined under given conditions (DIN 6583). Therefore, machinability describes all characteristics of a product, which influence the machining process.

The machinability of a product has to be seen in relation to the chosen

- machining process (e.g., turning, drilling, milling, etc.),
- cutting material (e.g., HSS, carbide metal, etc.) and
- process parameters (e.g., cutting rate, feed rate, cooling, etc.).

The machining process realizes plastic and elastic deformation of the product. The cutting tool penetrates the surface under the application of energy and separates the chips from the product. There are different machining processes, in this case the machining process 'turning', a process with geometric defined cutting edges, will be used. The basic principles can be transferred to the similar other machining processes e.g., drilling, milling and broaching (Westkämper & Warnecke, 2010).

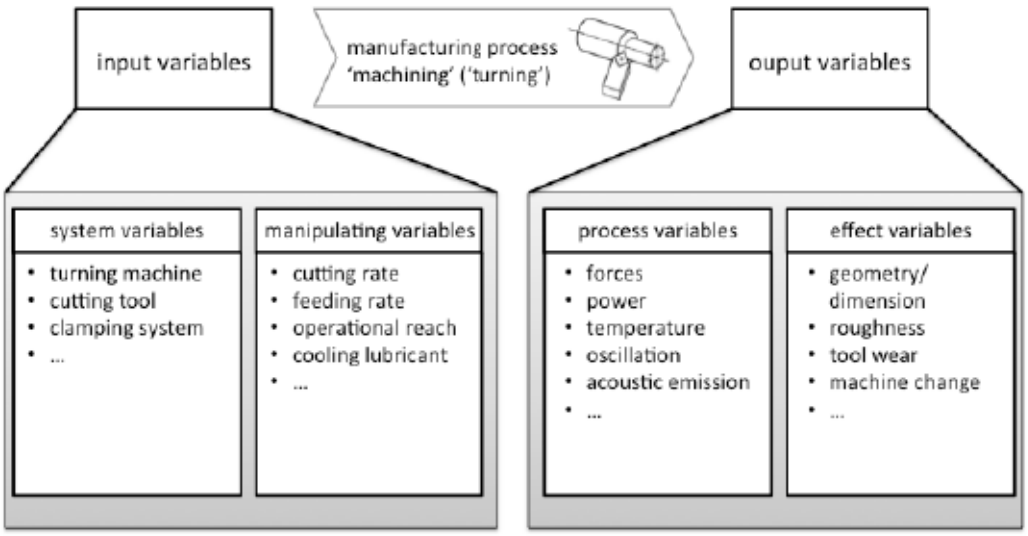

Figure 30: Variables of machining processes (based on Denkena & Tönshoff, 2011)

In Figure 30 a selection of influencing factors of the transformation are presented. The factors are firstly structured into input and output variables. ***Input variables*** summarize system variables and manipulating variables. The *system variables* consist of elements of the manufacturing process, which are fixed, at least for more than one iteration. This can be the model of the turning machine used or the type of clamping system (e.g., three jaw chuck). The *manipulating variables* on the other hand may change depending on the machining plan for the product. They may be adjusted manually or automatically through a programme. ***Output variables*** condense process variables and effect variables. *Process variables* are directly derived from the manufacturing process like occurring forces, the processing power or temperature. The *effect variables* present the results of the machining process concerning the product (the new product state/state characteristics), machine (e.g., wear, temperature), tool (e.g., wear) and cooling lubricant (e.g., temperature, contamination and chemical transformation) (Denkena & Tönshoff, 2011).

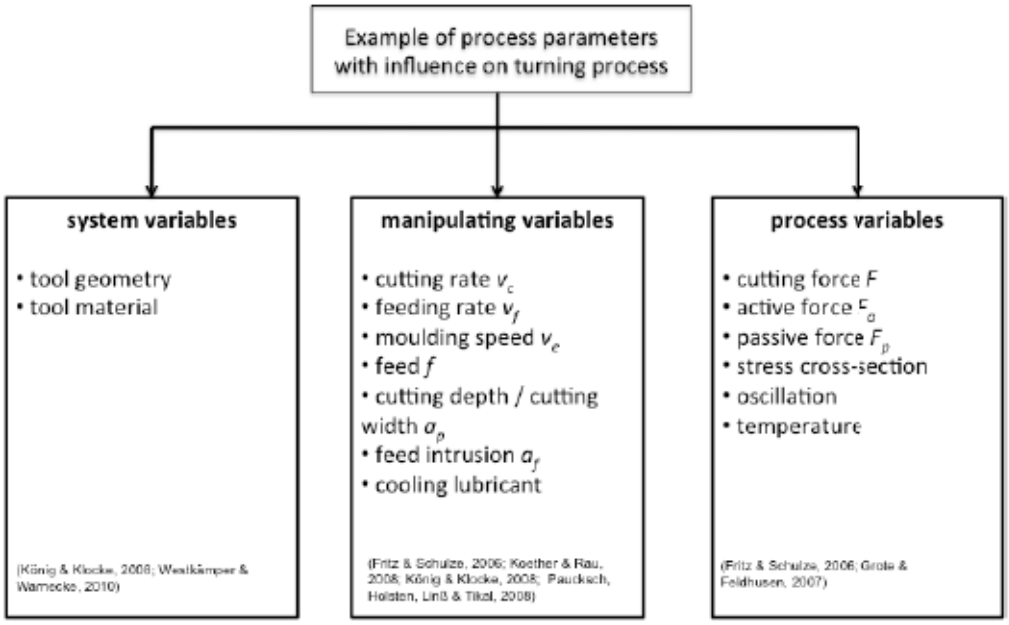

Figure 31: Example of process parameters with influence on turning process

Looking at the different variables, the system variables, manipulating variables, process variables and partly the effect variables are summarized under process parameters from this moment on. Only the product describing effect variables considered product state information. In Figure 31, a selection of manufacturing process parameters with a known influence on state transformation during the turning process are summarized. These process parameters with a known influence on the transformation of state during the turning process do not reflect all influencing factors. Even so it would be ideal to only focus on independent variables and thus causation rather than correlation this currently not possible. For example, the product state prior to the focus manufacturing process influences the process parameters directly and indirectly, and thus the state transformation. It is important to note that with 'prior to the manufacturing process' not only the product state directly

prior to the focus process, but also the various product states before that along the manufacturing programme may have an influence (section 4.4).

4.3.3 Categorization of product state transformation[14]

After the previous section detailed the transformation of state and influencing external factors, this section presents the different categories of product state transformation. Hereby, the focus is on possible challenges associated with the state transformation in manufacturing. Different categories of state transformation are established and illustrated. It is important to understand, that the categories are not exclusive. A state transformation and its influence on the different state characteristics and the product state itself may be described by more than one category. The categories of state transformation mostly focus on transformation of individual state characteristics. The categories of state transformation are established from the perspective of the process manufacturing programme owner and not from a universal perspective, e.g., some state transformations may be known to person A but not to person B. One has to understand that the aforementioned perspective is only valid for a certain point in time, as the population of state and the state transformation could be of dynamic nature and thus change over time. Also, as it is described later in this section, the knowledge of the population and state transformation is rather limited today. These indicators present already at this stage arguments for the later application of ML algorithms within this context (see section 4.5.2).

At first, a crucial category of state transformation directly connected to the main goal of a manufacturing programme, create added value to a product (Kalpakjian & Schmid, 2009), is introduced. This category focuses on the issue if the state transformation is ***intentional*** or not. This is related to the agreed upon quality definition for this research, connecting quality to the fulfillment of requirements. An example for this transformation categories is the following: In a manufacturing programme as described in Figure 13 the last manufacturing process, heat treatment adds value to the product by changing the hardness of the product to meet the customer requirements. The transformation of the state characteristic 'hardness' during the manufacturing process is intentional as it represents the planned output, a product with a higher hardness than before. A change of geometry is not the main goal of the process 'heat treatment' but can occur nevertheless during the process. This may be considered a ***non-intentional*** state transformation. It may be the case that a manufacturing process has only one intentional state transformation. This especially is often the case for manufacturing operations. However, there are manufacturing processes, which target more than one intentional state transformation. An example for such a manufacturing process is 'grind hardening', which targets the

[14] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2011a; Wuest et al., 2011b; Wuest et al., 2012c; Wuest et al., 2013b)

state characteristics 'hardness' and 'surface roughness' simultaneously (Brinksmeier & Brockhoff, 1996; Brockhoff, 1999).

Another category, very closely related with the one described above as intentional, can be described as ***planned***. Planned state transformations are transformations of state characteristics, which are anticipated and thus 'planned' by the process owner. Whether or not these transformations are contributing to the goal of adding value to the product and fulfill the customer requirements is not in the focus of this category. The importance lies that the process owner is aware of the change and may thus reacts accordingly to reach the quality goal of the manufacturing process. As long as state transformation occurs as planned, the output is meeting the quality goal (for the state characteristic in the focus). As soon as a state transformation, which is ***not planned***, occurs during a manufacturing process or operation, the outcome may vary from the quality requirements and thus jeopardize the quality goal. It has to be understood that this is an idealized scenario as when taking into account the other categories, the 'planning' based on e.g., customer requirements presents various pitfalls in itself. The customer requirements may be wrong, inaccurate or misunderstood. Also it may be that the transformation of customer requirements into a product concept can be wrong, inaccurate or misunderstood and thus the derived planned transformation may be problematic to begin with. Furthermore, the planned state transformations may not be possible disconnected from others, making this categorization a merely theoretical accentuation. Nevertheless the dynamic nature of state transformation is highlighted again.

Some state characteristics are not considered relevant and thus not part of the customer requirements. This is unveiled in a later section focusing directly on relevant state characteristics. For an example of planned state transformation, the above example of the manufacturing process 'heat treatment' is utilized again. As stated above, the intentional state transformation of this process is to change the 'hardness'. If a state transformation is intentional, it is always planned. However, if a state transformation is planned, it is not necessarily, even mostly not, intentional. During heat treatment, besides the hardness, other state characteristics may change. One of these is the geometry of the product, which may change during the heat treatment process and is not intentional. However, process planers are aware of the transformation, which will occur and thus plan it accordingly. This is for the 'plannable' geometry change, which does not involve geometry changes due to distortion for example. As introduced before, distortion is a common challenge occurring regularly during heat treatment. However, state transformation, which is not planned, is not necessarily not desirable. An example for an unplanned state transformation, which is considered desirable, occurred while performing repairs on 30l beverage KEG containers. The case was that the top of the KEGs was deformed and was to be repaired by means of reinstating its original form by a specially designed machine preforming a rolling process. Whereas the planned state transformation was to change the geometry, the physical characteristics of the KEG where

also transformed. Through the rolling process, the hardness of the material changed which was not planned but represented a desirable outcome of state transformation.

Today, many practitioners and academics are aware of challenges linked to distortion during heat treatment. The state transformation of geometry linked to distortion is ***known*** even so it is not planned. This represents the next transformation category. This category describes state transformations, which occurrence is known to the process owner. State transformations, which have no influence on the customer requirements, fall, e.g., under this category. This can be a change of color during the heat treatment process. The process owner knows a change of color may occur but he does not plan with it, as it does not have an effect on the product quality in this example. The known state transformations are the largest group and all planned and intentional ones are always also known. The ***unknown*** state transformations are a very important group and focus area within this research. As they are unknown, for whatever reasons, the amount cannot be quantified. However, it can be assumed that it is a very large number of state transformations that happens which are unknown by the process owner. Within this group certain state transformations may be of potential benefit for the process and product quality if they were known, planned or even intentional. However, it is a challenge, especially by looking at the whole manufacturing programme and the cross-process/operation relations, to identify these state transformations with this potential. In the later sections, the identification of currently unknown state transformations and their impact on the process and product quality will be investigated further.

As was established in the previous paragraphs, the few intentional state transformations are a sub-group of the larger group of planned state transformations. Those are in turn a sub-group of the largest group of known state transformations (see Figure 32). However, these groups are not fixed and the state transformation affiliation to a certain category can change depending on newly acquired knowledge or changes in e.g., process, environment or customer requirements.

There are two additional categories, which are important to categorize state transformations. One is describing if a state transformation is ***measurable*** or not. It is implied that measurable in this sense means the state characteristic that changes can be measured and the delta between input and output value linked to the process. If a state characteristic is measureable depends on various factors, like economic (e.g., being to expensive to measure), technical (e.g., not possible to measure without destroying the product) or knowledge reasons (e.g., the state characteristic is unknown, therefore cannot be measured). Intentional state transformations are mostly measurable as they represent the output defining the quality of a process. The same stands true for planned state transformations; even so in this category it may be more often the case that the right output is assumed. For the rest of the known state transformations it is open it depends on the individual case if they are measurable or not. Interestingly, unknown state transformations may be meas-

urable as well and may even be measured during the process. However, without knowledge of state transformation this information cannot be applied effectively.

The category '*measurable*' is a very crucial one, as it represents a basic requirement of the next state transformation category ***controllable***. Being measurable is a requirement in order to control a state transformation. Thus, every state transformation considered controllable is at the same time measurable. However, not every measureable state transformation is necessarily controllable. 'Unknown' state transformations (to the process owner), even so they might be measurable and actually being measured during the process, are not controllable as the contextual knowledge is missing which allows the process owner to connect the measurement to the state transformation.

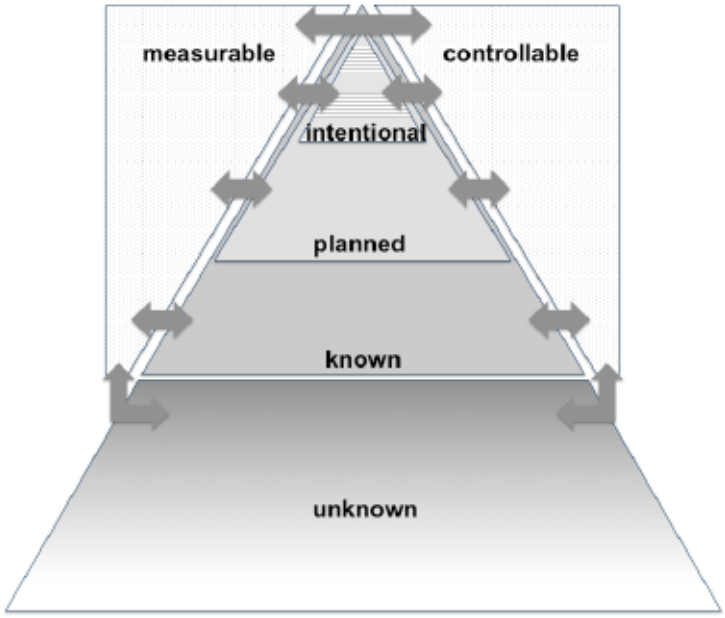

Figure 32: Theoretical distribution and linkage of state transformation categories (idealized)

The presumably large number of unknown state transformation (see Figure 32) does not mean that these state transformations and the state characteristics involved are not relevant. Looking at the whole manufacturing programme may even increase the likelihood of a necessary change of the categorization of certain state transformations as the influence across process and operation borders increases the need for knowledge about occurring transformations and transparency.

The application of pattern recognition described in later sections contributes to the goal of identifying and re-categorizing certain state transformations form unknown to known, planed or even intentional from a manufacturing programme perspective. Given the large number of assumed unknown state transformations and their potential impact on process and product quality this identification of currently unknown state transformations is one of the goals of the *product state concept* and will be described in more detail in following sections.

4.3.4 Relevant state characteristics[15]

The product state of an individual product within a manufacturing programme may be theoretically described at any time through a combination of state characteristics. Despite this deterministic approach, holistic knowledge concerning all state characteristics of a product during a manufacturing programme is neither worthwhile nor feasible. The question remains how the different sets of information can be described to distinguish them.

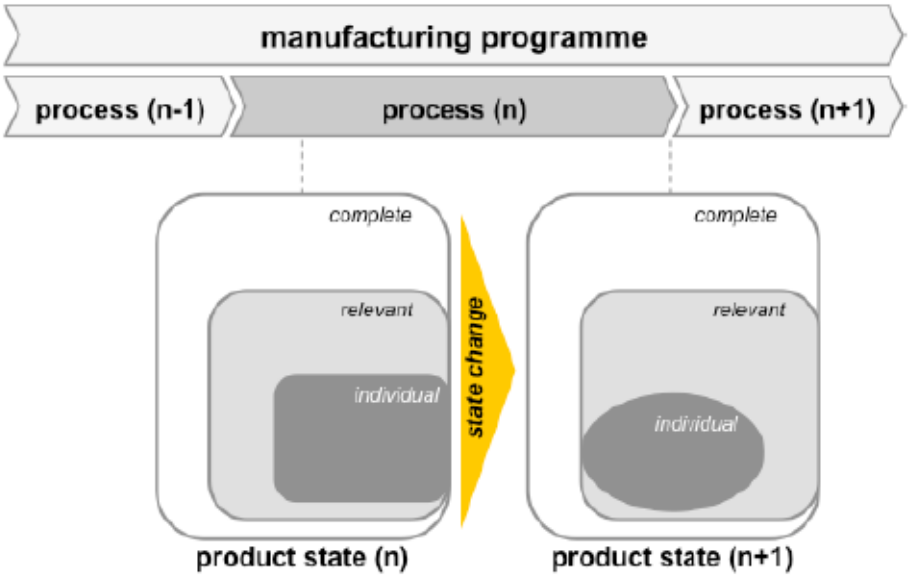

Figure 33: Theoretical information/data clustering of product state concept

Looking at those product state information from a theoretical perspective, three clusters of information/data sets can be identified: complete; relevant and individual (see Figure 33). The "complete" cluster resembles all information needed to describe every detail about the product and process, may it be relevant to achieve the desired outcome or not. This is however a purely theoretical set, as it contains a large amount of information with no impact on the manufacturing programme and thus the final product quality.

The "relevant" cluster on the other hand contains all information that is in one way or another relevant for the whole manufacturing programme. The challenge to identify a way to obtain this set of information is in the focus of this dissertation. The individual set of information is a subset of the relevant information set, representing the information relevant to the individual process. In order to simplify, in this and the following paragraph just the manufacturing programme and the manu-

[15] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2011a; Wuest et al., 2011b; Wuest & Thoben, 2012; Wuest et al., 2012c; Knoke et al., 2012)

facturing processes are used for illustration. Manufacturing operations, nevertheless just as important are not considered in this case to not increase complexity unnecessarily. However, the principle can be applied accordingly to operations.

The "individual" cluster resembles a subset of the relevant cluster as it is basically the relevant information for an individual process whereas the relevant cluster contains all needed information for the whole manufacturing programme. The diverse nature of the individual cluster is highlighted by the different shapes of the cluster in different processes in Figure 33. Individual information is important for monitoring and control of manufacturing processes in practice (see Figure 34). However, as this cluster is contained within the relevant information and the focus is on a holistic concept for a manufacturing system, within this dissertation the focus remains on the relevant information.

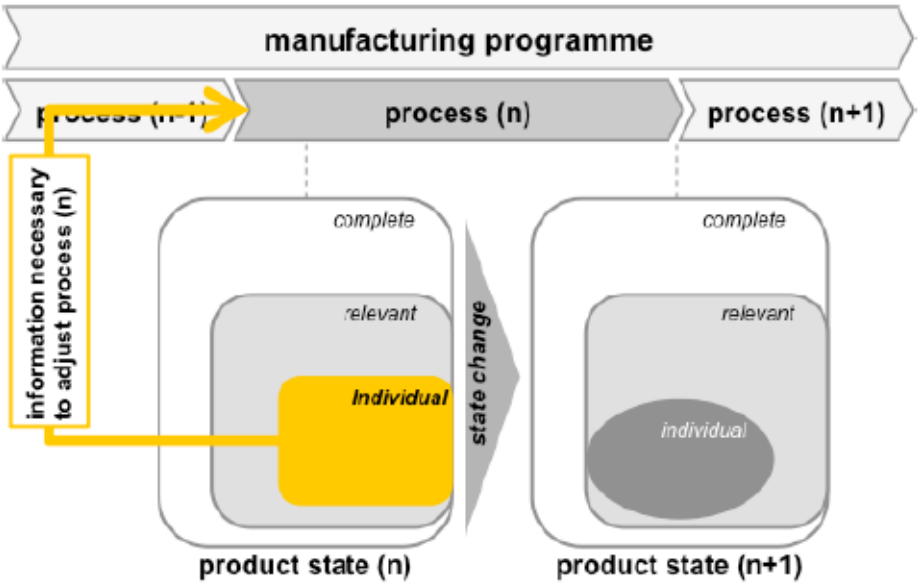

Figure 34: Individual set of information for manufacturing process adjustment

Practical reasons for not considering a state characteristic can be divided into three groups. They can either be technical (e.g., not measureable or measureable by destroying the product), financial (e.g., measurement is too costly), or caused by a knowledge gap (e.g., state characteristic is not known). However, some state characteristics may be characterized as relevant regarding their impact on the manufacturing process and the product state. It is therefore necessary to describe the product state based on an individual selection of relevant state characteristics. The accentuation is on individual because relevant state characteristics cannot be identified over all products and processes once and for all (Brinksmeier, 1991).

One way to identify relevant state characteristics is whether they include crucial information needed for each manufacturing process or operation. Therefore, a product state characteristic that neither impacts any manufacturing process or operation of the manufacturing programme nor influences other product state characteristics

may be disregarded. Knowing what the relevant state characteristics do may improve transparency and increase knowledge of the manufacturing programme itself. The state characteristics are often not independent, but relate to each other and form a complex (manufacturing) system.

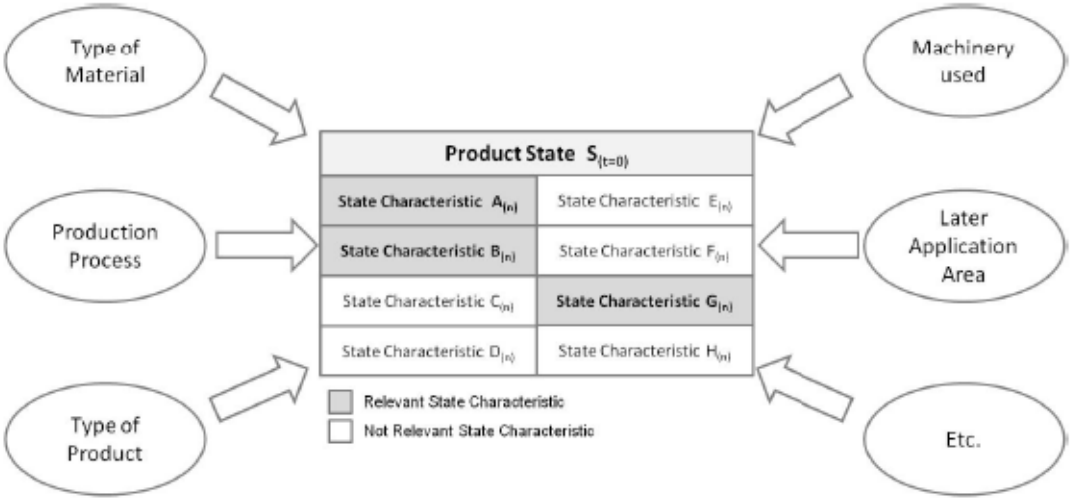

Figure 35: Exemplary parameters with influence on relevance of state characteristics (example, no claim for completeness)

There are many parameters influencing the relevance of state characteristics during a manufacturing process, e.g., type of product, type of material, type of production process, machinery used, application area of the product (e.g., low-cost or high-quality) and many more (see Figure 35). Ideally, the relevance reflects not only the manufacturing programme, process and operation but the whole product-life-cycle. The set of relevant information is derived for the whole manufacturing programme, not individual manufacturing processes or operations (see Figure 33).

In the following subsection, an approach to identify a set of relevant state characteristics is presented, before the next subsection focuses on the understanding and structure of relationships between product state characteristics, which represent a major challenge for the identification.

4.3.5 Identification of relevant state characteristics

In this section the question of how a set of relevant state characteristics may be identified given the previously illustrated understanding of product state, state characteristics and state transformation. The starting point is in accordance with the quality definition the goal of fulfilling the customer requirements (assumed they are transparent). As stated before, the set of relevant information has to incorporate the whole manufacturing programme (see Figure 33).

The first attempt to identify a set of relevant state characteristics of a manufacturing programme is focusing on deriving it by looking into customer require-

ments/quality of the product and the transformation through processes and operations. As it was mentioned before, there are different categories of state transformation, with the intentional transformation being the targeted one. However, with each intentional transformation a variation of other state transformations goes side by side. Therefore the first target-area of relevant state characteristics is ***state characteristics, which transform at least once during the manufacturing programme***. The reasoning behind this target-area is that the transformed state characteristics are relevant as they translate the manufacturing process effect and the development of the quality parameters.

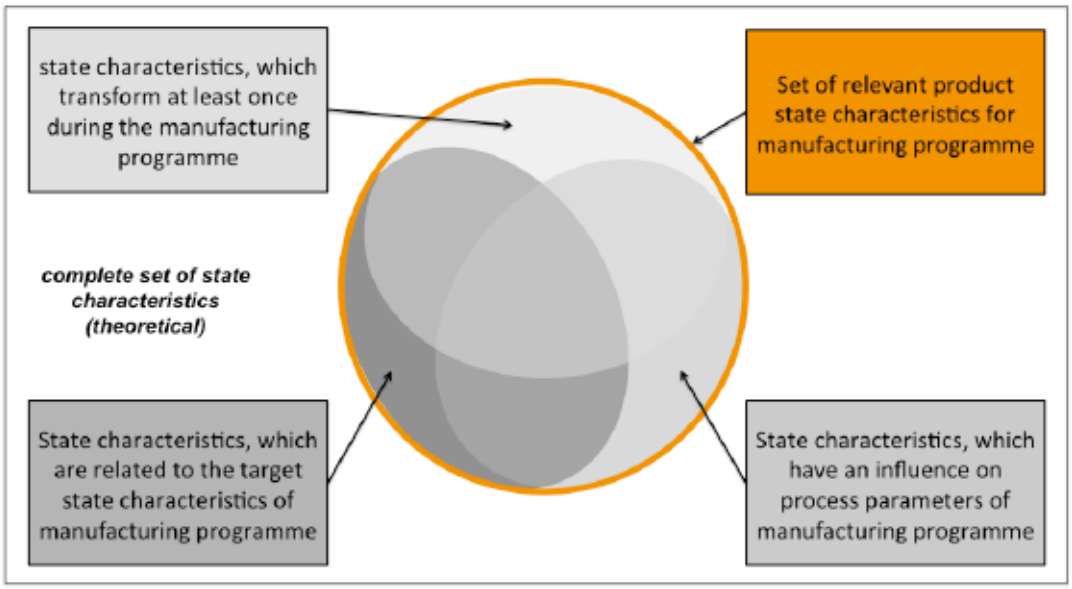

Figure 36: Theoretical framework of the set of relevant state characteristic

The second target-area is connected to relations between state characteristics. ***State characteristics, which are related to the target state characteristics*** are considered relevant throughout the manufacturing programme. The reasoning is, that in order to reach the quality goal and thus the customer requirements, the target state characteristics (and thus target state) have to be met in a way to ensure the functionality expected from the customer. All state characteristics, which are related to these, have an influence on the transformation and the manufacturing programme output.

The third target-area is focusing on the process and its influence on the transformation. ***State characteristics, which have an influence on process parameters,*** are considered relevant. State characteristics are also influenced by process parameters, however, that case is reflected in the first target-area 'state transformation'. This reflects the importance of the process parameters on the transformation of state. As with the other target-areas, the perspective is the whole manufacturing programme and thus cross-process/operation relations.

In Figure 36, the above-introduced target-areas used to identify relevant state characteristics are put into context to each other and all theoretically available state

characteristics of a product throughout a manufacturing programme. The illustration of the figure suggests that the number of state characteristics in the different target areas is of equal number. This is not necessarily the case; the numbers may be of similar size or vary significantly. Also the boundaries between the set of relevant state characteristics are not fixed. As mentioned before, there are several circumstances where there may be a knowledge gap, which can lead to a smaller than theoretically possible selection. Especially given the complex cross-process/operation intra-relations which will also be discussed in the following section. The figure is just a single snapshot in time before the boundaries change due to e.g., newly acquired knowledge about the process or a change in the set up.

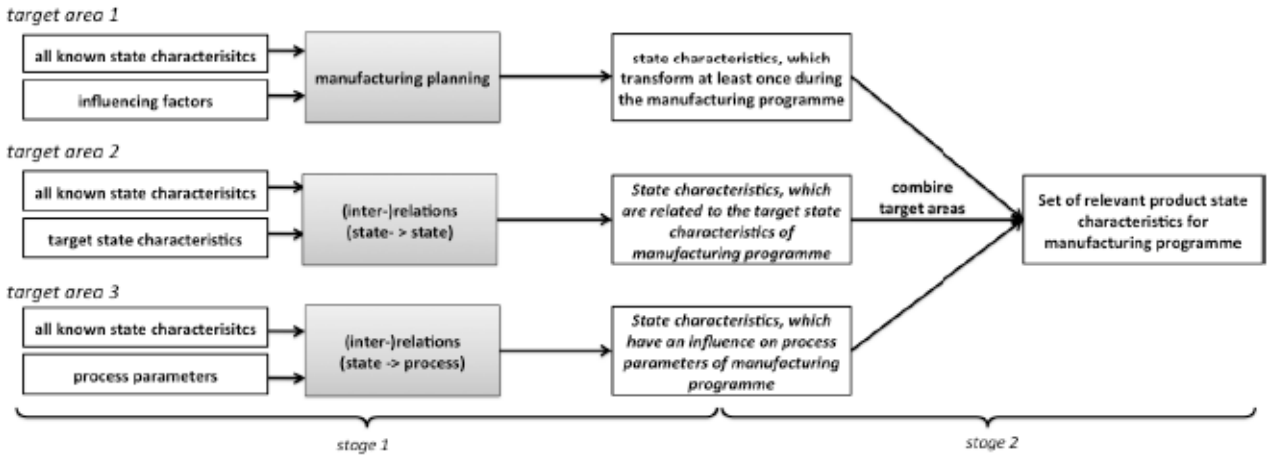

Figure 37: Two-stage process to identify set of relevant state characteristics

To combine the state characteristics of the three target areas towards a set of relevant state characteristics, a two-stage process is envisaged (Figure 37). All known state characteristics are an input in the three target areas during stage 1. In target area 1, the influencing factors are important to determine changing state characteristics. They may be mapped using a modified Ishikawa diagram (Figure 38).

In order to identify a state characteristics that transform during the manufacturing programme, the input state $Xn_{(t=0)}$ and output state $Xn_{(t=1)}$ are compared. If they are not equal of value a transformation of state is assumed and the characteristic is considered relevant in target area 1 (see Annex Figure 120).

Identifying relevant state characteristics according to target area 2 during stage 1 involves the process intra- and inter-relations of the target state characteristics and other state characteristics along the manufacturing programme. This is done without adding a qualitatively (high – low) rating or quantifying the process intra- and inter-relations at this point. Just the existence of a process intra- and inter-relation is considered. (see Annex Figure 121). However, given the existing knowledge gap of e.g., intra-relations (cross-process) between state characteristics, this task may be challenging.

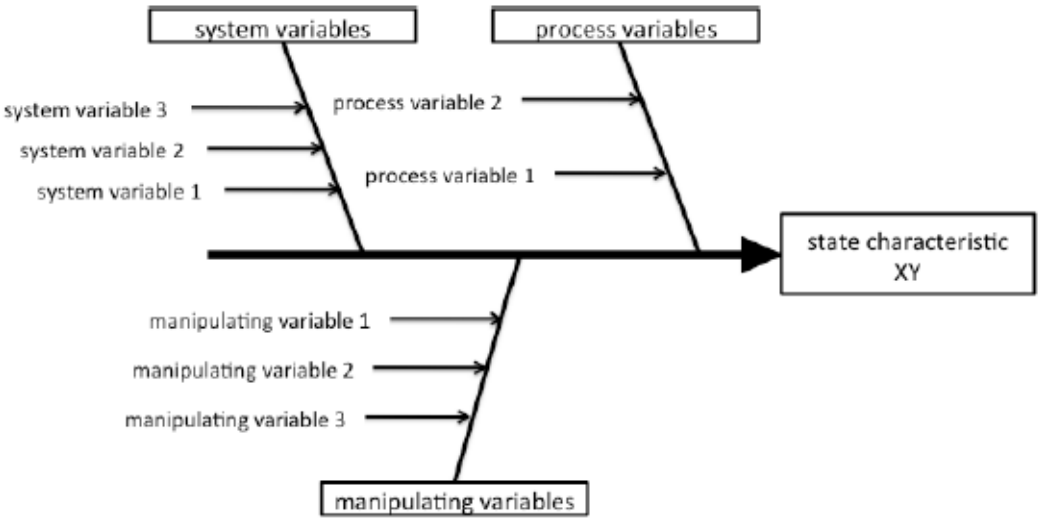

Figure 38: Ishikawa diagram in order to connect influencing factors to state characteristic

Target area 3, identifying relevant state characteristics by their influence on process parameters along the manufacturing programme, is based the existence of process intra- and inter-relations. The difference to target area 2 is that the process intra- and inter-relation is not between state characteristics themselves but between state characteristics and process parameters. Again, given the existing knowledge gap of e.g., process intra- and inter-relation between state characteristics and process parameters, this task may be considered challenging (see Annex Figure 122).

After identifying the relevant state characteristics for the different target areas in stage 1, the identified relevant state characteristics are combined in stage 2 to create a comprehensive set of relevant state characteristics for the specific manufacturing programme and product (see Annex Figure 123).

In this section a preliminary approach to identify relevant state characteristics was presented. By doing so, the importance of the process intra- and inter-relations between states/state characteristics and states/state characteristics and processes/process parameters is highlighted. Understanding these process intra- and inter-relations may be one lever to reach the set goal and increase the number of known relevant state characteristics and thus the transparency of the manufacturing programme. In the following section the process intra- and inter-relations are discussed in more detail before and different possible ways of describing and illustrate them are presented.

It has to be made clear that the identification of relevant state characteristic is not a one-time process, but a continuous effort to create a more complete set of relevant state characteristics describing the product state of a certain product along a specific manufacturing programme.

4.4 Process intra- and inter-relations among state characteristics

In this section, the process intra- and inter-relations of state characteristics are discussed. In this theoretical discussion, the term process intra- and inter-relations is chosen in order to reflect the general nature of the relationship in this context. The differentiation between the general relation, correlation and finally causation (causality) is not in focus here. Process inter-relation focuses on the relationship of state characteristics and/or process parameters within a process/operation, whereas process intra-relation highlights the cross-process relationships of the same kind occurring within multi-stage manufacturing programmes. However, this differentiation comes into focus in the actual application of ML techniques in the following section 5. In the previous section the importance these relations in regard to understanding the mechanics of product state transformation along a manufacturing programme has been pointed out. In a manufacturing system the different components are related others. Thus, relations exist between states and consequently between state characteristics. These relations can be of various form and character, direct or indirect, of importance within a single manufacturing process/operation or just valid when looking at the whole manufacturing programme.

Just, imagine an illustration of relations between a whole selection of relevant state characteristics and processes (process parameters) that provide a solid base for an information management system supporting in process quality control (Dijkman, 2009). The more processes and operations and relevant state characteristics that have to be considered, the more a possible illustration becomes complicated. In other words, the dimensionality of the problem increases as the number of the state characteristics increases. As a result, this will increase complexity instead of helping to increase transparency.

Next, the occurring relations between state characteristics along a manufacturing programme and ways of describing them are presented before visualization of relations within a manufacturing programme are introduced.

4.4.1 Describing process intra- and inter-relations of state characteristics[16]

In this section, the different forms of describing relations between state characteristics during a manufacturing programme are analyzed. At first the difference between an interrelation and a relation are depicted on in the following paragraphs.

[16] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2012c; Knoke et al., 2012)

Figure 39 shows product states and state characteristics (SC) along a multistage manufacturing programme. Product states frame manufacturing processes responsible for the state transformation. The product is described by discrete product state characteristics. The term relation describes the general connections between state characteristics. These relations can either be one-directional (dependent) (see a), c), d) in Figure 39) or bi-directional (interdependent) (see b) in Figure 39). The parameters of the manufacturing processes (e.g., cutting speed, damping pressure) influence the transformation of state characteristics. As shown in Figure 39 the manufacturing processes are framed by preceding and subsequent product states.

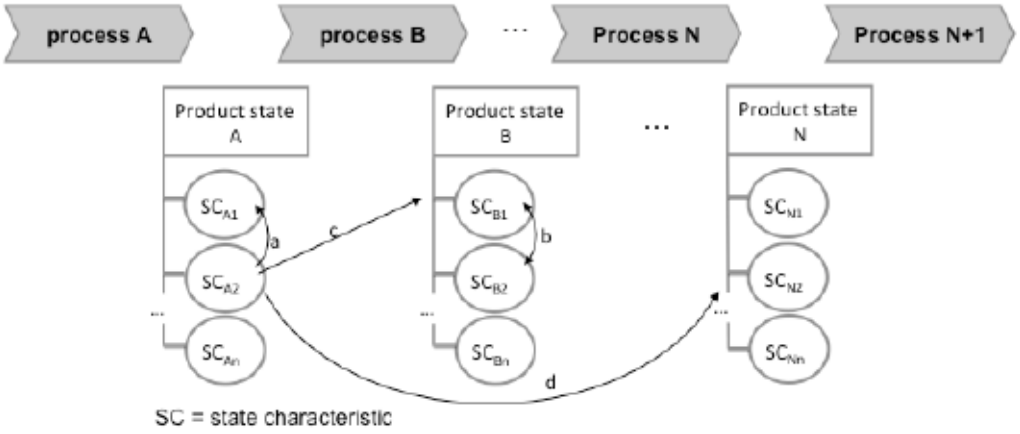

Figure 39: Visualization of process intra- and inter-relations between state characteristics

Interdependencies (see b) in Figure 39) can only occur within a definite product state while dependencies (see a), c), d) in Figure 39) cannot go against the process flow, so any potential shapes of the dependencies and interdependencies can be reduced. This is based on two axioms regarding the temporal restrictions of these connections:

- Dependencies can never go against process flow, since a state characteristic always has an existing value that only past or present effects can influence.
- Interdependencies can only exist between state characteristics of the same state and time, since a future effect cannot impact the past.

If a decision within the manufacturing process is considered because of an upcoming event, it is in fact not influenced by the future event but by expected requirements and other information existing at the present time of the decision. For example: A car within a manufacturing process is painted red not because a customer is expected to react positively to this specific color at the moment of exchange, but because he had ordered a red car in the past, and this information was already useable during the manufacturing process.

As described before, a state characteristic is dependent on state characteristics of previous states. These cross-state, and thus cross-process/cross-operation relations

can add up and may become increasingly complex. From an analytical perspective, the relations of state characteristics may theoretically be characterized as mathematical functions. For example, the dependency of a state characteristic SC1 on another state characteristic SC2 is expressed in the term SC1 = f (SC2). If interdependency be-tween these two state characteristics exists, they are described by a common function f (SC1, SC2). These functions can be described either by a mathematical term (e.g., the mass of a cylinder: $m = \rho * l * d^2 * \pi$) or a text (e.g., the overall error ratio is 3% in the dayshift and 5% in the nightshift).

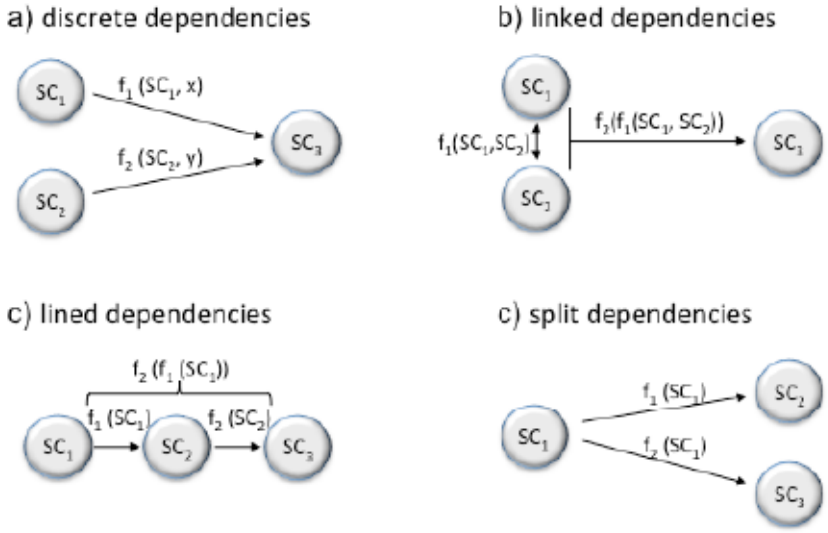

Figure 40: Different forms of dependencies between state characteristics

If dependencies between three or more state characteristics exist, four different characteristics can be identified. These types are visualized in Figure 40. In complex models, these types may appear in combination:

State characteristics with ***discrete dependencies*** (see a) in Figure 40) have independent influence on another state characteristic. This occurs on the condition of additional process parameters (x,y). Since SC3 within the functions SC3 = f1(SC1) and SC3 = f2(SC2) could be eliminated, therefore f1(SC1) = f2(SC2) would imply a direct connection. This causes the need of additional process parameters, which influence each function SC3 = f1(SC1, x) and SC3 = f2(SC2, y). ***Linked dependencies*** (see b) in Figure 40) are another form of the connection between state characteristics. In this case, the combination of two or more state characteristics impacts another. If two state characteristics SC1 and SC2 influence SC3 within a linked dependency, they share an interdependency f1(SC1, SC2), and SC3 can be described by the common function SC3 = f1(f2(SC1, SC2)). The sequence of multiple dependencies is defined as ***lined dependencies*** (see c) in Figure 40). If the dependencies SC2 = f1(SC1) and SC3 = f2(SC2) exist, they can be merged into a function SC3 = f2(f1(SC1)). Finally a state characteristic can also influence two or more other state characteristics. These ***split dependencies*** (see d) in Figure 40)

share a common origin and impact different state characteristics. E.g., the functions SC2 = f1(SC1) and SC3 = f2(SC1).

If three or more state characteristics share interdependencies, they may theoretically be described by a common function. Following this approach, the visualization of all connections is redundant and can be replaced by a chain of interdependencies, as shown in Figure 41. This may significantly simplify a model.

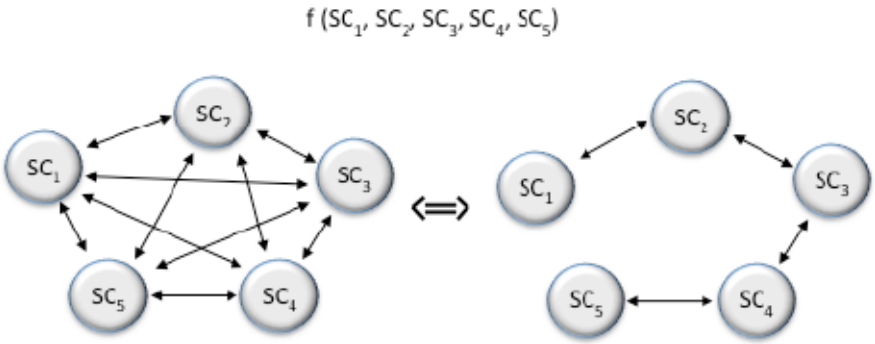

Figure 41: Optional visualization possibilities of multiple interdependencies

4.4.2 Visualization of relations[17]

In this section, the development of a visualization model of existing relations between state characteristics in a manufacturing programme is presented. The resulting three-layer model is introduced, looking at the whole manufacturing programme, the process and the individual state characteristic. Based on this, the development of the visualization approach is briefly discussed. These results are the basis for the following discussions of limitations and challenges of the visualization approach and consequently of the theoretical *product state concept* introduced to this point. The findings of this discussion are accordingly used in the concluding sub-section to derive requirements towards finding a suitable method able to handle the challenges and limitations of the theoretical approach.

[17] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2012c; Knoke et al., 2012; Wuest et al., 2014b)

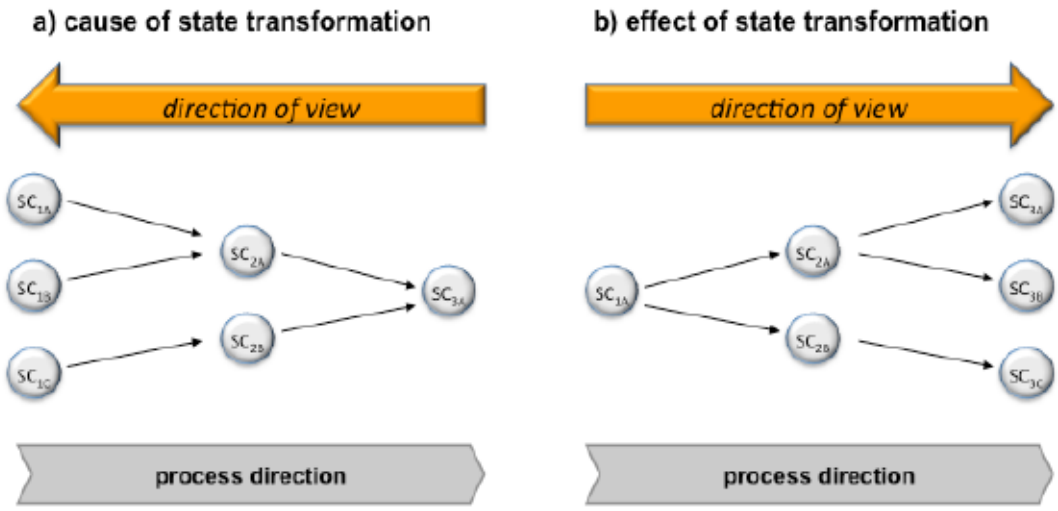

Figure 42: Theoretical application of a modeling of relations between state characteristics depending on the direction of view

The visualization is based on the previously introduced concept of linked state characteristics. This structure of the linked state characteristic provides two different approaches for application, both with a different perspective and goal. Whenever changes within a manufacturing process occur or have to be implemented, the model of state characteristic relations, when transferred to a manufacturing programme, can be applied. If the value of a state characteristic exceeds the acceptable range, the system might be used to create a model with all relevant influences on the state characteristic to identify the problem ('cause') (see a) in Figure 42). Alternatively, if a process parameter has to be changed, the visualization might be used for the opposite purpose: providing information about the 'effect' of the change (see b) in Figure 42).

Next, the basic principles of the *product state concept* is translated into a three-layer visualization before the application and benefits of the application of graph theory on the findings is briefly discussed based on the developed visualization model. First, the theoretical development and foundations of the complete three-layer model and subsequently each individual layer is presented. In order to put it into perspective, a short industrial example of the application of the model is discussed, which provides a first impression of the limitations, challenges and shortcomings of reaching the set goals in industrial practice of the theoretical *product state concept* modeling approach. These challenges and limitations are discussed in greater detail in the following subsection as a basis for the definition of requirements for the solution to-be developed.

The visualization model is developed following guiding modeling principles (Becker, 1998). Even though, these guidelines are sometimes criticized for its partly subjective criteria (Heinrich, Heinzl & Roithmayr, 2007), they are established as a supporting and guiding framework for process modeling in different domains

(Kobler, 2010). The six main principles are: correctness, relevance, economic efficiency, clarity, comparability and systematic composition (Rosemann & Schütte, 1997; Becker, 1998; Batini, Ceri & Navathe, 1992; Becker & Schütte, 2004). For further detail refer to Annex section 9.3.1.

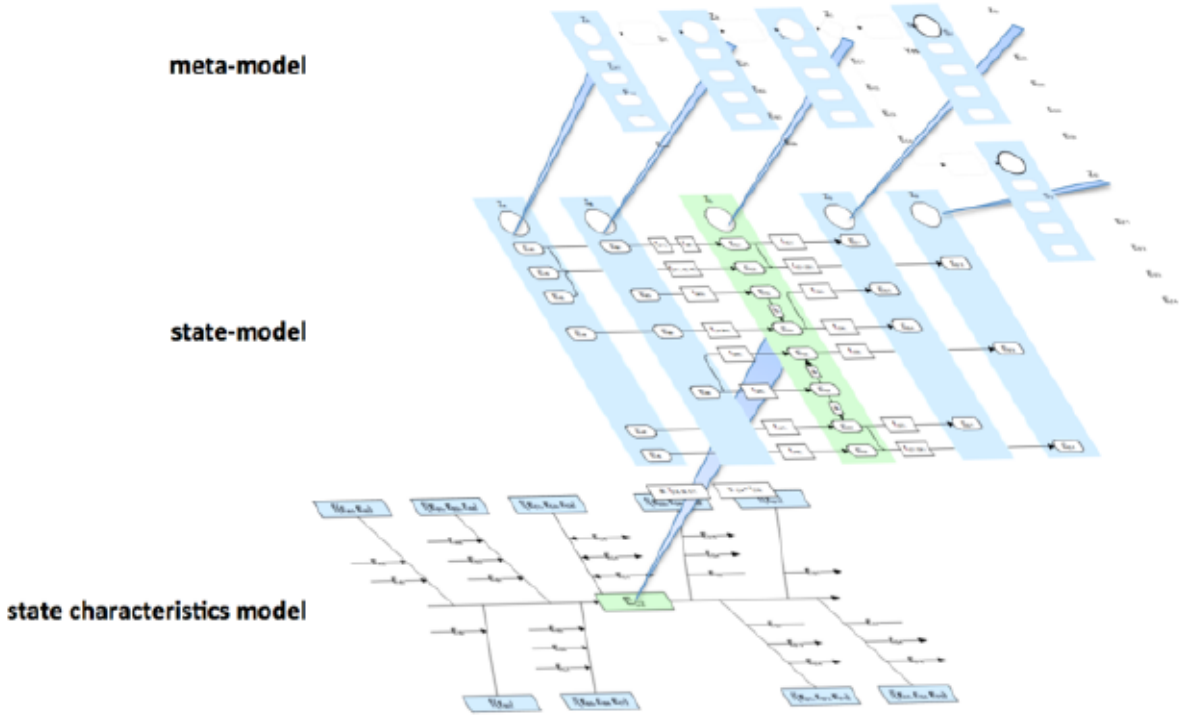

Figure 43: Relation of different model layers

Besides the principle of clarity, the information density has a major impact on the systematic composition. The information density depends on the amount of to-be illustrated information (e.g., knots, edges, objects, notations) and the spatial dimensions of the model. As cutting information is not always possible without jeopardizing the goal of the model, the stated challenges may be faced by applying a cascading of information in respect of different diagrams with different levels of detail (Sarich, Schutte & Vanden-Eijden, 2010). The visualization of the theoretical *product state concept*, contains the available information about the relation between state characteristics. Additionally is includes process parameters within a manufacturing programme. Therefore, the model may most likely become very complex and difficult to handle. To approach this issue, a model with three different hierarchical layers and accordingly, levels of detail may support the applicability of such an attempt by cascading, according to the above stated principles.

One possible approach is to split the model into a meta-model and two sub-models (see Figure 43):

- A ***meta-model*** that provides a general overview on all states and process steps with the aligned process parameters and state characteristics, along with the general process structure.
- A ***state-model*** that focuses on the relations of a single state or process step, and shows the relations of all process parameters or state characteristics of the focal state or process step.
- A ***state characteristic-model*** that visualizes all relations of a single state characteristic or process parameter, and may include the functions that describe its relations.

In the following elaboration, the focus is laid on the meta model (layer 1) as the main visualization option within the *product state concept* at this point. For further details on the state model (layer 1) and the state characteristics model (layer 3) refer to the Annex section 9.3.2.1 & 9.3.2.2.

In accordance to the principle of systematic composition, a meta-model is chosen to ensure the abstract overview of the model and its purpose/goal. The goal of the ***meta-model (layer 1)*** is to provide an overview over the manufacturing programme structure and connect relevant state characteristics to the different states along the manufacturing programme. Additionally the meta-model is illustrating existing (known) process intra- and inter-relations. The visualization form is based largely on the BPMN (OMG, 2010). This modeling annotation's strength is the clarity and intuitive nature of its symbols. However, the focus is laid not on the processes itself but on the product states before (input) and after (output). As the illustration of states is not represented in the original BPMN model, the symbols of 'events' are used, with the result of a bipartite graph (Weisstein, 2011).

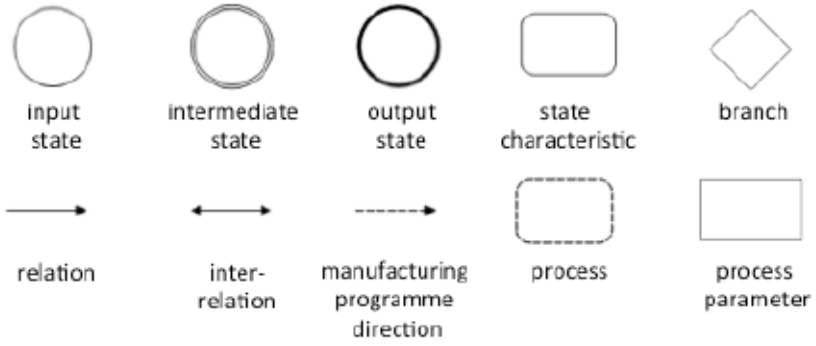

Figure 44: Symbols used in meta-model

The goal and purpose of the meta-model is mainly to provide an overview of the structure of the manufacturing programme and its processes/operations with the process intra- and inter-relations between the state characteristics in the focus. Therefore, the activities and edges of the processes are represented in broken lines (see Figure 44). Through this alternative symbol, the original symbol for activity of

the BPMN model can be applied to the state characteristics in the focus and the process intra- and inter-relations can be illustrated through arrows. The operators for branch and merge can be applied accordingly.

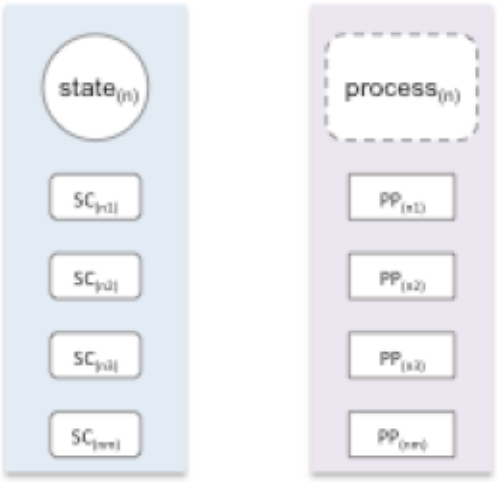

Figure 45: Illustration of state/state characteristics and process/process parameters

The allocation of State Characteristics (SCs) and product states (states) is realized through a rectangular frame in the background. In this context, a colored contour is preferred to a framed contour as the additional lines may add to the inherit complexity within the model and the arrows. The size and position is dependent on the number and position of the state characteristics to be contained (see Figure 45 left).

In case Process Parameters (PP) impact can be directly mapped to SCs, the processes can be visualized in similar form (see Figure 45 right). In order to distinguish product states and processes, a deviating color scheme should be chosen.

In Figure 46, the previously theoretically discussed different forms of dependencies between state characteristics (see Figure 40) are visualized by the above stated principles of the meta-model. In this figure, relations (dependencies) are visualized by a single headed arrow, whilst inter-/intra-relations (interdependencies) use two headed arrows as a representation. In case the edges (arrows) have to overcome vertical levels, this shall be accomplished within the colored contour representing the state/process (see a) in Figure 46). It is important to distinguish between connected and independent relations (dependencies). If the dependencies are independent, this shall be highlighted by including the number of overlaying relations (dependencies) (see c) Figure 46 and Figure 47). In case a relation (dependency) skips a state, the representing arrow shall be directed underneath the colored contour of the skipped state (see Figure 47). Thus, even with overlaying relations (dependencies) the distinct meaning of the illustration is ensured, highlighting the independence of the relation (dependency) again with a number.

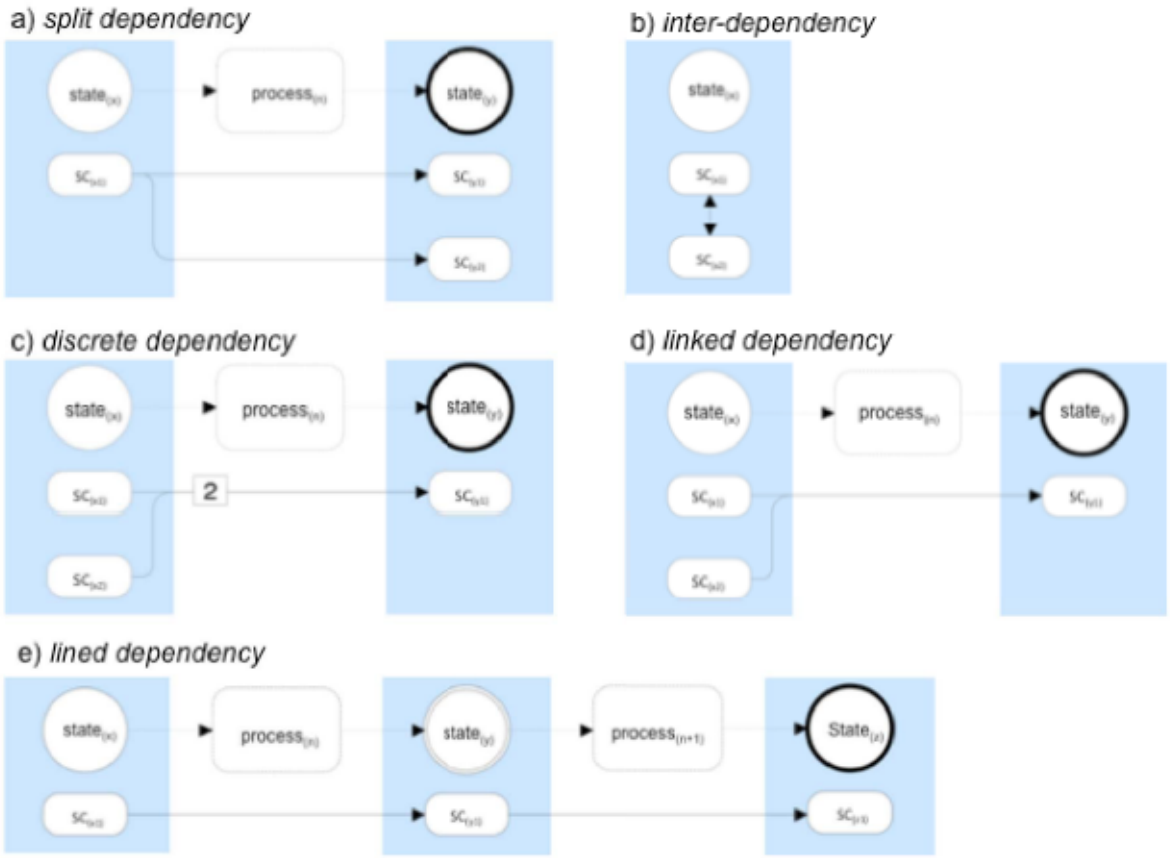

Figure 46: Selection of occurring dependencies within the meta-model

With increasing complexity it may be sensible to exclude the (inter-/intra-)relations ((inter-)dependencies) in the meta-model, reducing the meta-model purpose to providing only an overview of state characteristics, states and processes.

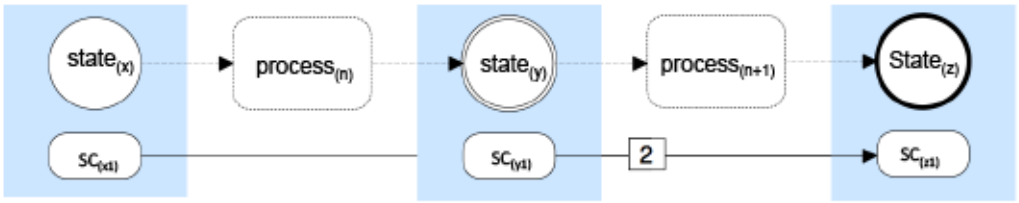

Figure 47: Illustration of independent dependencies with skipped state

In Figure 48 an exemplary manufacturing programme is illustrated applying the developed modeling annotations of the meta-model. The process structure is represented by the broken lined symbols and edges and the states accordingly by the previously introduced symbols. The color contours in the figure connect the state characteristics to the existing states and the process intra- and inter-relations are chosen randomly in this example.

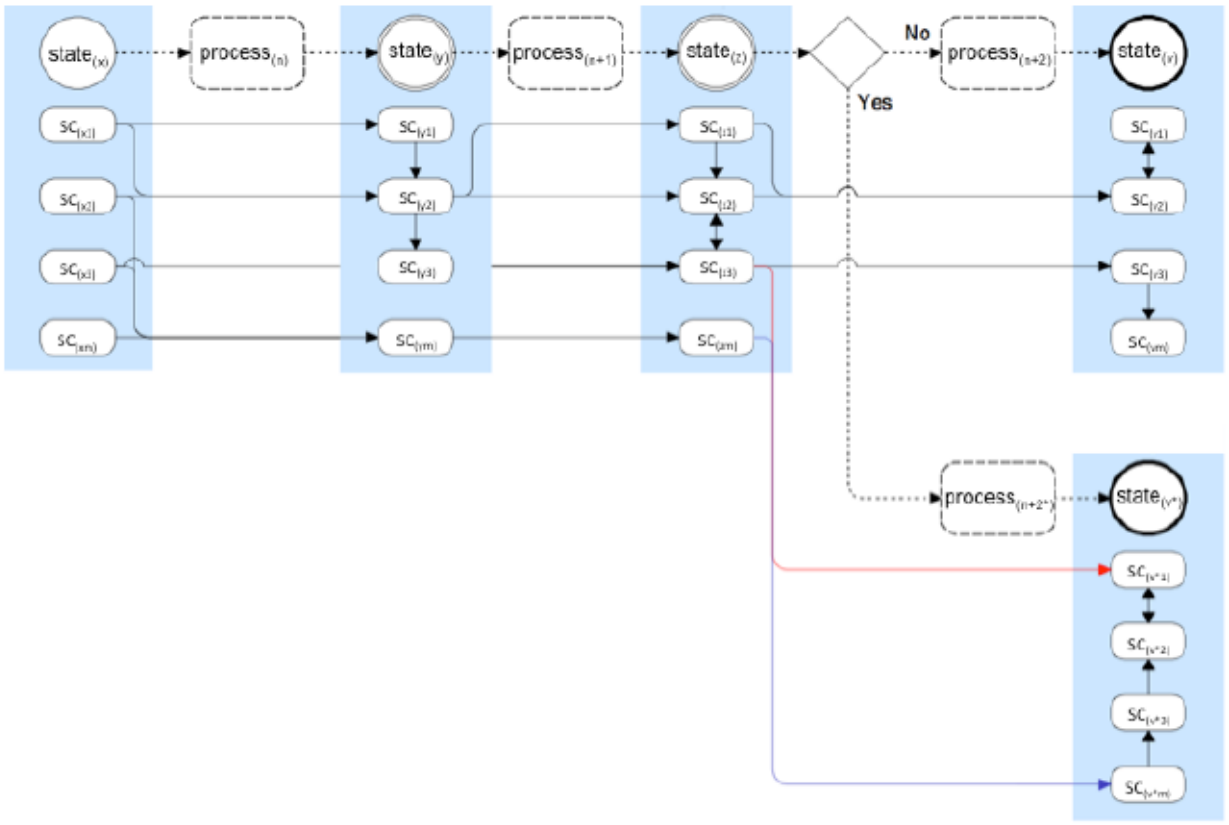

Figure 48: Exemplary illustration of meta-model (layer 1)

It can already be observed that even in this comparably simple model with just four state characteristics connected to each state with a reduced relation density, that the illustrated edges cannot be distinctly assigned without additional coloring (compare $SC_{(z3)}$ to $SC_{(v*1)}$ (in red) and $SC_{(z4)}$ to $SC_{(v*4)}$ in Figure 48) or that a relation, skipping a state, crosses other edges (compare $SC_{(x3)}$ to $SC_{(z3)}$ in Figure 48). This highlights that the main purpose of the meta-model may mainly be to providing an overview over the manufacturing programme, positioning of states within and the assignment of state characteristics to states. For more detailed modeling of the process intra- and inter-relations within the manufacturing programme a sub model needs to be developed.

After the meta model and its development was described previously, next a brief industrial example of the application of the theoretical product state framework and the developed model is presented. The industrial case is the manufacturing programme of a SME, producing products for the automotive industry (1st tier supplier). The main value adding processes are machining (turning and milling) and subsequent balancing. The company produces a large amount of identical products with high requirements towards product quality, as it is common in the automotive supply chain. The manufacturing programme and its processes are organized mainly in a job-shop production with a quality check at the end.

Before applying the visualization model, the process intra- and inter-relations and relevant state characteristics have to be analyzed. This demonstrates the first major limitation of the modeling approach in industry: the transparency and knowledge requirements towards the programme, the processes and the products are very high from the beginning. As in the case study, not all required information was available to the process owner, the modeling was undertaken under this information mismatch constraint. The goal was to analyze the manufacturing programme with its three processes and four states and create a transparent visualization based on the developed modeling approach with the knowledge available from the process owner and existent in literature.

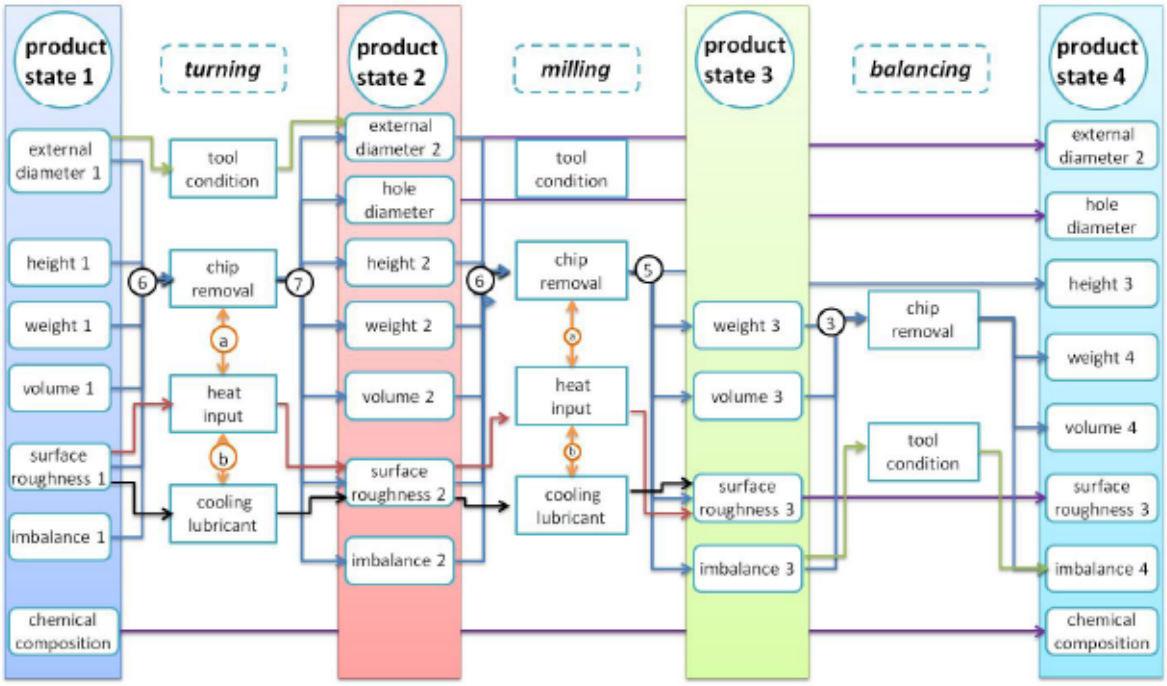

Figure 49: Exemplary ind. appl. of meta-model (adapted from Wuest et al., 2013b)

The result of the case analysis is depicted in Figure 49. Even though limited knowledge about relevant state characteristics, process influence and existing process intra- and inter-relations was available, the visualization became very complex from the beginning. Applying the different layers of the three-model was partly a failure due to the missing knowledge and thus difficulties to create a comprehensive model at the different levels. Especially the transfer functions descriptions were not available and thus the model was adapted to the local situation at the case company as can be seen in Figure 49. The model was reduced and mainly the meta-model notation was applied.

Taken the above stated situation into account, the derived visualization is still very complex and it is questionable if it is clearly understandable. Due to the knowledge gap and thus incomplete nature of the model, its contribution towards increasing transparency of the manufacturing programme is questionable as well. The im-

portant finding within these limitations around the knowledge gap is, that just very few companies actually possess this kind of knowledge about their processes and products. The large majority, especially SMEs, will face similar challenges, as did the company in this case study.

However, the following findings are worth mentioning despite the problematic nature of the case. The modeling approach is largely based on modeling process intra- and inter-relations, which could be either between state characteristics within a state or in between different states along the manufacturing programme. As has been found, to model these relations in this context, large volumes of product and process data/information are needed beforehand and thus make this approach impractical, at best. Not all relations are known as discussed in previous chapters, not to mention a possible quantification of the transfer function. Furthermore, there are still various correlations unknown, especially when the whole manufacturing programme is in the focus (system view) as it is the case here and not just a single manufacturing process/operation. An example for such knowledge about correlation is the field of distortion engineering where after years of research still not all mechanisms are identified (Zoch, 2012).

Much research and individual testing/experiments would be needed to get a first, partly satisfying result. At the same time this makes the approach inflexible, time-consuming and vastly expensive, thus, not applicable in a fast changing environment like industrial manufacturing. Theoretically, assuming that most dependencies, basically cause-effect relationships, between states or state characteristics are known, it still is very resource intensive to integrate them all in a model (see Figure 49). The feedback of the application was, that even so some knowledge gain and awareness was raised within the company, the process of deriving the relations and inter-/intra-relations and cumulating them in a visualization model was too time-consuming and prone to failure. As soon as one parameter of a process within the manufacturing programme is changed or the product itself or even the environment changes, the whole model may have to be redeveloped. And, if a company has different production lines with different products (variants), the model has to be developed for each individual product/production line.

The combination of unknown process intra- and inter-relations, fast increasing complexity and high-dimensionality in modern manufacturing and the time-consuming and resource binding process are not reflected adequately in the achievable results of the visualization model as presented above. These are some of the major limitation of the application of the theoretical *product state concept*, which will be derived in detail later within this section.

Within the development of the *product state concept* the utilization of graph theory as a suitable addition was analyzed. However, similar to the previously introduced visualization, graph theory application within this context has certain limitations.

The main limitation of applying graph theory in this context is the required level of understanding and knowledge about the manufacturing programme, its products and processes in greater detail. Unknown state characteristics and especially unknown process intra- and inter-relations may jeopardize any beneficial findings from the beginning as the initial model does not reflect the 'real world' in the detail needed. Another limitation is the effort needed to model all sub-graphs, depending on the complexity of the manufacturing programme, there can be a large number of sub-graphs needed, and the subsequent application of the algorithm. The algorithm design itself is a challenge but that is more a technical than a systemic one. Partly based on the first limitation raised here, the challenge to obtain an accurate set of (manufacturing) data for the purpose of modeling the manufacturing system is more of a general challenge for most analysis in manufacturing domain. In this case the added difficulty is the partly missing knowledge about what data is really needed. For a more detailed presentation of the graph theory application within the *product state concept*, please refer to (Wuest et al., 2014b).

In the next sections the limitations and challenges of the *product state concept* including its visualization, partly presented previously, will be summarized and elaborated in detail as a basis for the development of requirements towards a solution in this respect concluding section 4.

4.4.3 Limitations of describing process intra- and inter-relations[18]

This section will focus on the findings of the previous sections on the limitations and challenges of describing process intra- and inter-relations and, indirectly, of the theoretical *product state concept* are presented.

Modern multi-variate systems have considerable complexity with high-dimensional data (Apley & Shi, 2001; Zhang & Wang, 2009) with unknown or unclear cause-effect relationships in the process(-es) and with non-Gaussian data distributions, at times exhibiting seeming chaotic behavior (Chou, Polansky & Mason, 1998; Borror, Montgomery & Runger, 1999; Stoumbos & Sullivan, 2002), categorical (Wang & Tsung, 2007) or mixed (categorical and numerical) variables, and numerical data with different scales of measurement. Overall, the complexity and uncertainty inherit in most modern manufacturing programmes, can be considered the major challenge and limitations of the previously presented approach and also most conventional control and (off-line, predictive) scheduling approaches (Monostori, 2002). The *product state concept* and the visualization modeling have an inherent high complexity and high dimensionality (in this context high-dimensionality is understood as a multidimensional system with a large number of dimensions) (Suh, 2005; Lu & Suh, 2009; Elmaraghy, Elmaraghy, Tomiyama &

[18] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2013b)

Monostori, 2012). Optimization tools in this field need to be able to handle a large number of dimensions and variables in order to be useful in practice. And a monitoring technique without assumptions on the parametric forms of distributions is important in this context (Monostori, Váncza, & Kumara, 2006).

The importance of understanding process intra- and inter-relations of the product state for the theoretical *product state concept* was highlighted multiple times. However, even though the previous approaches were mainly just looking at the pure existence of process intra- and inter-relations, the identification is already complex. Given the experience with visualization and application of graph theory, the need for a more detailed description, be it quantitative or functional, of the process intra- and inter-relations adds to that already existing complexity. Determining if a relation is a correlation (or even a causal relationship) is knowledge hardly available, difficult to analyze but rather important to the concept.

Another limiting factor towards the successful application of the theoretical *product state concept* in practice it the matter of describing, mapping and illustrating the process intra- and inter-relations along a manufacturing programme is time consuming and may be considered un-flexible (Kano & Nakagawa, 2008). With each change occurring within the manufacturing programme, the visualization model needs to be adapted or at least checked for applicability. Taking into account the agile nature of modern manufacturing, manufacturing programmes change often and quickly, with constantly changing products (states), processes (parameters) and environmental conditions highlight the challenge this issue presents for practical application under financial and time restrictions. After all, "the performance of manufacturing companies ultimately hinges on their ability to rapidly adapt their production to current internal and external circumstances" (Monostori, 2002). The future solution ideally possesses or at least supports the adaptability and flexibility needed in today's manufacturing environment, maybe even in a (partly) automated processes.

A major limitation is the lack of ways to identify process intra- and inter-relations of importance, e.g., by adding a rating of the impact on the output quality. Especially as this is fundamental to the model generation and thus the generation of feasible results. The concept offers no support in identifying process intra- and inter-relations to this point. Availability of knowledge about existing process intra- and inter-relations is a pre-requirement. However, given the existing knowledge gap, this presents a major challenge when it comes to the applicability of the approach. This challenge is shared with many first principle models trying to explain why product quality issues appear. Kano & Nakagawa (2008) present an example for the steel industry, where "the relationship of operating conditions to product defects such as surface flaws and internal defects is not clear. The product qualities have been usually maintained by skilled operators on the basis of their experience and intuition. Although much effort has been devoted to clarify the relationship be-

tween operating conditions and product quality, the problem remains unsolved." Even though, the *product state concept* is not alone facing this challenge, a future solution should include ways to reduce the reliance on existing knowledge about process intra- and inter-relations and ideally provide help in identifying existing process intra- and inter-relations continuously throughout the manufacturing programme run.

It was mentioned before, a major challenge of most monitoring and control approaches in manufacturing is obtaining an accurate (quality) and sufficient (amount) set of (manufacturing) data. While in theory the issue of having too much data and thus having trouble to handle it is valid, in manufacturing practice, the lack of enough data represents a major issue. The reasons are manifold, starting by on a first account profane arguments like security issues of companies, over technical reasons to financial reasons, arguing that to obtain the data is too costly given the estimated benefits. Often the only possibility to apply a model or concept is to use the available data. Therefore, the future solution ideally makes use of all available manufacturing data of the manufacturing programme.

The whole theoretical *product state concept* is built around the relevant state characteristics and the process intra- and inter-relations to each other over the whole manufacturing programme. While the arguments for such an approach are strong in theory, given that precise first principle models are considered "the most reliable approach to quality monitoring" (Kano & Nakagawa, 2008), in today's practice, with the limiting factors like knowledge gaps, complexity and high dimensionality as described previously, there are many challenges hindering the successful reach of the set goals with the theoretical concept.

Overall, describing relations between state characteristics along a manufacturing programme is very complex and, if applied in industry, requires in-depth understanding, high levels of product and process knowledge and a high transparency of product, process and effects in order to realize its potential. Theoretically, if an application of the approach is possible, it may help to increase the final product quality and process efficiency by supporting the early identification of problems and allocation of in-formation to the right addressee. However, establishing such a model is binding significant resources if possible at all due to knowledge available. Furthermore, it is not very flexible which does not help justifying the resources needed to create it. Therefore, other means of identifying, evaluating and utilizing product state relations and their impact on product and process quality are needed.

Based on these findings it has to be researched how the knowledge gap can be attacked despite the limiting factors of high-dimensionality and complexity within the product state framework. Ideally a suitable methodology should contribute to solve or at least support the development of the *product state concept* in the theoretically sketched way, e.g., helping decrease the knowledge gap of state character-

istics and process intra- and inter-relations between states and state characteristics. Following, the requirements of such an approach and specific goals are discussed by first looking into the NP complete nature of the problem, followed by a discussion of the suitability of ML techniques based on the requirements.

4.5 Requirements of state driver identification

In this section the requirements for a future solution are discussed based on the previously determined challenges and limitations. Before the subsequent presentation of the suitability of ML techniques for a future solution approach, the NP completeness of the *product state concept* and its fundamental description of state/state characteristics process intra- and inter-relations are analyzed. However, firstly the main requirements towards a future solution approach and their corresponding limitations and challenges are summarized in Table 2.

Table 2: Limitations & challenges and resulting requirements of theoretical product state concept

Limitations/Challenges	Requirements towards future solution approach
High complexity of modern manufacturing programmes leading to *high-dimensionality* of data (in this context high-dimensionality is understood as a multidimensional system with a large number of dimensions)	Ability to handle high-dimensional problems and data sets with reasonable effort.
High complexity of modern manufacturing programmes leading to *multi-variate* nature of data	Ability to handle multi-variate problems and data sets with reasonable effort.
Lack of transparency and thus benefit for process owners once the model is developed as of the complex visualization of the model	Ability to reduce the possibly complex nature of the results and present transparent and concrete advice for practitioners (e.g., monitor state characteristic XX and process parameter YY at checkpoint ZZ)
Fast *changing processes/inputs* (process parameters/input parameters/products/environment) require constant adaptation and remodeling.	Ability to adapt to changing environment with reasonable effort and cost. Ideally a degree auf 'automated' adaptation to changing condition.
Knowledge gap of correlation/causality from the beginning.	Ability to further the existing knowledge by learning from results.
Limited availability of relevant information (manufacturing data) and unlikelihood of companies to initially add extra measuring points.	Ability to work with the available manufacturing data without special requirements towards capturing of very specific information at the start.
Describing process intra- and inter-relations and ideally correlation or causality towards each other is required for a successful model	Ability to identify relevant process intra- and inter-relations and ideally correlation and/or even causality towards each other.

After presenting an overview of the limitations and challenges of the theoretical *product state concept* and the requirements for a future solution approach, the next subsection will focus on the NP complete nature of the *product state concept*.

4.5.1 NP complete nature of product state concept

Given the many challenges and limitations and the resulting requirements towards a future solution approach, in this section the NP (nondeterministic polynomial ti-

me) complete nature of the optimization problem is discussed. Furthermore, examples of developed approaches handling similar limitations and challenges are briefly presented, indicating towards a possibility of utilizing AI and ML methods.

Based on the inherent complexity and high dimensionality of todays manufacturing programmes, the question is, if today's monitoring problems can be defined as NP-complete problems of complexity theory. Introduced by Cook (1971) in the paper "The complexity of theorem-proving procedures", the NP-completeness theory describes that NP-complete problems turn out to be not solvable in polynomial time. No efficient algorithm has been found to prove the contrary. The scientific majority believes that there is no such algorithm.

Lewis et al. describe a NP problem as a problem for which a valid (in polynomial time solvable) solution is being sought, but which may be solved (using the algorithms known so far) only in exponential time (Lewis, Horne & Abdallah, 1996). The traveling salesman problem is a classic example of a NP-complete case. Angel uses the example of the traveling salesman to show that local search algorithms in manufacturing are NP-complete (Angel & Zissimopoulos, 1998).

A literature review shows that many combinatorial optimization problems in production (Nearchou, 2010), such as scheduling, can be seen as NP-complete problems (Baker, 1988). The same can be said about issues in resource allocation (Udo, 1992), as well in flow-line and job-shop protocols (Lewis et al., 1996). Crama and Klundert provides additional evidence for the NP-completeness of scheduling issues, in terms of lower and upper bounds of processing windows (Crama & Klundert, 1997). NP-completeness can be thus also found in assembly line balancing (Nearchou, 2010) and in the just-in-time manufacturing (Baker, 1988). Tiwari, Patterson & Mabert cover NP-completeness at the organizational level and addresses specifically the distribution of tasks (Tiwari et al., 2009). However, Lewis et al. shows that reentrant flow protocols are solvable in polynomial time, which is one of the few exceptions in production-issues (Lewis et al., 1996). In order to handle such NP-complete problems, different algorithms and ML tools have been applied and shown promising results. So are genetic and biologically inspired algorithms are presented as a possible tool to overcome the hurdles of optimization problems in production (Aytug, Khouja & Vergara, 2003; Pham & Afify, 2005; Laili, Tao, Zhang & Ren, 2011; Ponsignon & Mönch, 2011; Shetwan et al., 2011). Additionally, e.g., Brun treats the issues of self-assembly and their NP-completeness using DNA-computation (Brun, 2007).

It has been argued that the limitations and challenges the theoretical *product state concept* faces indicate the NP complete nature of the optimization problem and thus the inability of first-principle models, like the visualization model to reach tangible solution (Kano & Nakagawa, 2008). Mapping complexities and relationships as they occur within this context is practically impossible and analyses with-

out a real multi-variate approach has little to no chance of success. Therefore, in the next section, the suitability of techniques, known for their ability to handle such issues in many cases, namely AI and ML is discussed.

4.5.2 Suitability of machine learning methods

Before looking into the suitability of ML based on the previously derived requirements towards a future solution approach, the used terms are briefly introduced (for more detailed description see section 5.1). ML is known for its ability to handle many problems of NP-complete nature, which often appear in the domain of intelligent manufacturing (Monostori, Hornyák, Egresits & Viharos, 1998; Srdoč, Bratko & Sluga, 2007).

The application of ML techniques increased over the last two decades due to various factors, e.g., the availability of large amounts of complex data with little transparency (Smola & Vishwanathan, 2008) and the increased usability and power of available ML tools (Larose, 2005). Nevertheless, the main definition of ML, allowing computers (artificial systems) to solve problems without being specifically programmes to do so (Samuel, 1959) is still valid today. ML is connected to other terms, like Data Mining (DM), Knowledge Discovery (KD), AI and others (Alpaydin, 2010). Today, ML is already widely applied in different areas of manufacturing, e.g., optimization, control and troubleshooting (Pham & Afify, 2005; Alpaydin, 2010) (for further details refer to section 5.1).

Many ML techniques (e.g., SVM) are designed to analyze large amounts of data and capable of handling high-dimensionality (>1000) very well. However, accompanying issues like possible over-fitting has to be considered (Widodo & Yang, 2007) during the application. If dimensionality proves to be an issue despite it being unlikely due to the power of the algorithms, there are methods to reduce the dimensions available, which claim to reduce the impact of the reduction of the dimensionality on the expected results (Kotsiantis, 2007; Manning, Raghavan & Schütze, 2009). The importance of using ML, in this case SVM (see section 5.2 for details) is that dimensionality is not a practical problem and therefore the need for reducing dimensionality is reduced. This implies the possibility of being more liberal in including seemingly irrelevant information available in the manufacturing data that may turn out to be relevant under certain circumstances. This may have a direct effect on the existing knowledge gap described previously (Pham & Afify, 2005; Alpaydin, 2010).

Besides the capability of ML to handle high-dimensionality, it is also capable of handling mutli-variate problems and data sets. Examples of successful application include monitoring and control problems based on multi-variate data (Yang & Trewn, 2004). The expected implications are similar to the ones described above, allowing the usage of previously not considered data and information to identify new information and knowledge of the manufacturing programme.

Applying ML in manufacturing may result in deriving pattern from existing data sets, which can provide a basis for the development of approximations about future behavior of the system (Nilsson, 2005; Alaydin, 2010). This new information (knowledge) may support process owners in their decision-making or used to automatically improve the system directly. In the end, the goal of certain ML techniques is to detect patterns or regularities that describe relations (Alpaydin, 2010).

Table 3: Summary of theoretical suitability of ML methods based on derived requirements

Requirements	**Theoretical ability of ML approaches to meet requirements**
Ability to handle high-dimensional problems and data sets with reasonable effort.	Certain ML techniques (e.g., SVM) are capable of handling high-dimensionality (>1000) very well. However, accompanying issues like possible over-fitting has to be considered (Widodo & Yang, 2007).
Ability to handle multi-variate problems and data sets with reasonable effort.	ML is capable of handling mutli-variate problems and data sets. Examples of successful application include monitoring and control problems based on multi-variate data (Yang & Trewn, 2004).
Ability to reduce the possibly complex nature of the results and present transparent and concrete advice for practitioners (e.g., monitor state characteristic XX and process parameter YY at checkpoint ZZ)	ML may be able to derive pattern from existing data and derive approximations about future behavior (Alaydin, 2010). This new information (knowledge) may support process owners in their decision-making or used to automatically improve a system.
Ability to adapt to changing environment with reasonable effort and cost. Ideally a degree auf 'automated' adaptation to changing condition.	As ML is part of AI, and thus be able to learn and adapt to changes, "the system designer need not foresee and provide solutions for all possible situations" (Alpaydin, 2010). Learning from and adapting to changing environments automatically is a major strength of ML (Simon, 1983; Lu, 1990).
Ability to further the existing knowledge by learning from results.	ML can contribute to create new information and possibly knowledge by e.g., identifying patters in existing data (Pham & Afify, 2005; Alpaydin, 2010).
Ability to work with the available manufacturing data without special requirements towards capturing of very specific information at the start.	ML techniques are designed to derive knowledge out of existing data (Alpaydin, 2010; Kwak & Kim, 2012). "The stored data becomes useful only when it is analyzed and turned into information that we can make use of, for example, to make predictions" (Alpaydin, 2010).
Ability to identify relevant process intra- and inter-relations & ideally correlation and/or even causality towards each other.	The goal of certain ML techniques is to detect certain patterns or regularities that describe relations (Alpaydin, 2010).

Given the challenge of a fast changing environment in manufacturing, ML, being part of AI and inherit the ability to learn and adapt to changes "the system designer need not foresee and provide solutions for all possible situations" (Alpaydin, 2010). Therefore, ML provides a strong argument why its application in manufacturing may be beneficial given the struggle of most first-principle models to cope with the adaptability. Learning from and adapting to changing environments automatically is a major strength of ML (Simon, 1983; Lu, 1990).

ML techniques are designed to derive knowledge out of existing data (Alpaydin, 2010; Kwak & Kim, 2012). Alpaydin (2010) emphasizes that "stored data becomes useful only when it is analyzed and turned into information that we can make use of, for example, to make predictions" (Alpaydin, 2010). This is especially true for manufacturing, given the struggle of obtaining real-time data during a live manufacturing programme run with the technical, financial and knowledge restrictions. This may also have an impact on the previously discussed issue of positioning of checkpoints (Wuest et al., 2014c). Whereas, it makes sense to carefully select checkpoints under the perspective of what data is useful, it may be obsolete given the analytical power of ML techniques to derive information from formerly considered useless data. This may result in the ability to determine more states along the overall manufacturing programme. If this is beneficial is an open question, which has to be researched. Given the ability of ML to handle high-dimensionality and multi-variate data, the technical side of analyzing the additional data provides no problem. However, in terms of capturing data it may still be a problem, specifically the ability to capture the data. Once the data is available, determining state drivers in very high-dimensionality situations is not considered problematic, nor is repeating it frequently. Table 3 provides a summary of the theoretical ability of ML techniques to meet the previously derived requirements of a future solution for the *product state concept*.

Overall, as Monostori, Márkus, Van Brussel & Westkämper (1996) emphasize, "intelligence is strongly connected with learning, and learning ability must be an indispensable feature of IMSs". ML provides strong arguments when it comes to the limitations and challenges the theoretical *product state concept* faces. Given the above stated analysis, ML techniques seem to provide a promising solution based on the derived requirements. Most of the requirements are directly addressed positively by ML. However, a more detailed analysis of available ML techniques as well as their strengths and limitations concerning the requirements has to be provided. Most of all, the possible compatibility with the theoretical *product state concept* and its perspective on the manufacturing programme has to be discussed further before a final judgment can be given. Furthermore, there are many questions to be answered like how ML techniques may handle qualitative information.

In the following subsection, the derived research hypothesis is introduced based on the presented findings. This is to be seen on a more conceptual level before a suitable ML technique for the problem at hand is selected and the integration in the *product state concept* is developed and evaluated.

4.6 Derived research hypothesis of the application of ML within the product state concept

One Key Success Factor (KSF) of the application of the *product state concept* as described in the previous section is transparency of the existing process intra- and

inter-relations between states. The nature of these process intra- and inter-relations can be described as resembling a correlation and/or causation/causality. *Correlation* can be defined as the dependence between two random quantities, e.g., X and Y. Kent also argues that the concept of information gain can be used to define a measure of correlation, if the dependence between the two quantities is modeled parametrically (Kent, 1983). Causation/causality on the other hand can be described, in simplest terms, as judging if X causes Y, with these cause-effect relations being fundamentally deterministic (Pearl, 2000).

Ideally it would be possible to select only independent variables thus focusing on causation/causality rather than correlation. However, that is not applicable given the existing knowledge gap, the large number of 'unknown' state transformations and the apparent complexity represented by high-dimensionality and multi-variate nature of the derived data. It is not even possible to identify all existing correlations between states and state characteristics within the manufacturing programme in reasonable time, making it applicable under the efficiency and flexibility constraints modern manufacturing faces today. Adding the existing impact of manufacturing processes and/or operations, environmental factors, etc. on state transformation, the identification of correlations or even causation becomes even more unlikely. In the long run, the differentiation between correlation and causation/causality may be relevant for the further interpretation and identification of state drivers along the manufacturing programme. Under the current circumstances, this objective is not achievable. However, in the long run it may be possible to arrange for the continuous learning of manufacturing programmes thus increasing the likelihood of discovering independent state variables and/or causal mechanism.

However, the previously introduced ML techniques, which allow for the analysis of such problems with similar constraints (limitations and challenges) offer a chance to reach the overall goal of increasing the transparency and increase the knowledge of the manufacturing programme. Applying supervised learning methods like e.g., a multi-variate classification method (SVM), allows to a large extent to include all available variables without having to define and map (known) influence or (inter-/intra-)relation. Using such a method, it should be possible to incorporate existing process intra- and inter-relations, known and unknown, within the analysis implicitly.

Within this research, the goal is to identify state drivers (or 'drivers of state') within a manufacturing programme. State drivers define the weighting of certain process parameters, state characteristics or events, which initiate a state transformation from 'good' to 'bad' along the whole manufacturing programme. State drivers therefore are not only representing process intra- and inter-relations but all influencing factors causing a state to change (transform). Based on that, the main research hypotheses for this dissertation are presented.

1) Hypothesis 1 'Capturing of process intra- and inter-relations by implication through ML'

Modern manufacturing programmes are complex and many process intra- and inter-relations between states, state characteristics and processes/operations cannot be mapped accurately due to increasing complexity and unknown relationships. Hence, it is desirable to capture existing relationships by implication without having to model them and their influences. Through the application of Supervised ML main drivers of the product and process state can be identified throughout the whole manufacturing programme by capturing and utilizing process intra- and inter-relations implicitly by incorporating all available product and process state data. This will positively impact the currently existing knowledge gap by furthering the understanding of the correlation mechanisms within the manufacturing programme.

2) Hypothesis 2 'Adaptability to changing conditions through ML'

Manufacturing programmes are set in constantly changing environments. Changing product and process parameters as well as shifting external influences demand that successful optimization approaches are able to adapt to these frequently changing conditions with minimal effort. Many of today's ML tools are highly adaptable to changing conditions with, at the same time, reduced demands in (computational) resources. By applying ML, changes of product states, process parameters and external factors can be continuously integrated in and thus, eventually reflected in the results of the analysis by contiguously updating the learning data set. This in turn will also contribute to the goal of reducing the existing knowledge gap about the manufacturing programme and its mechanisms by increasing the knowledge of the influence of changing conditions.

The above-presented hypotheses represent the general research direction which follows the identified requirements to bring the *product state concept* despite the identified limitations (e.g., lack of knowledge) to life. In the further course of the next section, the individual hypothesis are detailed further when the particular specification of the chose ML technique are defined (see section 5.3).

Therefore, in the next subsections a suitable ML technique is selected after a general introduction into the topic. Afterwards, a methodology for the application of SVM to identify state drivers in manufacturing programmes is developed. Subsequently, elaborating the evaluation of the application of the methodology, the derived specific research questions originated in the presented hypotheses are discussed in greater detail.

5 Application of machine learning to identify state drivers

It has been established in the previous sections that ML techniques may be generally suitable for the identified challenges of applying the *product state concept*. The successful identification of state drivers taking process intra- and inter-relations into account is essential for the application of the developed concept. To reach that goal, distinct research hypotheses focusing on a promising approach of applying ML within the *product state concept* were derived in the previous section.

In this section, the application of ML is investigated in further detail. First ML is briefly introduced in more detail with respect to the manufacturing domain. Based on this brief general elaboration, SVM algorithms are selected as a suitable ML technique to match the detailed requirements of the stated research problem. In the final subsection, the application of SVM is discussed towards its objective of identification of *state drivers* in manufacturing programmes. Within this last subsection, the application and evaluation approach of the SVM application are presented and the derived hypotheses are detailed based on the decision to use the SVM algorithm to conclude the section.

5.1 Machine learning in manufacturing

In this subsection the application of ML techniques in manufacturing is introduced. ML has been successfully utilized multiple times in various process optimization, monitoring and control applications in manufacturing in different industries (Gardner & Bicker, 2000; Pham & Afify, 2005; Alpaydin, 2010; Kwak & Kim, 2012). ML techniques were found to provide promising potential for improved quality control optimization in manufacturing systems (Apte, Weiss, & Grout, 1993), especially in "complex manufacturing environments where detection of the causes of problems is difficult" (Harding et al., 2006). However, often ML applications are found to be limited focusing on specific processes instead of the whole manufacturing programme or manufacturing system (Doltsinis, Ferreira & Lohse, 2012).

There are many different ML methods, tools and techniques available, each with distinct advantages and disadvantages. The domain of ML has grown to an independent research domain. Therefore, within this section the goal is to introduce ML techniques briefly and select a suitable algorithm for the previously established problem. The purpose is not to develop new tools or techniques to further the ML development. In order to achieve that goal, first, a brief general introduction to ML with regard of manufacturing application is presented. The subsequently identified supervised ML technique, SVM, is then detailed and the rationale behind the selection is discussed in greater detail, relating the arguments to the previously identified requirements.

5.1.1 Machine learning

The topic of ML firstly gained attention after Samuel (1959) published his paper "Some Studies in Machine Learning Using the Game of Checkers". Since then, not only did the research field of ML grow continuously but also it grew more divers. Today ML is omnipresent in our daily lives, e.g., through the use of various google products or certain public transportation systems (Smola & Vishwanathan, 2008). There are several journals specifically targeted to ML research available, some actively publishing continuously for more than 25 years.

The research domain of ML looks into the practice of preparing computers or artificial systems to act or react to certain events without being specifically programmed to do so (Nilsson, 2005; Smola & Vishwanathan, 2008). ML aims to solve (manufacturing) problems by applying knowledge that was acquired from analysis of (data of) earlier problems of similar nature to the to be solved problem (Priore, de la Fuente, Puente & Parreño, 2006). This capability is desirable or may be even necessary in some cases for various reasons. Nilsson (2005) states the following list of some existing reasons for the application of ML. All of those stated reasons can be directly mapped to the research problem at hand in this dissertation:

- Some tasks can only be defined by example *without a complete understanding of the existing relationships* (input – output). Therefore machines are required to adjust in order to succeed in creating correct outputs with given inputs by approximately 'learn' the implicit relationships.
- Possibility of *hidden relationships and correlations* in large piles of data, which ML may provide a tool to extract.
- Products are used in different environments and might not function as desired in some of them. This can be due to limited knowledge of the actual application area at the time of design or due to changing environmental factors, etc. However, *ML may enable products to adapt* to some of those previously not-anticipated circumstances.
- In the age of cheap data storage and sensor technology, it is possible that the *amount of data available exceeds the (economical and technical) ability* of humans to incorporate the contained information and knowledge in the programming. Preparing the artificial system to learn from the available sources independently through ML allows designers to make use of that knowledge.
- The previous point can be extended to *new information and knowledge discovered during the utilization* of a product. By incorporating ML techniques, the artificial system can make use of new developments and thus reduce the need for actively redesign or redevelop an existing system.

As mentioned before, the ML domain compromises several sub-domains and/or closely related domains with large overlaps. For instance, AI represents the theoretical and methodological foundation for learning of systems (Negnevitsky, 2005) and thus being a crucial part of ML. However, not all AI methods inherit the capability of learning (Negbevitsky, 2004). Whereas AI can be seen as the overarching domain, ML being a part of it (Whitehall, Lu, & Stepp, 1990), DM and KD/Knowledge Discovery from Databases (KDD) are more sub-domains describing a certain area of ML techniques. Kotsiantis (2007) identifies DM as the most significant application area of ML. MD and KD/KDD concepts focus on uncover and find hidden knowledge and information from (often large amounts) of available data. The unknown process intra- and inter-relations between states can be seen as such 'hidden knowledge'. DM and KDD have been applied successful to various problems in the manufacturing domain (Kwak & Kim, 2012). The areas of predictive maintenance, fault detection, condition monitoring, QM, operations, etc. are all examples where the ability of DM to identify hidden patterns is receiving increased attention (Harding et al., 2006). Within this dissertation, the term ML is used comprehensively instead of further differentiations in DM or KDD as the presented techniques all incorporate a learning component.

Within this dissertation, the focus will be laid on supervised methods described in more detail in the following subsection. Other methods like Reinforcement Learning (RL) (Wiering & Van Otterlo, 2012) and unsupervised ML (Manning et al., 2009) are not considered further due to the specific nature of the problem. The main assumption is that knowledgeable experts can provide feedback on the classification of states to identify the learning set in order to train the algorithm. However, in some cases this might not be possible or, in the future, desirable.

5.1.2 Supervised machine learning

In manufacturing application, supervised ML techniques are mostly applied due to the data-rich but knowledge-sparse nature of the problems (Lu, 1990). In addition, supervised ML may benefit from the established data collection in manufacturing for statistical process control purposes (Harding et al., 2006) and the fact that this data is mostly labeled. This is true also for the problem described in this dissertation. Basically, supervised ML "is learning from examples provided by a knowledgeable external supervisor" (Sutton & Barto, 2012). Supervised learning is often applied in manufacturing. This is partly due to the availability of a) expert feedback (e.g., quality) and b) the labeled instances. Supervised ML is applied in different domains of manufacturing, with monitoring and control being very prominent among them (e.g., Apte et al., 1993; Pham & Afify, 2005; Harding et al., 2006; Alpaydin, 2010; Kwak & Kim, 2012). The general process of supervised ML is illustrated in Figure 50.

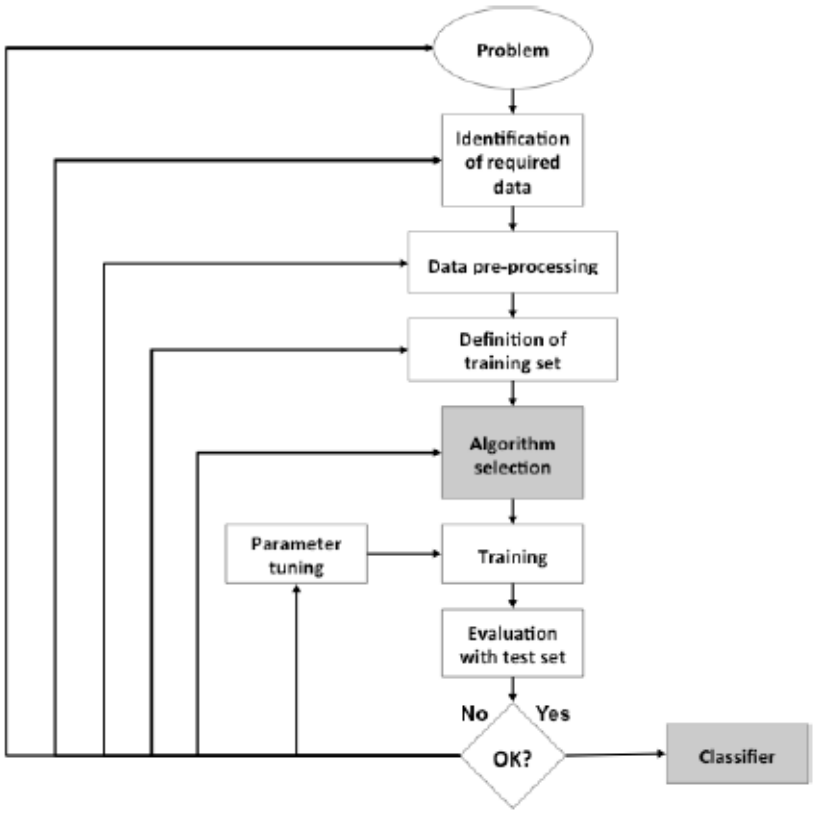

Figure 50: Generic process of supervised ML (Kotsiantis, 2007)

Based on a given problem, the required data is identified and (if needed) pre-processed. An important aspect is the definition of the training set, as it influences the later classification results to a large extent. Even so in Figure 50 it appears that the algorithm selection is always following the definition of the training data set, the definition of the training data also has to take the requirements of the algorithm selection into account. This is to some extent also true for the identification and pre-processing of the data as different algorithms have certain strength and weaknesses concerning the handling of different data sets (e.g., format, dimensions, etc.). After an algorithm is selected, it is trained using the training data set. In order to judge the ability to perform the targeted task, the trained algorithm is then evaluated using the evaluations data set. Depending on the performance of the trained algorithm with the evaluation data set, the parameters can be adjusted to optimize the performance, in case the performance is already good. In case the performance is not satisfying, the process has to be started over at an earlier stage, depending on the actual performance. Reliant on the application case, a rule of thumb is that 70% of the data set is used as a training data set, 20% as an evaluation data set (in order to adjust the parameters – e.g., bias) and the final 10% as a test data set, however, in practice often a 70% (training data) and 30% (test data) split is utilized.

There are several established supervised ML algorithms available. Each of these algorithms has specific advantages and limitations concerning the application in manufacturing (see Table 4).

Table 4: Comparing learning algorithms (Kotsiantis, 2007)

	Desicion trees	Neural Networks	Naïve bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)	*** (not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***
Explanatory remarks: **** stars represent the best performance & * star represents the worst performance						

A major challenge is to select a suitable algorithm for the requirements of the research problem. As described before, as a first step, a general applicability of a ML algorithm with the requirements may be derived from more general comparisons (e.g., presented by Kotsiantis (2007)). This may be conducted to rule out unsuitable ML algorithms. However, due to the individual nature of most research problems and the specific characteristics of ML algorithms as well as their adapted 'siblings', it is not advisable to base the decision for a ML algorithm solely on such a theoretical and general selection. In order to identify a suitable ML algorithm for the problem at hand, the next step involves a careful analysis of previous applications of ML algorithms on research problems with similar requirements. The research problems do not have to be located within the same domain. A major issue in this selection is the matching of the identified requirements, which in this case include the ability to handle multi-variate, high dimensional data sets and the ability to continuously adapt to changing environments (updating the learning set).

The selected ML algorithms to be applied to the identified research problem within this dissertation are SVM. The detailed argumentation and the identified comparable structured problems with matching requirements are presented in the following subsection in greater detail.

5.2 Selection of suitable machine learning technique

In this section, the ML technique SVM is presented as the algorithm of choice to apply on the identified research problem. Details concerning the choice and suitability of SVM are illustrated in the later section 5.2.2 (also see Wuest et al., 2013b). Burbidge, Trotter, Buxton & Holden (2001) found SVM to be a "robust and highly accurate intelligent classification technique well suited for structure–

activity relationship analysis". SVM can be understood as a practical methodology of the theoretical framework of Statistical Learning Theory (STL) (Cherkassky & Ma, 2009). SVM have a proven track record for successfully dealing with non-linear problems (Li, Liang & Xu, 2009). SVM can be combined with different kernels and thus adapt to different circumstances/requirements (e.g., Neural Networks; Gaussian) (Keerthi & Lin, 2003).

First, SVM, as a supervised ML algorithm is described in greater detail, presenting the main principles, technical background and its main application fields. Following, the reasoning for the choice of SVM with regard to the identified requirements of the research problem is shown. As previously stated, in this section, existing applications of SVM on similar problems (with regard to the requirements) are referred to within the argumentation. After introducing SVM and the rationale behind its choice, the application and evaluation approach of the technique is described in the following section 5.3.

5.2.1 Support vector machines (SVM)

SVMs were introduced by Cortes & Vapnik (1995) as a new ML technique for two-group classification problems. The idea behind it is that input vectors are non-linearly mapped to a very high dimensional feature space (Cortes & Vapnik, 1995). Lately, SVM as a relatively new supervised ML algorithm (Kotsiantis, 2007) received increasing attention within the research community due to their ability to balance structural complexity and empirical risk (Khemchandani & Chandra, 2009). The theoretical background of SVM is presented in the next subsection, followed by an introduction of different application fields of this algorithm.

5.2.1.1 Theoretical background[19]

SVM as a classification technique has its roots in STL (Khemchandani & Chandra, 2009; Salahshoor, Kordestani & Khoshro, 2010), has shown promising empirical results in a number of practical manufacturing applications (Chinnam, 2002; Widodo & Yang, 2007) and works very well with high-dimensional data (Sun, Rahman, Wong & Hong, 2004; Ben-hur & Weston, 2010; Wu, 2010; Salahshoor et al., 2010; Azadeh, Saberi, Kazem, Ebrahimipour, Nourmohammadzadeh & Saberi, 2013). Another aspect of this approach is that it represents the decision boundary using a subset of the training examples, known as the support vectors.

SVM are linear two-class classifiers (Ben-hur & Weston, 2010). The basic idea behind SVM, is the concept of a maximal margin hyperplane. A linear SVM can be trained explicitly to look for this type of hyperplane in linearly separable data.

[19] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest, Irgens & Thoben, 2012b)

However, the method can also be extended to non-linearly separable data using kernels. Through application of kernels, the former linear classifier can be extended to serve as a non-linear classifier. This will be explained later on.

A linear SVM is based on a decision boundary, called hyperplane that divides a set of data points into two classes. These two classes are described as either positive (+1) or negative (-1) examples. Figure 51 illustrates a decision boundary ($\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$) between two linear separate sets of positive ($\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b > 0$) and negative ($\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b < 0$) class (Ben-hur & Weston, 2010).

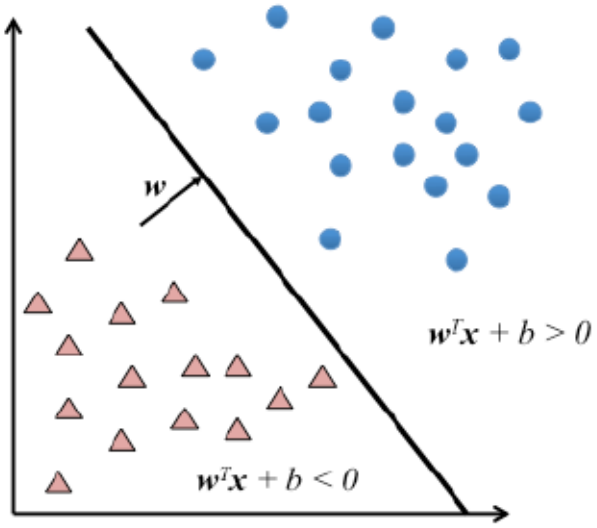

Figure 51: Linear classifier with decision boundary $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$ (Ben-hur & Weston, 2010)

The following technical definition of linear SVM follows Ben-hur & Weston (2010) if not indicated otherwise. In this example, $\boldsymbol{x}$ is understood as a vector with components x_i. The term $\boldsymbol{x}_i$ symbolizes the i^{th} vector in a data set, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, with y_i describing the label $\boldsymbol{x}_i$ is associated with. The scalar product, which is required in order to define a linear classifier is defined as $\boldsymbol{w}^T\boldsymbol{x} = \sum_i w_i\, x_i$. Based on this foundation, a linear classifier is built around the discriminant function:

$$f(x) = \boldsymbol{w}^T\boldsymbol{x} + b \qquad (1)$$

In this function, vector $\boldsymbol{w}$ is defined as the weight vector whereas b is known as the bias. By considering $b = 0$, the set of points x such that $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = 0$ represent all points perpendicular to $\boldsymbol{w}$ which go through the origin. Hence it forms a line (two dimensions), a plane (three dimensions), and more generally, a hyperplane (n dimensions). The vector $\boldsymbol{w}$ in this aspect represents a decision hyperplane normal vector and is commonly named weight vector in SVM literature (Hamel, 2009; Manning et al., 2009). Thus both bold notations indicate vectors whereas T stands for transpose and the bias b translates the hyperplane away from the origin, when $\neq 0$. Consequently, the hyperplane can be described as:

$$\{x : f(x) = \boldsymbol{w}^T \boldsymbol{x} + b\} \quad (2)$$

Dividing the space in two, this hyperplane allows to classify points by the sign of the discriminant function *f*(x) as positive and negative (Ben-hur & Weston, 2010).

Figure 52 shows a data set containing examples that belong to two different classes, represented as squares and circles. The data set is also linearly separable; i.e., there is a hyperplane such that all the squares reside on one side of the hyperplane and all the circles reside on the other side. A linear decision boundary between regions (see Figure 51) defines a classifier as linear. Inevitably such clear-cut results are not always available in real applications and suitable compromise solutions are used in order to allow a certain amount of misclassification.

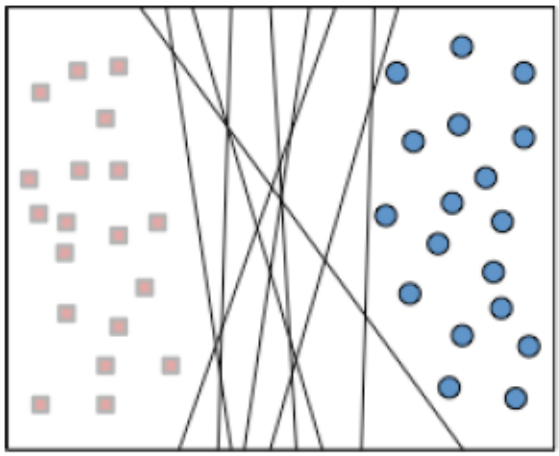

Figure 52: Possible decision boundaries for a linear separable data set (based on Hamel, 2009)

There are infinitely many hyperplanes possible. Although their training errors may be zero, there is no guarantee that the hyperplanes will perform equally well on previously unseen examples. The classifier must choose one of these hyperplanes to represent its decision boundary, based on how well they are expected to perform on test examples.

SVM are utilizing the concept of a maximal margin separation (see Figure 53) (Lessmann, Sung & Johnson, 2009). The algorithm tries to maximize the distance between the decision surface and data points separating the two classes (circles and squares) (Cristianini & Shawe-Taylor, 2000; Manning et al., 2009). The algorithm tries to place the decision surface maximally far away from the data points (Manning et al., 2009). In Figure 53, the optimal decision surface B is assumed to be the chosen decision surface with the largest margin. The supporting hyperplanes $b_{(-1)}$ and $b_{(1)}$ show the distance to the closest data points (support vectors).

The maximization of the margin reduces the upper bound of the (expected) generalization error, i.e., error of future data (Vapnik, 1995; Kotsiantis, 2007; Lessmann et al., 2009). The decision function is defined by a sub-set of the data (training da-

ta) defining the position of the hyperplane (Manning et al., 2009). "These points are referred to as the support vectors (in a vector space, a point can be thought of as a vector between the origin and that point)" (Manning et al., 2009). As SVM is a supervised ML algorithm, this training data is generally selected by experts.

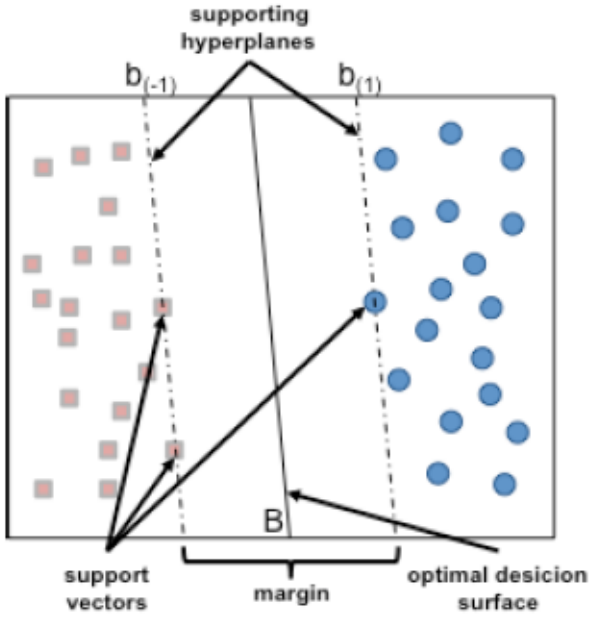

Figure 53: Margin for decision boundary (based on Hamel, 2009)

The problem with SVM classifiers is to establish relevant training/learning data for the case at hand so that the computation of the support vectors can be completed. Given that most process control and analysis task display a degree of dynamism, the use of SVM is not immediately clear for such applications as it is likely that the generation of support vectors will have to be done frequently and without knowledge of detailed performance results of the process.

The geometric representation of the maximum margin is briefly described, mainly based on Ben-hur & Weston (2010). For this example, it is assumed that the data is separable. Therefore, for a given hyperplane, x_+ and x_- are defined as the nearest data points to the hyperplane of the two classes (positive and negative as mentioned earlier). Then the length of the weight vector $\boldsymbol{w}$ is denoted by its norm $\|w\|$ and given by $\sqrt{w^T w}$. The unit vector $\widehat{w}$ in the direction of $\boldsymbol{w}$ can be obtained by $w/\|w\|$ having $\|\widehat{w}\| = 1$. Based on these preconditions, the margin of hyperplane $\boldsymbol{f}$ within the data set $\boldsymbol{D}$ may be seen as:

$$m_D(f) = \frac{1}{2}\widehat{w}^T(x_+ - x_-) \tag{3}$$

Within this formula $\widehat{w}$ is representing a unit vector in the direction of w and the distance of x_+ and x_- are assumed to be equal. Therefore the following equations may be set up:

$$f(x_+) = w^T x_+ + b = 1 \tag{4}$$

$$f(x_-) = w^T x_- + b = -1 \tag{5}$$

Adding the previous equations (4, 5) in the decision function (3), and divide it by $\|w\|$, a function as follows can be obtained:

$$m_D(f) = \frac{1}{2}\widehat{w}^T(x_+ - x_-) = \frac{1}{\|w\|} \tag{6}$$

As mentioned before, the data set is assumed to be linearly separable so a hard margin SVM can be applied. Later in this section, this is modified in order to handle non-separable data. A main functionality of SVM is the maximum margin classifier which is represented by a discriminant function maximizing the geometric margin $1/\|w\|$. As maximizing $1/\|w\|$ is the equivalent to minimizing $\|w\|^2$, the constrained optimization problem can be formulated as follows:

$$\underset{w,b}{minimize} \qquad \frac{1}{2}\|w\|^2 \tag{7}$$

$$subject\ to: \quad y_i(w^T x_i + b) \geq 1 \quad i = 1, \ldots, n$$

The resulting equation is based on the linear separability of the data set. However, in practice, data sets are not always linear separable. Additionally, when it happens to be linear separable, the achievable maximum margin is greater if the classifier allows misclassification of a certain number of data points, which is called soft margin SVM. In order to integrate a certain allowed classification error, the inequality constrain in (7) is replaced by

$$subject\ to: \quad y_i(w^T x_i + b) \geq 1 - \xi_i \quad i = 1, \ldots, n \tag{8}$$

This addition allows certain points (called slack variables: $\xi_i \geq 0$) to be either within the margin (called margin error: $0 \leq \xi_i \leq 1$) or misclassified ($\xi_i > 1$). Hence, a data point is misclassified when the value of the slack variable is exceeding 1. This allows for calculating the total number of misclassified data points ($\sum_i \xi_i$). Including this in equations (7,8) allows for representing a cost element (also known as penalizing element) in the maximum margin optimization function:

$$\underset{w,b}{minimize} \qquad \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{9}$$

$$subject\ to: \quad y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad i = 1, \ldots, n$$

In this function (9), the relative importance of maximizing the margin and minimizing the error is represented by the constant C > 0. By applying the Lagrange multiplier method, a dual formulation can be obtained, expressed by variables $\propto_i$ (Ben-hur & Weston, 2010):

$$\underset{\propto}{maximize} \quad \sum_{i=1}^{n} \propto_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i\, y_j\, \propto_i \propto_j x_i{}^T x_j \tag{10}$$

$$subject\ to: \quad \sum_{i=1}^{n} y_i \propto_i = 0, \quad 0 \leq \propto_i \leq C$$

Dual formulation, or duality, aims to convert "a linear model in the original (possibly infinite dimensional) 'feature' space into a dual learning model in the corresponding (finite dimensional) dual 'sample' space" (Zhang, 2002). The dual formulation allows to expand the weight vector $\boldsymbol{w}$ in terms of input data:

$$w = \sum_{i=1}^{n} y_i \propto_i x_i \tag{11}$$

Given $\propto_i > 0$, the points x_i are located on or within the margin in case a soft margin SVM is applied and are called support vectors. The number of data serving as support vectors with regard to the total number of data points is used as an upper bound of the error rate of the classifier (Ben-hur & Weston, 2010).

SVM so far have been presented as a classifier for linearly separable data with the addition of slack variables. As previously mentioned, SVM can be also used on non-linear data sets. Actually, SVM have a proven track record for successfully dealing with non-linear problems (Li, Liang & Xu, 2009). It has been found that non-linear classifiers provide better accuracy in many applications. However, they lack the advantage of linear classifiers, e.g., utilizing (relatively) simple training algorithms and scaling with regard to the number of examples. It has been shown, that through dual formulation, the SVM optimization problem is depending on the data through dot products. This allows to replace the dot product through kernel function which is non-linear and thus performing large margin separation in the feature-space of the kernel (Ben-hur & Weston, 2010). The dot product, also known as inner or scalar product, describes the generation of a single number out of two (equally long) sequences of numbers, e.g., a vector, through an algebraic operation (Manning et al., 2009).

As can be seen in Figure 54 a) & b), a SVM with a polynomial kernel allows to separate two classes more accurately than a soft margin linear SVM in this example. Through applying kernels, SVMs are able to "classify points by assigning

them to one of two disjoint half spaces, either in the pattern space or in a higher-dimensional feature space" (Khemchandani & Chandra, 2009).

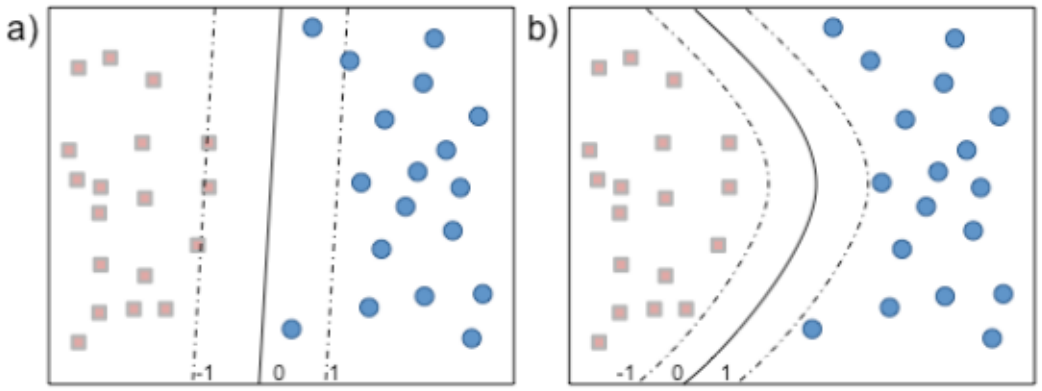

Figure 54: a) soft margin SVM with a linear kernel b) SVM with a polynomial kernel (based on Ben-hur & Weston, 2010)

Through applying a more complex kernel, the potential to achieve better results may be increased (Chinnam, 2002). On the other side, choosing the wrong kernel for the problem may reduce the performance of the classifier. As there are many different kernels for SVM available it is important to choose a suitable one, in accordance with the requirements of the data, in order to achieve the best classification results. A more complex kernel adds to the computation cost of the algorithm compared to a linear SVM (Ulaş, Yıldız & Alpaydın, 2012). However, given the recent development in ICT, most optimization problems of high-dimensionality, even with complex kernels, are mostly computable in a reasonable timeframe. Today, available software tools like 'RapidMiner' or 'WEKA' allow the user to apply different kernels within their solution without the need to do any programming themselves. Thus providing the opportunity to adjust the algorithm, including parameters and kernels, more easily based on the performance and requirements of a certain problem and the available data.

Overall, Ben-hur & Weston (2010) summarize the specific challenges for applying SVM in form of a list of decisions that have to be made prior to the application: "how to preprocess the data, what kernel to use, and finally, setting the parameters of the SVM and the kernel" (Ben-hur & Weston, 2010). For further in-depth reading on the technical background and a more detailed insight of the mathematical models behind the algorithm and kernels, the following publications are suggested (Burges, 1998; Sánchez, 2003; Larose, 2005; Bishop, 2006; Smola & Vishwanathan, 2008; Hamel, 2009; Manning et al., 2009; Alpaydin, 2010).

5.2.1.2 Application fields

SVM as a classifier technique has a very broad application field. Basically SVM can be applied wherever classification is needed. With regard to the research prob-

lem of this dissertation, certain domains where SVM was successfully applied are presented. While the focus is on manufacturing application, other domains with problems of similar nature are also included.

A major application area of SVM in manufacturing is monitoring (Chinnam, 2002). In particular within that domain, ***tool/machine condition monitoring***, fault diagnosis and tool wear are domains where SVM is continuously and successfully applied (Azadeh et al., 2013; Sun et al., 2004; Widodo & Yang, 2007; Salahshoor et al., 2010). Quality monitoring in manufacturing is another field where SVMs were successfully applied (Ribeiro, 2005).

An application area of SVM with an overlap to manufacturing application is ***picture/image recognition*** (e.g., character and face recognition) (Salahshoor et al., 2010; Widodo & Yang, 2007; Wu, 2010). In manufacturing, this applicationcan be utilized to identify (classify) damaged products (e.g., surface roughness) (Çaydaş & Ekici, 2010). Other application areas include handwriting classification (Scheidat, Leich, Alexander & Vielhauer, 2009). ***Time series forecasting*** is also a domain where SVM optimization is often applied (Tay & Cao, 2002; Guo, Sun, Li & Wang, 2008; Salahshoor et al., 2010).

Besides manufacturing and image recognition, SVMs are often used within the medicine domain. Among the many areas of application within this domain, the use of SVM in ***cancer research*** stands out (Furey, Cristianini, Duffy, Bednarski, Schummer & Haussler, 2000; Guyon, Weston, Barnhill & Vapnik, 2002; Rejani & Selvi, 2009). Other medical application areas are e.g., drug design (Burbidge et al., 2001) and detection of microcalcifications (El-naqa, Yang, Wernick, Galatsanos & Nishikawa, 2002).

Further application areas include but are not limited to credit rating (Huang, Chen, Hsu, Chen & Wu, 2004), food quality control (Borin, Ferrão, Mello, Maretto & Poppi, 2006), classification of polymers (Li et al., 2009) and rule extraction (Martens, Baesens, Van Gestel & Vanthienen, 2007). These examples from various industries and optimization problems highlight the wide applicability and adaptability of the SVM algorithm.

5.2.2 Rationale of SVM application for identification of state drivers in manufacturing systems[20]

The previous section presented the SVM algorithm's technical background and its different application fields. This section will focus on the suitability of SVM as a supervised ML algorithm for the previously stated research problem. First, the

[20] The content of this section has been partly published in accordance with (Universität Bremen, 2007) in (Wuest et al., 2013b)

main advantages and challenges of SVM regarding the problem requirements are presented in more detail. Following, two research problems of similar nature to the one at hand and how SVM was successfully applied to solve their optimization problem are highlighted. This way the choice as well as the suitability and applicability of an SVM algorithm on the identified problem of identification of state drivers in manufacturing are made comprehensible.

Before coming to the distinct advantages of SVM for the research problem, the main challenges one has to face when applying SVM are introduced. It has been found, that in order to achieve the high accuracy, a ***large sample size*** is required by SVM (Kotsiantis, 2007). SVM are also known for obvious over prediction when the available number of data examples is too small (Sun et al., 2004). However, given the derived problem of identifying state drivers in manufacturing, the sample size may be considered to be large enough to not cause any problems in most cases. As has been stated in previous sections, for the *product state concept*, continuous manufacturing with a large output is assumed. Therefore, this challenge may be considered as not relevant in this application scenario.

Over-fitting is commonly accepted as a draw back of SVMs under certain circumstances (Kotsiantis, 2007). However, other researchers found no indication for over-fitting problems for SVMs (with simple dot product kernels) (Chinnam, 2002). Thinking about over-fitting problems within this approach, it has to be considered that SVM is basically very resistant against over-fitting given that the training data has no massive class imbalance (Scheidat et al., 2009) and a specific hyperplane is chosen among the many separating the data (Vapnik, 1998). The chosen kernel and the nature of the data influence the risk of over-fitting when applying SVM (Cawley & Talbot, 2010). In this case, the training data may not be assumed to having a massive class imbalance, thus the over-fitting risk is assumed not to be problematic. However, once the individual application and its parameters are fixed, it has to be analyzed concerning the tendency to and risk of over-fitting.

As previously mentioned, another main problem of applying SVM algorithms is the large influence of ***choosing a suitable kernel and/or setting the right parameters***. In both cases a non-suitable choice has a significant impact on the SVMs optimization performance (Azadeh et al., 2013). This is a very common challenge similar to most supervised ML algorithms. As the software tools become more user friendly and the computational efficiency increases, today it is possible to compare test runs with different kernels and parameter settings in order to select a suitable alternative which allows to achieve a high classification performance. The selection of and decision for a suitable configuration can be done by utilizing k-fold cross validation. Typically n=10 provides good results with a reasonable effort. In order to do so, various configurations (e.g., different kernels, cost elements variations) are applied to the learning set and run through n-fold cross validation (n=10) until a good solution is determined.

After presenting the major challenges of SVM application in the previous paragraphs, the following paragraphs introduce major advantages. Overall SVM are found to "find an optimal tradeoff between structural complexity and empirical risk" (Khemchandani & Chandra, 2007). One major advantage of SVM over other supervised ML algorithms is that the solution of the classification problem is relatively straightforward and even though it may involve non-linear training, the output as an objective function is convex. In general the number of training points, still being relatively large and depending on the size of the training set, is larger than the number of basis functions (Bishop, 2006). This highlights the ***high interpretability and comprehensibility of the results*** for the practitioner compared to other algorithms like NN (Pham & Afify, 2005; Kotsiantis, 2007). This factor presents an important argument when thinking about applying the algorithm in a manufacturing environment.

Besides the relatively easy to interpret results, SVMs are capable of ***handling high-dimensional and multi-variate data*** (Sun et al., 2004; Kotsantis, 2007; Ben-hur & Weston, 2010; Wu, 2010; Salahshoor et al., 2010; Azadeh et al., 2013). Given that this is one of the major requirements of the identified research problem, this advantage is a strong argument for the choice of SVM as a classifier.

One, if not the most important, advantage is the proven ***high performance in practical applications*** of SVM algorithms. It has been found that "SVM generalization performance either matches, or is significantly better than, that of competing statistical and machine learning methods" (Chinnam, 2002). This is always an advantage for the application of SVM on classification problems.

Besides achieving high performance, the ***wide applicability*** of SVM algorithms is another advantage. SVM can be combined with different kernels and thus adapt to different circumstances/requirements (e.g., NN) (Chinnam, 2002; Keerthi & Lin, 2003). Furthermore, the wide applicability is supported by the factor that SVM inherit a high flexibility in modeling diverse sources of data (Ben-hur & Weston, 2010). As noted in the manufacturing domain, more specific advantages of SVM are described in recent literature. So found Chinnam (2002) that "SVMs are extremely good at recognizing shifts in correlated and non-correlated manufacturing processes". Burbidge et al. (2001) found SVM to be a "robust and highly accurate intelligent classification technique well suited for structure–activity relationship analysis". The advantage of SVM of allowing to "take advantage of prior knowledge of tool wear and construct a hyperplane as the decision surface so that the margin of the separation between different tool state examples is maximized" (Sun et al., 2004), underlines the suitability of SVM. Adding the fact that the "classification performance for every tool state can be adjusted" (Sun et al., 2004), presents a strong argument for SVM application.

After the main challenges and advantages have been discussed with regard to the identified research problem, existing problems of similar nature which have been approached using SVM are presented next. The goal is to highlight the suitability of SVM algorithms for problems of such kind and identify lessons learned on the application in order to incorporate those in the application and evaluation approach described in the next subsection.

The selected publications are all looking into similar problems: selecting examples of importance, calling it feature selection or gene selection method. The general idea is to identify 'relevant' factors which are either able to represent a system through generalization (feature selection) or are important to monitor as they may allow to predict a certain (future) outcome/behavior (gene selection). This corresponds highly with the set goal of this dissertation to identify state drivers, which are in return relevant state characteristics within the *product state concept*.

The first publication identified with a similar research problem is ***"A gene selection method for cancer classification using Support Vector Machines"*** (Guyon et al., 2002). The research background of this publication is that DNA micro-arrays allow the screening of large amounts of genes simultaneously in order to determine genes, which are either active, hyperactive or silent in normal or cancerous tissue. This corresponds highly with the previously introduced concept of 'good' and 'bad' states (see section 4.6) which applies to the 'normal' and 'cancer' tissue (with cancer tissue having different possible specifications). In this study, the research problem of selecting a small subset of genes from the large amount of available data using training examples from cancer and normal patients is addressed by applying SVM with Recursive Feature Elimination (RFE). The goal is to identify a set of genes, biologically relevant to cancer (Guyon et al., 2002). In this study, linear SVM are applied as they correspond with the nature of the investigated data set, the DNA micro-arrays.

The result of this research is that by applying SVM, it is possible to extract a small subset of relevant, highly discriminant genes as a basis for building a very reliable cancer classifier. It has been proven that SVM perform very effective for discovering informative features or attributes. Compared to other available methods for gene selection, the approach presented in the paper shows qualitative and quantitative advantages and outperforms "other methods in classification performance for small gene subsets while selecting genes that have plausible relevance to cancer diagnosis" (Guyon et al., 2002). During the study it has been found, that the experienced performance improvement of the SVM application are rooted in the SVM feature selection which provides the basis for the decision function whereas the way the decision function itself is trained, was found to be less important. Another finding is that SVM achieves better performance than other methods given a smaller selection of examples (genes) and is able to deal with high-dimensionality (number of features) and small number of training patterns (number of patients in

this case). An important finding is the need for preprocessing of the data as it has a strong impact on SVM. In this case, the scales have been made comparable by subtracting the mean and dividing the result by the standard deviation for every individual feature. The authors of the study cross checked the top ranked features selected by the SVM classifier and found that they are all known for their plausible relation to cancer (in contrast to other methods) in previous medical/biological research (Guyon et al., 2002).

Besides the advantages shown in the previous paragraph, two important findings that distinguish the application of SVM from other methods were found.

- SVM as multivariate classifiers *make use of information between features.* This is very important in the case of applying a similar approach on product state data as it has been shown in section 4.4 that process intra- and inter-relations exist and have to be taken into consideration.
- With the applied method, the *decision function is only based on support vectors that are "borderline" cases* (instead of all examples trying to map a typical case) (Guyon et al., 2002). In other words, the dominant parameters (drivers) that were found to have a significant influence on the classification (cancer/no cancer) are emphasized. This may also be a factor to consider in the following application within the *product state concept*.

Supporting the rationale behind this approach of applying SVM is that similar research also in the field of cancer research was undertaken by (Fung & Mangasarian, 2006; Huang, Zhang, Zeng & Bushel, 2013).

The second publication ***"Feature Ranking Using Linear SVM"*** (Chang & Lin, 2008) is actually based on the first one. In this study, again a linear SVM is combined the (SVM specific) feature ranking method introduced in (Guyon et al., 2002) and compared to a number of different feature ranking methods. The main advantage of feature ranking being that it supports the gain of knowledge about a data set and allows to identify relevant features (Chang & Lin, 2008). The findings indicate that the performance of this (relatively) simple method is very high and even outperforms several more complicated casual discovery methods. However, the method used ranks features based on their relevance and does not directly increase the knowledge on underlying causal relationships. Therefore, in this case the performance on non-manipulated data sets is found to be much better than on manipulated ones. Another important factor is that the study was undertaken within the so called 'causality challenge' which provided the data set and goal of making predictions on manipulated testing sets. This again corresponds with the requirements of the research problem of this dissertation, the identification of state drivers within the *product state concept*. However, the currently missing causality representation of this method may need to be addressed later.

Overall, it has to be stated that there are several other publication available besides the two previously presented ones looking into feature selection applying SVM (e.g., Bradley & Mangasarian, 1998; Bi, Bennett, Embrechts, Breneman & Song, 2003; Fung & Mangasarian, 2004; Mangasarian & Wild, 2007; Abe, 2010). These additional publications may also serve as argumentation for the selection of SVM for the given research problem, however, the author decided that the added benefits are marginal as the presented three can be considered sufficient in supporting the argument for the choice of SVM for the given research problem.

The rationale behind selecting SVM as a suitable approach for the given research problem of this dissertation was discussed in this section. It can be concluded, that SVM, as a classification method based on maximizing the margin between two groups of data points is theoretically suitable for the task of identifying state drivers within a manufacturing programme. The maximum margin hyperplane (and its weight vector $\boldsymbol{w}$), as a population separator and state classifier defining whether or not the x_i is positively classified is the key advantage of the SVM algorithm.

In the product state application scenario, the challenge lies in transferring the (inherited) relationships of product and process state characteristics in the algorithm and to interpret the results accordingly. The hyperplane, being constructed in the multidimensional space is able to reflect these relationships, meaning the timeliness of operations/processes from early state to a final state. Thus, SVM utilizing the hyperplane allows for classification in multi-dimensional space and furthermore to derive the driving parameters (or features/attributes) which are responsible for a change in class (this directly related to hypothesis 1 & 2). When applied to a product state description of a manufacturing programme, these driving parameters may represent state drivers which are (partly) responsible (or have a strong impact) on a change in class, which in this case would translate to a change between desirable or undesirable state ('good'/'bad'). Following, the application of SVM to identify state drivers is presented in greater detail.

5.3 Application of SVM for identification of state drivers

This section is structured in two major parts, one presenting the conducted application and evaluation and the second will provide an outlook on the derived results and how they may be interpreted. First, the previously introduced hypothesis 1 is adapted and discussed in more detail leading to the formulation of two sub-hypothesis, hypothesis 2 is adapted and hypothesis 3 is introduced. Next, the application of SVM on a manufacturing programme within a *product state concept* perspective is illustrated. This is structured around the previously adapted hypotheses and presents the structure of the research conducted in the subsequent section.

The second part looks at the expected outcomes of the application and evaluation in order to create awareness from the beginning on what may be expected. This is

important on the one hand to manage expectations and on the other to provide guidance on how the results are to be interpreted.

A crucial part of the approach is to evaluate the application of SVM on product and process state data and the ability to identify (known and unknown) process intra- and inter-relations between states and state characteristics (*hypothesis 1*). In order to achieve this, the hypothesis has to be detailed further. Two distinct sub-hypotheses can be derived from ***hypothesis 1*** 'Capturing of process intra- and inter-relations by implication through ML'. The overall hypothesis is updated to ***'Capturing of process intra- and inter-relations by implication through application of SVM'***. Hereafter, these two sub-hypothesis are described in greater detail.

Hypothesis 1.1 'Application of SVM allows the identification of state drivers of individual processes'

In hypothesis 1.1 the focus is on individual processes or operations. Again, process is used comprehensively throughout this section to reduce complexity. The individual process will be monitored using product and process state data based on the output of that process. The final result of the overall manufacturing programme is furthermore based upon the final quality assessment of the finished product. The manufacturing programme is seen as an entity of the manufacturing processes.

Within this hypothesis, the increasing complexity introduced to the progressing product state from process to process is not reflected in the observation as it focuses on individual processes. Therefore, process intra-relations which may have a significant influence across process/operation borders may be overlooked. This may prohibit the ability to detect and identify state drivers and unacceptable process drifts during intermediate stages of the process.

By analyzing the manufacturing programme using SVM, state drivers of an individual process/operation can be identified. The created state vectors indicate their influence on the state change by crossing the hyperplane (change of prefix +/-).

This hypothesis' focus on individual processes as a complementary approach is considered valid, in combination with, the following hypothesis 1.2. Hypothesis 1.2 reflects the importance of cross-process (inter-)relations better, which is a fundamental pillar of the *product state concept*'s view on manufacturing systems.

Hypothesis 1.2 'Combining different processes allows the identification of relevant drivers at different phases of the manufacturing programme'

Hypothesis 1.2 concentrates on identifying the process intra-relations across process borders within the manufacturing programme. As established in section 4.4, the states and state characteristics and their process intra- and inter-relations have to be analysed from a systems perspective. The goal is to provide a realistic moni-

toring and control of the product's progress towards its final product state and acceptable quality. The plan to achieve this is to make use of intermediate quality observations collections which may be reflected in accumulating state vectors for different stages. For example, in a manufacturing programme with three processes, the accumulated vectors are {process 1}; {process 1; process 2}; {process 1; process 2; process 3}. It is assumed that operational quality influences are incorporated at each stage so that the increase in complexity is captured and can be analysed stage-by-stage. This utilizes the previously stated finding that dependencies never go against the process flow and interdependencies between state characteristics can only exist within a state (see section 4.4). Concluding, hypothesis 1.2 reflects the *product state concept*'s overall understanding of a manufacturing programme and how it influences the final product's quality.

By creating accumulated state vectors, combining individual processes along the manufacturing programme and applying SVM, relevant state drivers reflecting (ex- and implicit) process intra- and inter-relations (system view) can be identified.

As this hypothesis is assumed to incorporate intra-relations (cross-process) of states and state characteristics to a higher degree than hypothesis 1.1, hypothesis 1.2 is considered the main research focus of this dissertation. However, there might be more sophisticated approaches of accumulating the state vectors throughout the manufacturing programme. This will be looked into during this research.

Hypothesis 2 'Adaptability to changing conditions through ML' is looking into how the proposed method reflects the need of a manufacturing programme for adaptability is not further divided as it is already focused enough to be evaluated in the following section 6. By utilizing the SVM algorithm, the hyperplane is the learning mechanism which can be updated/re-computed with high frequency and low computational effort. Its major practical limitation is the need for updated learning data. So this hypothesis is updated to ***'Adaptability to changing conditions through application of SVM'***.

In addition to hypotheses 1 and 2, the following paragraph introduces hypothesis 3. Hypothesis 3 reflects the future potential of the findings and is split in two sub-hypotheses similar to hypothesis 1. ***Hypothesis 3*** states ***'Through application of the SVM approach, defect products can be identified'***. This is further specified in *hypothesis 3.1* stating *'the trained SVM system is able to detect faulty products in the manufacturing programme'*. Connecting this hypothesis to hypothesis 1.1, *hypothesis 3.2* states *'a connection to the identified state drivers can be established within the set of (within the manufacturing programme) identified defect products'*.

Organizing the product and process data according to the *product state concept*, in a cluster of subsequent states, is the basis for this research and thus the approach illustrated in the next section. As was described in previous sections, there is still a

knowledge gap when it comes to the existing process intra- and inter-relations within a multi-stage manufacturing programme.

To test the previously identified hypotheses, three scenarios are analyzed in the following section. Two are based on publically available data sets, resembling a chemical manufacturing process (Kuhn & Johnson, 2013) and a manufacturing programme for semiconductors (Mccann, Li, Maquire & Johnston, 2010). One scenario is based on a mechanical engineering manufacturing process from the aviation industry provided by Rolls-Royce. In this case, no further information concerning the parameters or products in focus can be provided due to anonymity requirements. Each result will be compared to the results of the other scenarios in order to verify the made assumptions concerning the wide applicability of the approach. In all scenarios, the approach will be tested within a 'real world' application to verify its applicability in practice.

It has been established that modern manufacturing programmes often display chaotic behavior (Monostori, 2002). One reason for this can be that they tend to have very high dimensionality, at times extremely high dimensionality, and consequently the cause-effect mechanisms are hidden and the important process driving parameters are thus unknown and may indeed change with time such that parameters P which are important at time t will have changed to parameters P' at time $t+r$. This may indeed happen where the complete manufacturing programme is constructed from a number of interdependent processes/operations. Consequentially different process analysis and control methods are needed from the established orthodox ones. This in turn contributes to the chaotic nature of the manufacturing programme in that the process' outputs (product state) seem to be varying inconsistently with expected values for the given inputs. It is highly likely that this is a perception by the observer and that the outputs in fact are driven by cause-effect mechanisms as yet undefined/un-discovered. The implication of this is that such processes will seem to enter and exit process states in a random fashion and even the actual process states may seem to be random and undefined, adding to the chaotic perception by the observer (Figure 55).

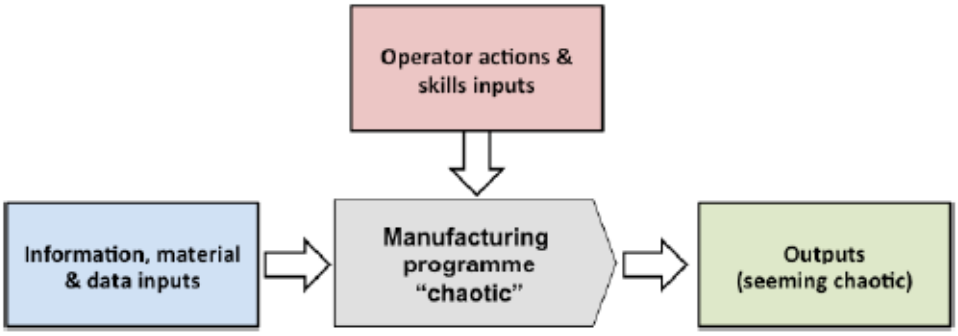

Figure 55: Chaotic nature of manufacturing programmes

If it is possible to bring some order into the seemingly chaotic output, then it may be possible to identify the set of product and process state. If product and process states can be identified, then it would be possible to identify the associated inputs and from this determine the actual state 'drivers' (driving variables) which are found to influence the process results (product states). Each process state is thus associated with input and output variables' values and can then be classified as good or bad. Given that the drivers for each good or bad product state is known, the seemingly chaotic manufacturing programme could be perceived as ordered and thus controllable (Figure 56).

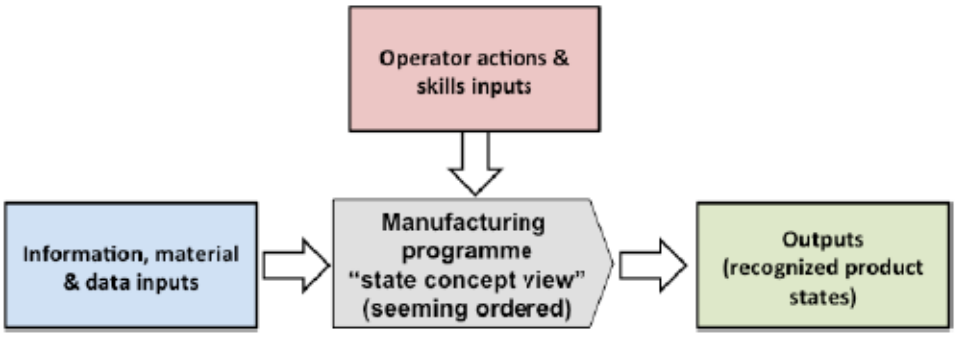

Figure 56: Order manufacturing programme according to the product state concept

The assumption is that processes are likely to operate across a relatively constant set of process states. This means that one may assume that the product will enter a specific state dependent upon the input to the process. For complex chain of manufacturing programmes such input should include the human participants responsible for the effectuation of the process. The process concept would thus become (see Figure 57):

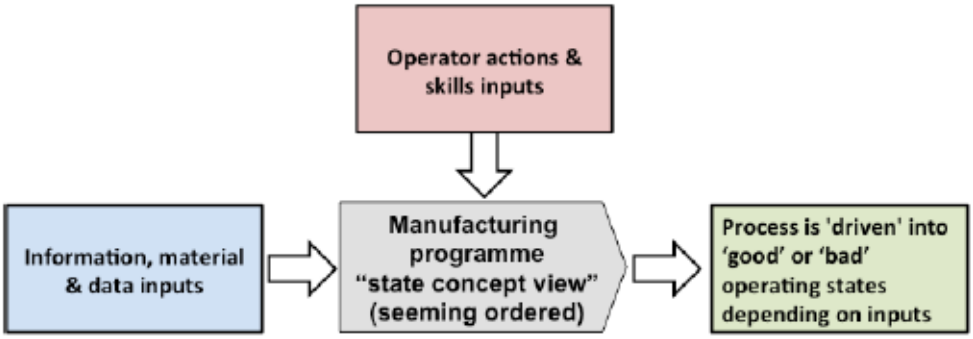

Figure 57: Final product state is driven by previous product and process states of the manufacturing programme (based on Wuest et al., 2013b)

The conceptual approach for the application of the approach on manufacturing data is presented in the following paragraphs. As each of the three evaluation scenarios is different and serves different purposes, each will vary to some extent from the following generic methodology. Also the different steps, though numbered in Figure 58, are not followed strictly as they may be run parallel or in reverse order in

the three scenarios. This is due to the nature of the data and the specific focus of the individual evaluation. For example, scenario I & II focus on hypothesis 1.2 and have multiple processes within the manufacturing programme.

Each of the three scenarios is introduced briefly before the available data sets are organized according to the *product state concept*. In this first step, the three scenarios differentiate thus far, that scenarios I & II focus on the processes and their interconnection within the manufacturing programme whereas the third scenario focuses on the individual and final product state.

The data sets are pre-processed so that the SVM algorithm may be applied. For more details refer to section 9.2 in the Annex. Again, the pre-processing is different for each of the three scenarios and reflects the nature of the data and the goal of the evaluation. Important steps of data pre-processing include the replacement/handling of missing values, creating synthetic process based on existing process parameters and standardization/normalization of the data set.

In a next step, first hyperplanes of the classifier may be computed. Of the available selection the most suitable parameter (incl. kernel) configuration for the available data set is to be selected using n-fold cross-validation. The typical parameter for the cross-validation is n=10. This provides a first impression of the classification power through a confusion matrix and individual weights of the $\boldsymbol{w}$ vectors. As soon as a suitable parameter configuration is chosen, the classification power of the algorithm may be tested by use of the test data set by the trained (using the learning set) SVM algorithm.

After the data is prepared and the classifier is set up and running, the testing of the research hypotheses is executed. There will be different evaluations for each scenario. For example, in scenario I & II different accumulated vectors will be derived and analyzed by the SVM algorithm to test hypothesis 1.2. To do so, SVM feature evaluation according to Guyon et al. (2002) is applied to derive the weights for the individual features and rank them accordingly. However, the specific approach is described in more detail within the evaluation scenario set up in section 6.

Another part of the evaluation is described by the next activity. Here the complete data set is split into a learning (70%) and a test (30%) set (Borovicka, Jirina, Kordik & Jirina, 2012). In this case there is no need for an evaluation set which is needed for some algorithms and in general accounts for around 10% of the whole data set. The learning and test set are decided based on the specific situation. Depending on the available knowledge, different methods may be applicable. In this case, the splitting is done randomly, however keeping the ratio of the original set concerning the two classes. The goal of this evaluation is to show the classification performance of the classifier on formerly unknown examples. The general methodology is summarized in Figure 58.

Next, the to be expected results will be briefly introduced in order to prepare the reader on what to expect. This will not entail any information on values or findings, but more the visual and information output of the applied approach. The structure is oriented on the previously presented hypothesis and the presented approach. The goal is to prepare the reader in what to expect from the following evaluation. However, it has to be understood that this section is not replacing the discussion of the results afterwards.

Overall, the reasoning behind the evaluation using three scenarios from different domains with data sets of different complexity is to show the general applicability of the developed approach. Whereas scenario I, the Rolls-Royce manufacturing programme resembles the targeted area of mechanical engineering, scenario II, the chemical manufacturing process gives a different perspective. Scenario III, the semiconductor manufacturing process, was chosen to show the challenges a real world data set may present regarding data pre-processing (e.g., missing values), classification and general structure of the data (hypothesis 2).

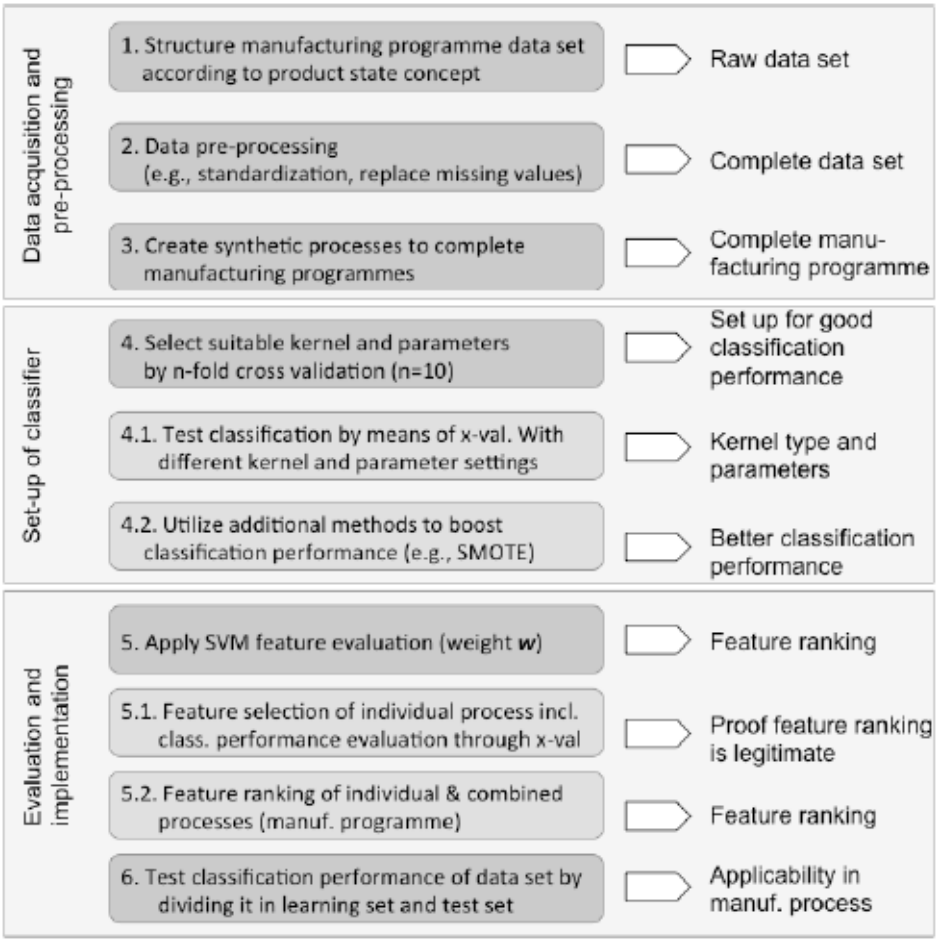

Figure 58: General application approach for evaluation

The results of the pre-processing may not provide additional arguments to answer the raised research question (hypotheses). Therefore they are not part of the main

body of the dissertation and presented in the Annex (see Annex section 9.2). However, they provide a necessary step of every ML based approach.

The second area of results provides the classification performance of the different data sets. This part is crucial as it provides evidence that the approach is applicable to real life application cases. Part of these results is also the application of a feature ranking/selection as described in Guyon et al. (2002). In this case, feature selection is applied, reducing the number of attributes (features) of the data sets. Then the classification performance of the reduced (variations) data sets is analyzed. In this regard, especially the comparison of classification performance by means of cross-validation 'pre-feature selection' and 'post-feature selection' provides evidence that relevant state drivers can be identified by the approach (hypothesis 1.1).

After the previously sketched results show that it is possible to identify relevant features (state drivers) for individual processes, the next challenge is to show that it is also possible to identify relevant state drivers cross-process which reflect the process intra- and inter-relations highlighted in the previous sections. Those are a key point of the *product state concept* and thus it is essential that the approach is able to include them. This is done by applying feature ranking to combinations of processes in addition to the individual processes as described before. By doing so, the applied SVM feature weights indicate the important features the same way as they do for individual processes. However, this way they incorporate the (implicit and explicit) cross-process intra-relations which have an influence of the product state. The results are then different rankings showing the relevant features for both individual and combined processes (manufacturing programme) and by analyzing and comparing them, especially shifts of importance along the program, indicates the inclusion of important process intra- and inter-relations (hypothesis 1.2).

Looking at the classification performance of the model, trained by the learning set and applied to the test set, shows the ability of the approach to create a model which may be implemented in a manufacturing programme to identify quality problems at an early stage (hypothesis 3).

6 Application of SVM to identify relevant state drivers

In this section, the previously derived hypotheses are evaluated by developing and analyzing three scenarios. The section is structured as follows: at first the scenarios are briefly introduced (for more detail refer to Annex section 9.2). The following two subsections focus on the application of the previously introduced research plan on the three scenarios. However, it has to be noted that the scenarios were not evaluated following the presented sequence during the analysis phase. The presented sequence (scenario I-III) does not resemble the timely sequence of evaluation of the different scenarios. Therefore, it is possible that the background of and justification for some of the methods, tools and applications are explained in later sections even so they are applied beforehand. In such cases, reference is given to the more detailed explanation in later sections. The next section 7 presents and discusses the evaluation results and illustrates the limitations of the approach.

6.1 Introducing scenarios I, II and III

In this section the three evaluation scenarios are introduced and the available data for each scenario is presented and analyzed. After the three processes and accompanying data sets are presented individually, necessary pre-processing steps were conducted. Since this is not part of the main application approach, this is expanded on in the Annex (see section 9.2). The pre-processing entails among other things, replacing missing values (scenario II & III) and the generation of additional data (scenario I & II). The result of the pre-processing are three data sets ready for the application of SVM algorithms in order to identify state drivers. The three data sets complement each other in terms of the evaluation focus areas and goals.

The *first scenario 'RR'* (details in Annex section 9.2.1) is based on a mechanical manufacturing process of a highly stressed product. The scenario is set in the aviation domain and is provided by Rolls-Royce. The 'real world' data set resembles a complex process in the manufacturing programme. It is supplemented by two additional synthetic processes named 'Dick' and 'Harry' which are generated based on the characteristics of the original data set.

The *second scenario 'CHEM'* (details in Annex section 9.2.2) is similarly designed and set in the chemical manufacturing domain. The original data set is complemented by two additional synthetic processes, based on the characteristics of the original process. Both scenarios aim to show how the structuring according to the *product state concept*, intelligent combination of processes and application of SVM, allows the identification of state drivers throughout the manufacturing programme (hypothesis 1.2). As can be seen in Figure 59 a), scenario I and II focus on different areas of the manufacturing programme for the evaluation of the hypotheses. The data sets resemble the manufacturing programme as well as the individual processes. Therefore, different analyses can be conducted, e.g., combining differ-

ent process vectors to accumulated process vectors (e.g., process 1 & 2 – highlighted in dashed red line). Having distinct 'real world' scenarios from different domains allows to additionaly evaluate the applicability of the approach in different environments of manufacturing. This is extended further by the following third scenario, set in the semiconductor manufacturing domain.

This *scenario III 'SECOM'* (details in Annex section 9.2.3) represents another 'real world' manufacturing programme, based on process data from the semiconductor manufacturing domain. The main purpose of this scenario is to apply the approach on a data set that presents a challenge for pre-processing and application due to its highly imbalanced, high dimensional nature, additionally containing a large amount of missing values. This is common in real world data sets and thus the developed approach needs to evaluate its applicability towards such data sets. However, this scenario shall also provide further evidence of the broad practical application potential of the develop approach given the different domain it is set in. Furthermore, scenario III will be used to support the evaluation of hypothesis 1.1. The evaluation focus of this scenario is on the complete manufacturing programme (highlighted in continuous orange line) and partly on an individual process (highlighted in dashed orange line) (see Figure 59 b). The term 'partly' describes the assumption that even though the complete data set describes a manufacturing programme, it can also be perceived as a process containing various operations. This may be of relevance for the later comparison of the results between this scenario and scenarios I & II.

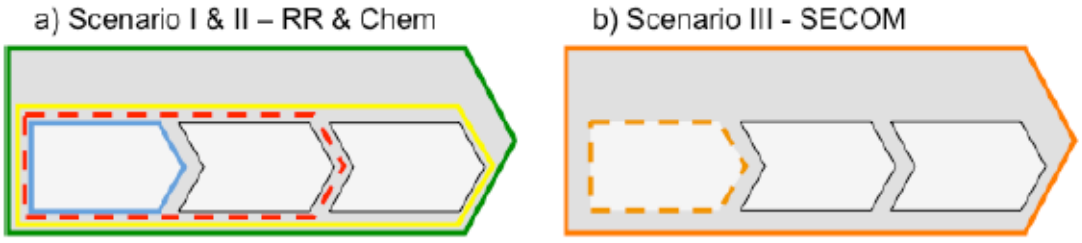

Figure 59: Summary of focus areas of evaluation scenario I, II & III

For details of the data pre-processing, it is advised to refer to the Annex. The application of the developed approach is evaluated in the following subsections.

6.2 Scenario I – Rolls-Royce

In this section, the described approach of applying supervised ML (SVM algorithm) to identify relevant information in form of state drivers (relevant features) within the Rolls-Royce data set as described within the approach is presented. At first the classification performance of the Rolls-Royce data set is evaluated using first a linear kernel and later an ANOVA kernel. This is done in detail for the TOM(RR) process as it represents real world data. For the synthetic and combined vectors, a basic evaluation is applied in order to evaluate the suitability of the fea-

ture ranking method. Following, a feature ranking based on SVM weight vectors $\boldsymbol{w}$ is performed and the classification performance of the data set with different reduced feature sets is compared to the original full features data set based on the TOM(RR) process. Then a feature ranking is applied for the other synthetic processes as well as the combined vectors. Finally, the classification performance of the model on previously unknown data is analyzed using a (random) split in a learning (70%) and test (30%) set. However, for this scenario, the results of the analysis have specific limitations which will be specified in the respective section.

6.2.1 SVM kernel & parameters for hyperplane by x-validation

At first the classification performance of the data set is tested through 10-fold cross-validation. For a more detailed description on how this is applied technically, refer to section 6.4.1. In a first step, the basic linear kernel is applied with original parameters as provided by RapidMiner (v5.3).

accuracy: 72.32% +/- 1.68% (mikro: 72.32%)

	true Defect140	true PASS	class precision
pred. Defect140	1774	837	67.94%
pred. PASS	324	1260	79.55%
class recall	84.56%	60.09%	

Figure 60: Confusion matrix showing the classification performance of x-val. for TOM(RR) with a linear kernel (orig. parameters)

The classification results of this test show acceptable results even so they are partly below a threshold of 80% (see Figure 60). The target threshold of 80% over all class prediction and class recall percentages is used for the evaluation within this dissertation. This is deemed to represent a good classification results for the application case within the three presented scenarios. It is however not possible to judge a classification performance over all domain, data sets, etc. as the threshold may vary significantly for what is deemed a good or a bad result. In general, this has to be determined based on the case at hand and the circumstances (Witten, Frank & Hall, 2011).

accuracy: 78.24% +/- 1.91% (mikro: 78.24%)

	true Defect140	true PASS	class precision
pred. Defect140	1840	655	73.75%
pred. PASS	258	1442	84.82%
class recall	87.70%	68.76%	

Figure 61: Classification results of RR data set (TOM(RR)) by x-val. after basic parameter optimization on linear kernel (C 1.5)

After some basic optimization of the linear kernel parameters (C 1.5) (see Figure 125 in Annex), the classification results improve considerably, nearing the target

threshold (see Figure 61). These acceptable classification results running with the basic linear kernel indicate that the later feature ranking based on SVM algorithm determined weight $\boldsymbol{w}$ is applicable for this data set's structure.

After this first run with the linear kernel, a parameter optimization is run to find the best fitting parameters for the data set and improve the classification performance significantly. The identified parameter set with the best classification performance is an ANOVA kernel with the following specs (different from orig.): kernel degree 3.0 and C 1.0. The results are significantly higher than the target threshold of 80% for class recall as well as class precision (see Figure 62).

accuracy: 98.45% +/- 0.47% (mikro: 98.45%)

	true Defect140	true PASS	class precision
pred. Defect140	2097	64	97.04%
pred. PASS	1	2033	99.95%
class recall	99.95%	96.95%	

Figure 62: Classification results of RR data set (TOM(RR)) by x-val. after basic parameter optimization on ANOVA kernel (kernel degree 3.0 & C 1.0)

After the original data set provided by Rolls-Royce is analyzed and shows good classification performance with the chosen SVM algorithm, the additional synthetic data sets of the individual processes DICK(RR) and HARRY(RR) as well as the combined vectors TD(RR) and TDH(RR) are analyzed according to their classification performance with a linear kernel. This is also done to show the general suitability of the feature ranking based on the weight vectors $\boldsymbol{w}$.

a) TOM(RR) (x-val; orig. para.; DOT kernel)

Accuracy:	72.40%	
1774	834	68.02%
324	1263	79.58%
84.56%	60.23%	

b) DICK(RR) (x-val; orig. para.; DOT kernel)

Accuracy:	99.90%	
2693	2	99.93%
2	1498	99.87%
99.93%	99.87%	

c) HARRY(RR) (x-val; orig. para.; DOT kernel)

Accuracy:	99.93%	
2327	2	99.91%
1	1865	99.95%
99.96%	99.89%	

d) TD(RR) (x-val; orig. para.; DOT kernel)

Accuracy:	99.33%	
1869	12	99.36%
16	2298	99.31%
99.15%	99.48%	

e) TDH(RR) (x-val; orig. para.; DOT kernel)

Accuracy:	99.09%	
2156	18	99.17%
20	2001	99.01%
99.08%	99.11%	

Figure 63: X-val classification performance of the processes and combined vectors

The results are summarized in the following Figure 63 and show overall a very good classification performance for all processes and combined vectors. This was expected due to the chosen process of synthesizing the processes and combined vectors and assigning labels. Therefore, the later applied feature ranking method is assumed suitable for the complete data set (original and synthetic) of the Rolls-Royce scenario.

The very good classification performance by cross-validation of all processes and combined vectors using a linear kernel shows that a feature ranking by the weight vector $\boldsymbol{w}$ using an SVM classifier is applicable. In the following sub-section, such a feature ranking method is applied for all processes and combined vectors.

6.2.2 Feature ranking using SVM classifier

In this section, the features of the different processes and combined vectors are ranked according to their weight vector $\boldsymbol{w}$. This method is based on Guyon et al. (2002) and described in greater detail in section 6.3.4. As the WEKA[21] feature ranking function does not provide an output of the actual weight values, in this scenario the feature ranking function of RapidMiner (v5.3) is additionally utilized as illustrated in Figure 64. The feature ranking derived by the WEKA function is detailed subsequently in Table 5. Following a short comparison of the results of the two approaches (WEKA and RapidMiner) is described.

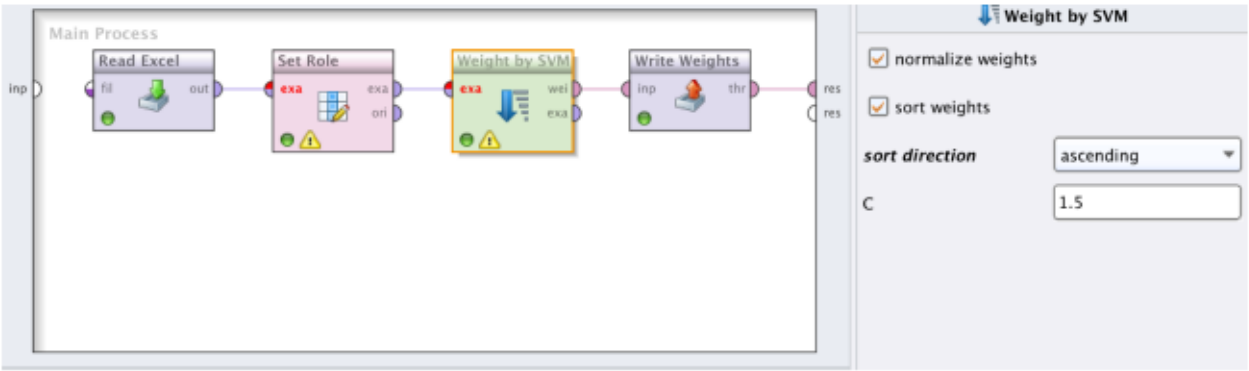

Figure 64: Feature ranking by SVM in RapidMiner (v5.3)

The RapidMiner (v5.3) function does not allow for the same customization as does the WEKA version. However, as has been previously established (see Figure 61), the C value was identified as the optimizing lever for linear kernels and this parameters can be adjusted in the function (see Figure 64).

The resulting feature ranking including the values of the weight vector $\boldsymbol{w}$ are depicted in Table 15 in the annex. The weight vector $\boldsymbol{w}$ values are normalized [1;0]. Based on this ranking, the feature selection is done. The chosen variants are FS10; FS15; FS20; FS30 and FS50. The variants are chosen not based on their weight value at this point, but for the comparability within and between scenarios. The later discussed variant with 57 features is based on the weight value, as the values of the features ranking no. 58 to no. 85 is under 0.1. For the respective data sets the classification performance is tested by 10-fold cross validation using the previous-

[21] WEKA 3: Data Mining Software in Java issued under the GNU General Public License (http://www.cs.waikato.ac.nz/~ml/weka/).

ly determined parameters (see Figure 62). The results are illustrated in the following Figure 65.

85			FS10			FS15		
Accuracy:	98.45%		Accuracy:	92.42%		Accuracy:	93.92%	
2097	64	97.04%	2040	260	88.70%	2058	215	90.54%
1	2033	99.95%	58	1837	96.94%	40	1882	97.92%
99.95%	96.95%		97.24%	87.60%		98.09%	89.75%	
FS20			**FS30**			**FS50**		
Accuracy:	95.07%		Accuracy:	96.59%		Accuracy:	97.66%	
2072	181	91.97%	2084	129	94.17%	2097	97	95.58%
26	1916	98.66%	14	1968	99.29%	1	2000	99.95%
98.76%	91.37%		99.33%	93.85%		99.95%	95.37%	

Figure 65: Comparison of classification performance by x-val for TOM(RR) with variations in no. of features (RapidMiner (v5.3))(ANOVA; kernel gamma 1; kernel degree 3; C 1)

The results show that even so the classification performance of the full features data set is the highest, even the data set with a significantly reduced feature set (10) reaches very good classification performance. The variations with 30 and 50 highest ranking features reach almost the performance of the full data set. This confirms that the feature selection is able to select the state drivers rather accurately. Looking at the weight values, one more evaluation is run with the 57 highest ranking features. Those features have a normalized weight value of 0.1 or above. Figure 66 shows that the results are nearing the results of the full feature set:

accuracy: 97.97% +/- 0.47% (mikro: 97.97%)

	true Defect140	true PASS	class precision
pred. Defect140	2097	84	96.15%
pred. PASS	1	2013	99.95%
class recall	99.95%	95.99%	

Figure 66: X-val classification performance on TOM(RR) with 57 highest ranking features

The results of the previously applied RapidMiner (v5.3) feature selection function are compared with the WEKA feature ranking function which is used in scenario II & III. The WEKA tool is designed based on Guyon et al. (2002) and thus directly applicable to the task at hand. Thereafter, the same evaluation is run with the feature ranking derived from the WEKA function for process TOM(RR) (see Table 5). A more detailed description of the WEKA feature ranking function and the fundamental method described by Guyon et al. (2002) is presented in section 6.3.4.

As can be directly observed, the feature ranking derived by the WEKA function (see Table 5) is rather different to the one derived by RapidMiner (v5.3) (see Table 15 annex). Next, the classification performance during the same test configurations as shown before are performed with the data set with reduced feature sets (based on the WEKA feature ranking). The results are presented in Figure 67.

Table 5: Feature ranking of TOM(RR) by WEKA

Rank	feature	Rank	feature	Rank	feature	Rank	feature
1	para.51	23	para.46	45	para.23	67	para.56
2	para.21	24	para.48	46	para.24	68	para.30
3	para.50	25	para.31	47	para.25	69	para.40
4	para.33	26	para.38	48	para.83	70	para.77
5	para.6	27	para.11	49	para.73	71	para.79
6	para.36	28	para.42	50	para.41	72	para.65
7	para.14	29	para.20	51	para.37	73	para.69
8	para.9	30	para.16	52	para.17	74	para.71
9	para.47	31	para.13	53	para.26	75	para.4
10	para.59	32	para.84	54	para.52	76	para.7
11	para.29	33	para.53	55	para.39	77	para.72
12	para.55	34	para.3	56	para.87	78	para.74
13	para.61	35	para.43	57	para.8	79	para.70
14	para.60	36	para.22	58	para.12	80	para.80
15	para.44	37	para.62	59	para.10	81	para.82
16	para.45	38	para.19	60	para.27	82	para.86
17	para.34	39	para.54	61	para.15	83	para.68
18	para.32	40	para.57	62	para.18	84	para.78
19	para.64	41	para.81	63	para.28	85	para.76
20	para.2	42	para.75	64	para.63		
21	para.5	43	para.49	65	para.67		
22	para.35	44	para.58	66	para.66		

When comparing the classification performance results of the different data sets with varying number of features, the tendency is similar for the feature ranking derived by RapidMiner (v5.3) (see Figure 65 and Figure 66) and WEKA (Figure 67).

85			FS10			FS15			FS20		
Accuracy:	98.45%		Accuracy:	91.99%		Accuracy:	94.06%		Accuracy:	96.04%	
2097	64	97.04%	2026	264	88.47%	2056	207	90.85%	2083	151	93.24%
1	2033	99.95%	72	1833	96.22%	42	1890	97.83%	15	1946	99.24%
99.95%	96.95%		96.57%	87.41%		98.00%	90.13%		99.29%	92.80%	
FS30			**FS50**			**FS57**					
Accuracy:	96.92%		Accuracy:	98.19%		Accuracy:	98.33%				
2090	121	94.53%	2096	74	96.59%	2097	69	96.81%			
8	1976	99.60%	2	2023	99.90%	1	2028	99.95%			
99.62%	94.23%		99.90%	96.47%		99.95%	96.71%				

Figure 67: Comparison of classification performance by x-val for TOM(RR) with variations in the number of features (WEKA) (ANOVA; kernel gamma 1; kernel degree 3; C 1)

In Figure 67 the classification performance of similar feature selection variations (10; 15; 20; 30; 50; 57 features) using 10-fold cross-validation with an SVM classifier (ANOVA kernel) is evaluated. In this case the feature selection is based on the feature ranking derived with the feature ranking function of WEKA based on Guyon et al. (2002). Overall the results show also very good classification performance results, similar to the previous evaluations of variants based on the RapidMiner (v5.3) feature selection function. It shows that the classification performance improves the more features are employed by the data set. However there is a slightly better performance noticeable for the WEKA ranking for data sets with

15 and more features compared to the RapidMiner (v5.3) ranking. Just for the data set with 10 features the RapidMiner (v5.3) version outperforms the WEKA version by a fraction. Overall, the results of both variations can be considered very good.

In the following paragraphs, one more comparison analysis for the two different feature ranking functions is conducted. For each version, a data set with the 20 highest ranking features and the 20 lowest ranking features is compared (see Figure 68). Interestingly for the RapidMiner (v5.3) version, it shows that the data set with the 20 lowest ranked features performs better than the version with the 20 highest ranked features. This is rather unexpected. The WEKA version on the other hand shows a significantly better performance for the 20 highest ranked feature data set with an accuracy of 96% over 74% for the 20 lowest ranked feature data set.

RapidMiner (v5.3)						WEKA					
highest FS20			lowest FS20			highest FS20			lowest FS20		
Accuracy:	95.07%		Accuracy:	95.23%		Accuracy:	96.04%		Accuracy:	75.14%	
2072	181	91.97%	2072	174	92.25%	2083	151	93.24%	1599	544	74.62%
26	1916	98.66%	26	1923	98.67%	15	1946	99.24%	499	1553	75.68%
98.76%	91.37%		98.76%	91.70%		99.29%	92.80%		76.22%	74.06%	

Figure 68: Comparison of class. Perf. by x-val for TOM(RR) for WEKA and RapidMiner (v5.3) version with 20 highest and lowest ranked features (ANOVA; ker. gamma 1; ker. degree 3; C 1)

Comparing the number of similar features contained in the different variations of the WEKA and RapidMiner (v5.3) ranking, the FS10 variant also stands out with a very low overlap percentage of 20% followed by the 'lowest FS20' variant with 30% whereas all other variants show overlaps of 40% and above (see Figure 69).

FS variant	**FS10**	**FS15**	**FS20**	**FS30**	**FS50**	**FS57**	**lowestFS20**
No. of identical features	2	6	8	16	34	44	6
Percentage of overlap	10.00%	40.00%	40.00%	53.33%	68%	77.19%	30.00%

Figure 69: No. of features contained in both rankings of WEKA and Rapidminer (v5.3)

As for the original Rolls-Royce data set TOM(RR) expert knowledge is available, the ranking of features by WEKA was approved by the experts. The WEKA ranking was found to share a higher compliance with the existing expert knowledge of the relevant process parameters than the Rapidminer (v5.3) ranking. This confirms the previous suspicion of the WEKA ranking method being superior to the Rapidminer (v5.3) one.

The WEKA analysis confirmed already known relevant process parameters for the TOM(RR) processes. More importantly the conducted analysis also identified a new and potentially most important influence by including the implicit process inter-relations. In this case the process intra-relations could not be confirmed in a similar fashion by the RR experts as the processes DICK(RR) and HARRY(RR) and the combined vectors TD(RR) and TDH(RR) are supplemented by synthetic

data. The experts however acknowledged the potential those analyses promise and are interested in exploring the applicability within their manufacturing system further. Based on this results, for this scenario, the WEKA version is used in the throughout the following evaluation scenarios for feature ranking purposes.

6.2.3 Classification on previously unknown data

As has been previously mentioned, the results of this sub-section, looking into the classification performance of previously unknown data for the RR data are not necessarily generable or comparable to those of scenario II & III. The reason for this is that SMOTE oversampling was applied prior to the provision of the data set by Rolls-Royce. As SMOTE does add additional examples to the minority class that are inspired by the existing population, the data of the test set cannot be considered unknown.

However, as is discussed in the following results section (see section 7.1.5), the classification performance on the test set may still indicate how the data set may behave when more examples of the minority class are available under the assumption that the future minority examples are not too diverse.

For the application of this method, the TOM(RR) data set is randomly split in a learning set (70%), used to train the model and a test set (30%) on which the model is applied subsequently. The feature selection in this section is based on the feature ranking done by the RapidMiner (v5.3) SVM weight function. For further details on the technical aspects of the process refer to section 6.3.2.

85			FS10			FS15		
Accuracy:	97.30%		Accuracy:	91.10%		Accuracy:	92.77%	
641	2	99.69%	608	35	94.56%	618	25	96.11%
32	583	94.80%	77	538	87.48%	66	549	89.27%
95.25%	99.66%		88.76%	93.89%		90.35%	95.64%	
FS20			**FS30**			**FS50**		
Accuracy:	94.28%		Accuracy:	96.03%		Accuracy:	96.66%	
619	24	96.27%	632	11	98.29%	638	5	99.22%
48	567	92.20%	39	576	93.66%	37	578	93.98%
92.80%	95.94%		94.19%	98.13%		94.52%	99.14%	

Figure 70: Comparison of classification performance of previously unknown data for TOM(RR) with variations in no. of features (ANOVA; kernel gamma 1; kernel degree 3; C 1)

The analysis of classification performance on previously unknown data, depicted in Figure 70 show results that are similar to the ones obtained by cross-validation (see Figure 65). For the full feature set, all four percentages are in the mid- to high-nineties and thus have to be considered very good. The variations with different numbers of highest ranking features show that the classification performance results are slightly lower than for the full feature set. Nevertheless, even the data set reduced to the 10 highest ranking features shows very good classification results

significantly higher than the target threshold of 80%. This indicates that even with reduced features the outcome can be predicted with a high accuracy. The limitations of the RR data set for the learning/test evaluation as stated before also apply for this test. In the evaluation scenario the analysis of a chemical manufacturing programme is presented.

6.3 Scenario II – Chemical Manufacturing Process

In this section, the previously introduced chemical manufacturing process is evaluated with regard of the proposed research hypothesis. At first, the classification performance of the data set is evaluated using by applying a cross-validation (10-fold). Following, a suitable SVM classification algorithm is identified (regarding kernel and parameter choice) and if needed, additional measures are taken (e.g., oversampling of minority class, etc.) in order to optimize the classification performance. Then, the identification of relevant attributes (features), the state drivers is applied by the feature selection method proposed by Guyon et al. (2002). In a first step a feature ranking is derived, sorting the features according to their weight vector $\boldsymbol{w}$, reflecting their importance. In order to evaluate the correct choice, the classification performance different variations concerning the amount of features of the data set is compared to the performance of the full data set. In order to analyze the applicability of the approach in practice and to evaluate the ability of the approach to identify formerly unknown failure examples, a performance evaluation is conducted by splitting the data set in a learning (70%) and test (30%) subset. Finally, feature ranking is applied to all individual processes and combined process vectors. The results of this will be evaluated in the following section 7.

6.3.1 SVM kernel & parameters for hyperplane by x-validation

A 10-fold cross-validation is applied on the TOM(CHEM) data set using RapidMiner (v5.3) as shown in Figure 74. In the first run, the original settings of the SVM function is used (kernel type DOT). The classification performance results of the derived confusion matrix (see a) in Figure 71) show a low classification performance of the minority class (negative). As can be observed, the results are uneven, but still significantly below the target threshold of 80% for both class recall and class prediction.

To show that the synthetic processes are applicable to classification by SVM algorithms with linear kernels and thus for feature ranking as described later, the results of the 10-fold cross-validation for DICK(CHEM), HARRY(CHEM), TD(CHEM) and TDH(CHEM) are presented in b) to e) in Figure 71. The classification performance of TD(CHEM) with a linear kernel is not very good. However, it is deemed acceptable within the scenario. The further application process is however based on TOM(CHEM) as it resembles real world data with all the challenges associated with real world data.

a) TOM(CHEM) (x-val; orig. para.; DOT kernel)

Accuracy:	84.12%	
118	22	84.29%
5	25	83.33%
95.93%	53.19%	

b) DICK(CHEM) (x-val; orig. para.; DOT kernel)

Accuracy:	95.88%	
136	1	99.27%
6	27	81.82%
95.77%	96.43%	

c) HARRY(CHEM) (x-val; orig. para.; DOT kernel)

Accuracy:	99.41%	
148	1	99.33%
0	21	100.00%
100.00%	95.45%	

d) TD(CHEM) (x-val; orig. para.r; DOT kernel)

Accuracy:	57.06%	
75	41	64.66%
32	22	40.74%
70.09%	34.92%	

e) TDH(CHEM) (x-val; orig. para.; DOT kernel)

Accuracy:	95.88%	
138	5	96.50%
2	25	92.59%
98.57%	83.33%	

Figure 71: Results of x-val classification performance of synthetic CHEM processes linear SVM

The next step in improving the classification model is to optimize the parameters of the SVM classifier applied. Optimizing parameters can improve the classification results significantly depending on the data set structure.

The optimization can either be done manually by continuously adapting the individual parameters, by observing and reacting to the changing results, or automated according to relevant parameters. In order to find the optimal classification model for the TOM(CHEM) data set, the automated optimization method was applied. The reasons are that the automated method allows for a structured and thorough testing of all parameter combinations with a direct feedback loop from the associated results.

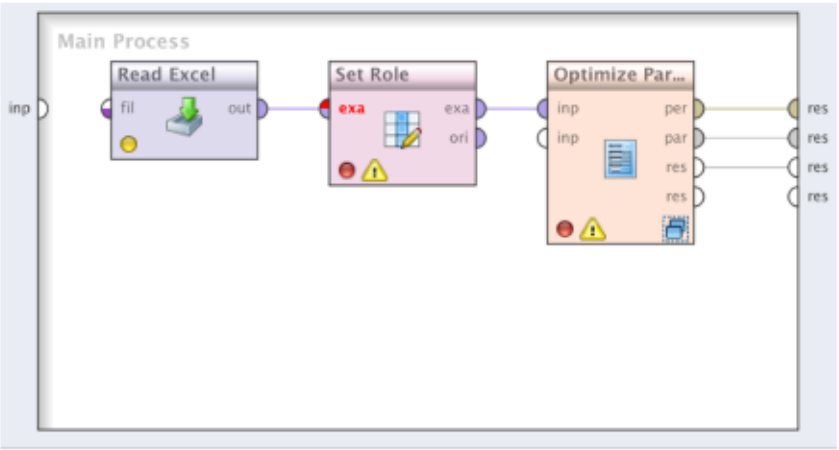

Figure 72: RapidMiner (v5.3) x-val process incl. optimization routine (top-level)

RapidMiner (v5.3) allows to incorporate a parameter optimization including the cross-validation process (component "optimize parameter (Grid)"). The function acts like a shell around the cross-validation component as can be seen in Figure 72, showing the overall process, Figure 73 and Figure 74 illustrating the containing components within the optimization component.

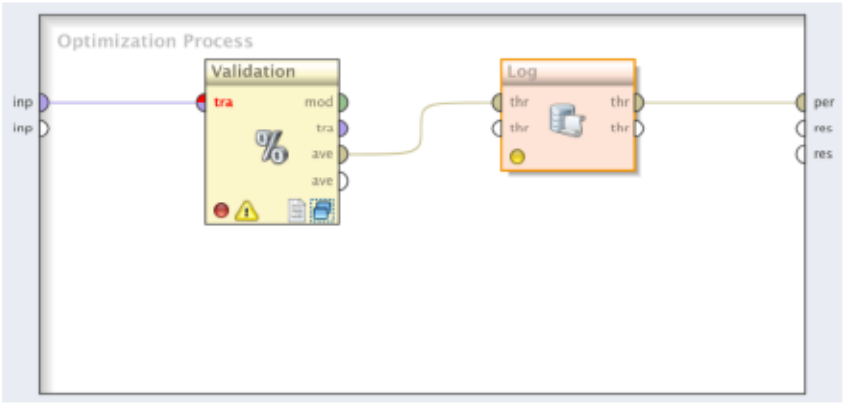

Figure 73: RapidMiner (v5.3) CHEM x-val process incl. log routine (second level)

The optimization component allows to select the parameters which shall be optimized for all components within the 'shell'. In this case, the parameters of the SVM classifier are in focus. For each parameter to be optimized, the range of optimization, number of steps and the scale (e.g., linear) may be chosen. Alternatively it is also possible to pre-define a list of values to be tested. As each parameter and each individual step adds to the number of calculations exponentially, it may be sensible to divide the optimization in several runs. This stands especially true as for each alternative n validations (n=10) are calculated. In this case, first an optimization finding the most suitable kernel type was conducted before the individual parameters were targeted. The final optimization routine calculated over 19000 operations to find the optimal combination of parameters for the data set.

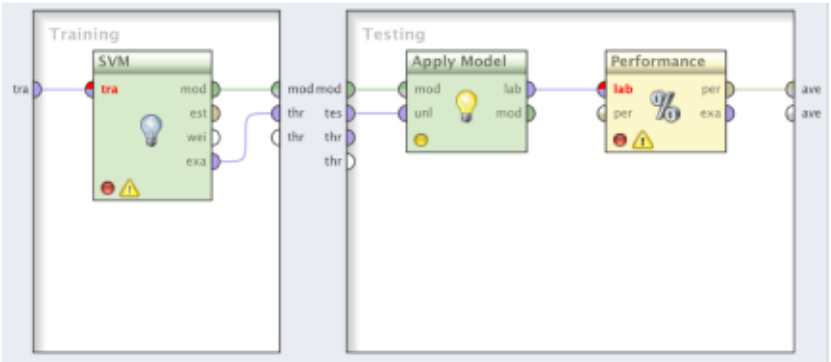

Figure 74: RapidMiner (v5.3) CHEM x-val process with SVM classifier (third level)

The log routine (see Figure 73) of the optimization routine tracks the different combinations of parameters and their performance during the process (see Figure 75). This is updated throughout the process in real time. The results of the final optimization run are compromised by the parameter set found to realize the optimal classification performance for the tested data set (see Figure 76) and a confusion matrix highlighting the classification performance using the optimal parameters set (see Figure 77). It can be observed that the classification performance improved compared to the first cross-validation (see Figure 71). In particular, looking at class

recall performance of the negative class. However, the performance is still under the target threshold of 80% for all indicators.

Learning r...	Kernel	Performance	Opti para	Kernel Gamma	Kernel Degree	L_pos	L_neg
-1	anova	1	SVM.kernel_	1.666666666666666	2.333333333333333	0.800000000000000	1.2000000000000
-1	anova	1	SVM.kernel_	3	1	2	2.7000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	1	0.400000000000000	0.6000000000000
-1	anova	0.941176470588235	SVM.kernel_	1.666666666666666	1.666666666666666	0.400000000000000	0.6000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	3	0.400000000000000	0.6000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	1	0.600000000000000	0.9000000000000
-1	anova	0.941176470588235	SVM.kernel_	1.666666666666666	3	0.400000000000000	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	3	1	0.600000000000000	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	2.333333333333333	1	0.800000000000000	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	3	1	0.800000000000000	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	1	1	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	1.666666666666666	1	1	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	1.666666666666666	1.800000000000000	1.2000000000000
-1	anova	0.941176470588235	SVM.kernel_	3	2.333333333333333	0.400000000000000	1.5000000000000
-1	anova	0.941176470588235	SVM.kernel_	3	1.666666666666666	0.600000000000000	1.8000000000000
-1	anova	0.941176470588235	SVM.kernel_	2.333333333333333	2.333333333333333	0.800000000000000	1.8000000000000
-1	anova	0.941176470588235	SVM.kernel_	1	1	1.200000000000000	1.8000000000000

Figure 75: Optimization of SVM parameter (ANOVA)

Result Overview | ParameterSet (Optimize Parameters (Grid)) | PerformanceVector (Performance) | Log

Text View | Annotations

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 81.18% +/- 6.86% (mikro: 81.18%)
ConfusionMatrix:
True:   positive        negative
positive:       111     20
negative:       12      27
-----precision: 79.35% +/- 21.98% (mikro: 69.23%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       111     20
negative:       12      27
-----recall: 57.00% +/- 29.43% (mikro: 57.45%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       111     20
negative:       12      27
-----AUC (optimistic): 0.865641025641026 +/- 0.106987621175594 (mikro: 0.865641025641026) (positive class: negative)
-----AUC: 0.865641025641026 +/- 0.106987621175594 (mikro: 0.865641025641026) (positive class: negative)
-----AUC (pessimistic): 0.865641025641026 +/- 0.106987621175594 (mikro: 0.865641025641026) (positive class: negative
]
SVM.kernel_gamma        = 3.0
SVM.kernel_degree       = 3.0
SVM.L_pos       = 0.8
SVM.L_neg       = 1.7999999999999998
```

Figure 76: Results of parameter optimization (ANOVA)

The difficulty of the classifier to correctly classify the minority class may be caused by the unbalanced data (minority ratio of 27.6%). In such cases, there are several established methods available. One that was found explicitly powerful is oversampling of the minority class using the Synthetic Minority Oversampling TEchnique (SMOTE) method (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Tang et al., 2009; Chawla, 2010; Farquad & Bose, 2012). The advantage of this technique is that it is specifically designed to avoid overfitting when oversampling is used. SMOTE operates in a feature space instead of a data space (Chawla, 2010). SMOTE oversampling by 200% to 500% of the minority class shows promising

results in improvement of classification of unbalanced data sets (Chawla, 2002). In some applications, SMOTE oversampling of 1000% and more showed good results in previous research (Akbani, Kwek & Japkowicz, 2004; Yun, Nan, Da & Bing, 2011). The application of the SMOTE method in WEKA using the build in function is described in more detail in scenario III SECOM (section 6.4.1).

accuracy: 84.12% +/- 8.34% (mikro: 84.12%)

	true positive	true negative	class precision
pred. positive	111	15	88.10%
pred. negative	12	32	72.73%
class recall	90.24%	68.09%	

Figure 77: Cross-validation performance with optimized parameters as shown in Figure 76 (ANOVA kernel type 3; kernel degree 3; C -1; Lpos 0.8; Lneg 1.8)

In this case the minority class was enhanced by 100% using the SMOTE method as it is incorporated in WEKA. The results of the subsequent cross-validation of the enhanced data set show significantly improved results. All indicators are above the target threshold of 80% (see Figure 78).

accuracy: 87.16% +/- 9.26% (mikro: 87.10%)

	true positive	true negative	class precision
pred. positive	103	8	92.79%
pred. negative	20	86	81.13%
class recall	83.74%	91.49%	

Figure 78: X-val of TOM(CHEM) with SMOTE (100%) and same previously identified optimal parameters (ANOVA kernel type 3; kernel degree 3; C -1; Lpos 0.8; Lneg 1.8)

However, the previous parameter optimization which is the basis for the chosen parameters used for the cross-validation of the enhanced data set was based on another data set with need for a higher penalty for negative misclassification. The enhanced data set is more balanced and thus may profit from different parameter settings. Optimizing the parameters and subsequent cross-validation for the enhanced data set shows that the results improve further (see Figure 79) presenting now a very good classification result.

accuracy: 89.94% +/- 7.31% (mikro: 89.86%)

	true positive	true negative	class precision
pred. positive	109	8	93.16%
pred. negative	14	86	86.00%
class recall	88.62%	91.49%	

Figure 79: X-val of TOM(CHEM) with SMOTE (100%) and optimized parameters (ANOVA kernel type 3; kernel degree 3; C 1; Lpos 1; Lneg 1)

6.3.2 Classification on previously unknown data

In this scenario, three possible approaches to select the learning and test set are applied and illustrated in order to evaluate the classification performance on previously unknown data. First the learning and test set is selected using different approaches before their classification performance is compared.

6.3.2.1 Definition of learning set – Random

The approach described in this subsection is based upon the annotation of the data set in positive (>=39 Yield) and negative (<39 Yield) examples. Two variations are utilized, one defining the learning set as 70% of the positive and 70% of the negative examples, chosen at random by a RapidMiner (v5.3) sampling process.

From the *negative* examples, the following 15 (out of 51; exact ratio 29,4%) were chosen as the ***test data set***: *(example no.)* 26; 29; 30; 62; 87; 105; 136; 139; 141; 152; 154; 155; 157; 164; 165.

From the *positive* examples, the following 37 (out of 125; exact ratio 29,6%) were chosen as the ***test data set***: *(example no.)* 2; 7, 10; 11; 12; 14; 17; 19; 23; 31; 40; 41; 46; 47; 53; 59; 61; 67; 70; 72; 83; 84; 97; 113; 118; 122; 128; 130; 140; 142; 143; 144; 151; 168; 172; 174; 176.

The learning set is composed from the remaining positive and negative examples. The two separate tables (positive and negative examples) for each learning set are combined to a single one before proceeding to the next step.

The other uses the same process but the inverted ratio of 30% for the learning set and 70% for the test set. This reflects the reality in some application cases better. This allows using the same sets as before by just relabeling the test for learning and vice versa.

6.3.2.2 Definition of learning set – timely

Resembling a manufacturing process, the learning and test set are selected in a timely manner (timely sequence in process) in this subsection (compare Figure 92). The reasoning is that the first 70% of all examples which are in timely succession, represent the learning set. As the classifier is trained by these, the following 30% of the examples are the test set. This test set is representing new examples which are monitored based on the classifier model trained by the previous examples. In this case, the ratio of negative and positive examples is not the same as it is in the random selection.

The learning set consists of the first 119 examples, incorporating 22 negative examples (18.5%). The test set resembles 51 examples following in timely succes-

sion. Of those, 25 are negative, leading to a ratio of 49%. This shows that the smaller test set (30%) includes more negative results than the learning set (70%). This is representing an extra challenge for the classifier for identifying negative results correctly as the learning set possibly does not represent a majority of the possible negative results. Therefore, the constructed hyperplane may have difficulties with correctly classifying previously unknown failure examples.

6.3.2.3 Definition of learning set – Cluster Analysis

In this subsection the definition of the learning set by applying a cluster analysis is illustrated. The rationale behind this approach is that the two extreme clusters within a data set may present a good data set for the learning phase in case no expert knowledge is available to select suitable examples for the learning set.

The identified clusters within the data set are:

- ***Cluster one:*** *(example no.)* 136, 121, 153, 135, 165, 137, 166, 167.
- ***Cluster two:*** *(example no.)* 7, 23, 24, 36, 65.

It can be seen that the learning set (13 examples) is significantly smaller than the test set (163 examples). Which may have an influence on the classification results.

6.3.3 Compilation of SVM operation and output data

In order to create the classifier model, the algorithm is trained using the selected learning set. The process is modeled in RapidMiner (v5.3) (see Figure 80). After reading the learning data set, including the positive/negative labels in the systems, roles are assigned. In this case, the roles are the label (positive/negative) and the identifier (No. of example). Following, the SVM is applied to create the model (***w*** vectors for hyperplane) and the model is exported in a *.mod file as well as forwarded to the results output of RapidMiner.

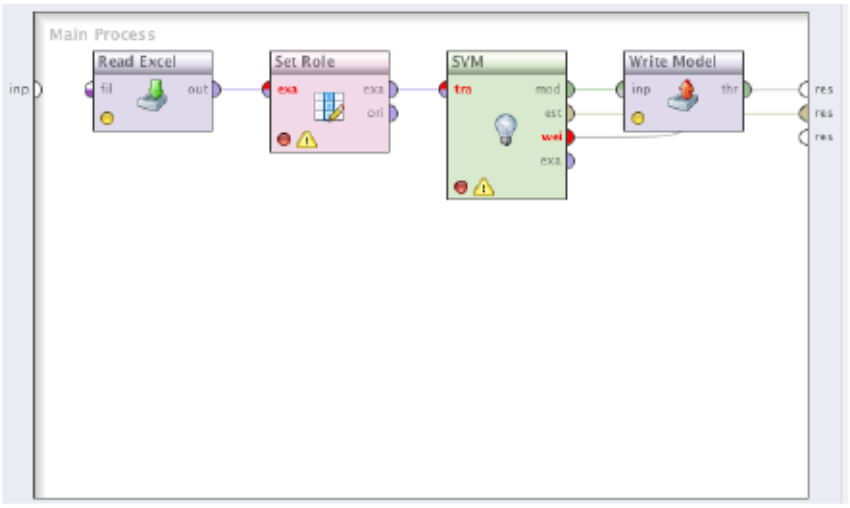

Figure 80: RapidMiner (v5.3) model generation process

The selected parameters for the SVM algorithm are the same as were identified as optimal for the data set in the focus during the 10-fold cross-validation described before. In this case, the cross-validation using the SVM classification was executed with the same parameter identified before (see Figure 79).

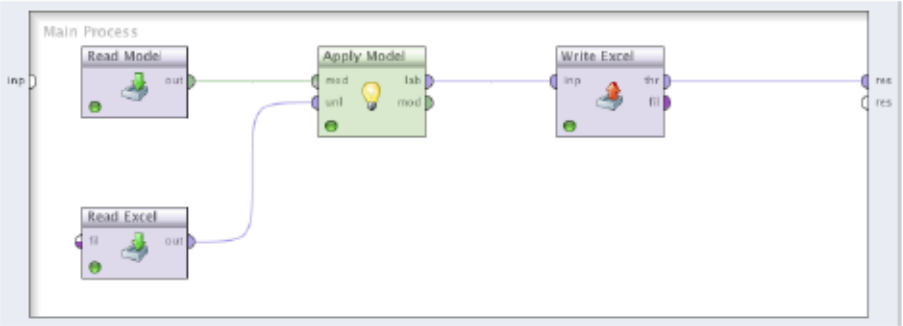

Figure 81: Applying the trained model on test data set.

Once the model is successfully generated and stored in the specified *.mod file, the next phase can be tackled. In order to evaluate the created model, it is applied to the test data set. In this case, the test data set has no labels and the examples are classified according to the created model. The process, modeled in RapidMiner (v5.3), is illustrated in Figure 81.

Result Overview | ExampleSet (Read Excel)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (124 examples, 3 special attributes, 57 regular attributes)

Row No.	confidence...	confidence...	prediction(...	BiologicalM...	BiologicalM...	BiologicalM...	BiologicalM...	BiologicalM...	Biological
1	0.4798309	0.5201690	Positive	0.9690885	0.1515659	0	0.5765566	0.0001493	0.16666
2	0.4803557	0.5196442	Positive	0.8810299	0.7885906	0.4938909	0.0000769	0.0001457	0.66773
3	0.4791249	0.5208750	Positive	0.8810299	0.7885906	0.4938909	0.0000769	0.0001457	0.66773
4	0.4803365	0.5196634	Positive	0.3669388	0.9205816	0.7180451	0.8783988	0.0001112	0.74866
5	0.4816880	0.5183119	Positive	0.8880543	0.6426174	0.3919172	0.0000878	0.0002036	0.56602
6	0.4672660	0.5327339	Positive	0.7295192	0.9356823	0.7091165	0.8815817	0.0001472	0.75186
7	0.4706663	0.5293336	Positive	0.7295192	0.9356823	0.7091165	0.8815817	0.0001472	0.75186
8	0.4804093	0.5195906	Positive	0.9079003	0.8031319	0.6127819	0.0000497	0.0001293	0.65122
9	0.4728853	0.5271146	Positive	0.9536222	0.8982102	0.6000939	0.8821508	0.0001738	0.74387
10	0.4774966	0.5225033	Positive	0.8769888	0.7511185	0.5267857	0.8821929	0.0001781	0.63312
11	0.4725050	0.5274949	Positive	0.4983285	0.7030201	0.4816729	0.0000984	0.9949024	0.60809
12	0.4809858	0.5190141	Positive	0.3978503	0.7197986	0.4844924	0.0000994	0.9970939	0.62300
13	0.4831614	0.5168385	Positive	0.3978503	0.7197986	0.4844924	0.0000994	0.9970939	0.62300
14	0.4777229	0.5222770	Positive	0.8854942	0.8366890	0.6447368	0.0000512	0.0001067	0.67465
15	0.4651600	0.5348399	Positive	0.9845336	0.5234899	0.3764097	0.3680494	0.0001355	0.51703
16	0.4792103	0.5207896	Positive	0.9085561	0.9390380	0.7739661	0.0001553	0.0002505	0.85410
17	0.4779591	0.5220408	Positive	0.8861289	0.4133109	0.1889097	0.9461026	0.0001648	0.35303

Figure 82: Exemplary results of the application of the trained model in RapidMiner (v5.3)

It starts with loading the respective model (*.mod) and the test data set in form of an adjusted excel table with removed labels. The operator ‘apply model’ then classifies the examples according to the learned model and illustrated an output with the predicted classification for each example (see Figure 82).

Additionally, an excel table is created as an output providing an extra column with the predicted classification (positive/negative). This allows to compare the predic-

tion to the actual classification for each example. The results for the three variations are summarized in the following Figure 83:

random (LS 70% & TS 30%)			**time sequence**			**extreme clusters**			**random** (LS 30% & TS 70%)		
Accuracy:	78.85%		Accuracy:	68.63%		Accuracy:	78.53%		Accuracy:	78.23%	
9	6	60.00%	11	14	44.00%	17	28	37.78%	14	22	38.89%
5	32	86.49%	2	24	92.31%	7	111	94.07%	5	83	94.32%
64.29%	84.21%		84.62%	63.16%		70.83%	79.86%		73.68%	79.05%	

Figure 83: Accuracy of SVM classifier models for learning/test set variations (ANOVA; kernel gamma 3; kernel degree 3; C 1)

It can be observed that the classification results of all variations are significantly worse than the cross-validations classification results. This was to be expected. As suspected, the variations with a smaller learning set did not perform as well as the ones using more examples to train the SVM classification algorithm model. This corresponds with the common practice of using a 70% (learning) and 30% (test) split for evaluation purposes.

Another explanation why the variant using cluster analysis did not perform that strongly may be that it is due to the lack of expert knowledge which is needed to identify the relevant clusters. This is not available for this data set.

The results for the test of formerly unknown data (test set 30%) indicate that an implementation of a SVM classification model implemented in a monitoring system may be possible. By constantly growing the learning set with an increasing number of negative results the performance is expected to improve over time.

In the next section, the identification of relevant state characteristics (state drivers) using feature selection for individual processes and combined process vectors is evaluated (hypotheses 1.1 & 1.2).

6.3.4 Feature ranking using SVM classifier

After the classification performance is acceptable for the CHEM data set with the adjusted kernel settings and parameters, a feature selection is conducted. As described previously in section 5.2.2, a SVM based feature selection (Guyon et al., 2002) is applied to rank the features based on their importance. The importance is determined by the weight vector $\boldsymbol{w}$ of each feature.

The feature selection method is based on a linear kernel, therefore the classification performance of such a kernel for the specific data set is relevant. The first result can be seen in Figure 71. Applying some parameter adjustment, the classification results are acceptable (see Figure 84). Hence, the applicability of the proposed feature selection method based on linear SVM on the CHEM data set is assumed.

accuracy: 84.35% +/- 4.61% (mikro: 84.33%)

	true positive	true negative	class precision
pred. positive	104	15	87.39%
pred. negative	19	79	80.61%
class recall	84.55%	84.04%	

Figure 84: Results x-val DOT kernel (SMOTE 100%; C -1.0; conv. eps. 0.005; Lpos 1.8; Lneg 2.0)

The open source software toolkit WEKA includes an implemented feature evaluation function called 'SVM Attribute Eval' (see Figure 85) (Witten et al., 2011; Eibe, Hall & Holland, 2014). This function resembles the methodology described by Guyon et al. (2002) and allows an easy application on available data sets. One limitation of the WEKA function is that the weight vectors $\boldsymbol{w}$ are not available as an output.

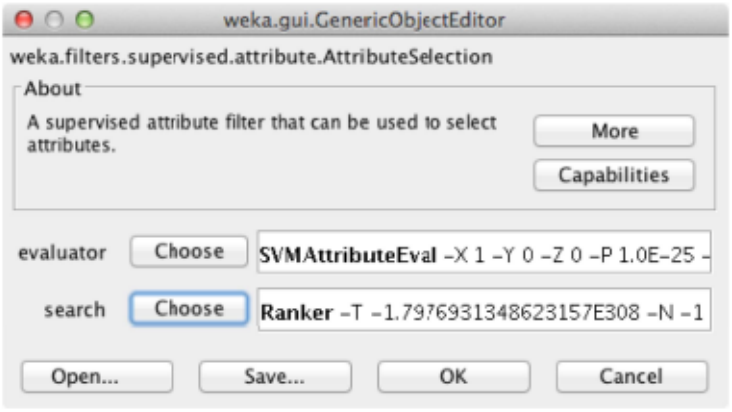

Figure 85: WEKA SVM feature evaluation function

However, WEKA has different requirements when it comes to importing data sets. The data set needs to be in the *.arff format. RapidMiner however allows for easy transformation of *.csv and *.xlsx data sets in the respective format.

Once loaded, the software allows applying the feature evaluation with relative ease. A few steps are needed before the evaluation can be started. First, the existing special column 'Identifier', supplying a unique number to each example has to be removed before a feature ranking can be conducted. Furthermore, the attribute 'SVM' has to be manually selected before application of the function as in this data set, the 'SVM' attribute is not at the end of the attribute list, which would allow for an automatic recognition.

There are two components which have to be selected together after choosing 'AttributeSelection' in WEKA. One is the 'SVMAttributeEval' and the other is called 'Ranker'. The former is responsible for determining the feature weights (= importance/relevance) and the latter for preparing the ranking based on the previously determined weights (see Figure 85).

Once these pre-processing steps are done, the parameters of the function can be adjusted. One of the most important ones is the 'numToSelect'. This parameter describes what number of the highest ranked features shall remain in the data set after the evaluation is completed. For ranking purposes this is set to '-1', leading to a complete ranking by importance (weight) and no elimination of features. By doing so, all attributes are rearranged starting with the one of highest weight and ending with the least important one (lowest weight).

For the evaluation of the classification performance after feature selection, the ranker threshold is set to 5; 10; 15; 20 and 30 features. By doing so, different data sets are derived who's classification performance in cross-validation is individually assessed (see Figure 86). This allows to evaluate the performance of the feature selection in identifying the most relevant features (state drivers).

	No SMOTE			100%			200%		
57	Accuracy:	81.18%		Accuracy:	82.49%		Accuracy:	83.33%	
	118	27	81.38%	105	20	84.00%	95	16	85.59%
	5	20	80.00%	18	74	80.43%	28	125	81.70%
	95.93%	42.55%		85.37%	78.72%		77.24%	88.65%	
5	Accuracy:	78.24%		Accuracy:	78.70%		Accuracy:	82.20%	
	117	31	79.05%	91	15	85.85%	85	9	90.43%
	6	16	72.73%	31	79	71.82%	38	132	77.65%
	95.12%	34.04%		74.59%	84.04%		69.11%	93.62%	
10	Accuracy:	82.94%		Accuracy:	82.95%		Accuracy:	84.47%	
	116	22	84.06%	98	12	89.09%	90	8	91.84%
	7	25	78.13%	25	82	76.64%	33	133	80.12%
	94.31%	53.19%		79.67%	87.23%		73.17%	94.33%	
15	Accuracy:	83.53%		Accuracy:	84.33%		Accuracy:	85.61%	
	119	24	83.22%	108	19	85.04%	93	8	92.08%
	4	23	85.19%	15	75	83.33%	30	133	81.60%
	96.75%	48.94%		87.80%	79.79%		75.61%	94.33%	
20	Accuracy:	84.71%		Accuracy:	82.95%		Accuracy:	85.23%	
	119	22	84.40%	107	21	83.59%	91	7	92.86%
	4	25	86.21%	16	73	82.02%	32	134	80.72%
	96.75%	53.19%		86.99%	77.66%		73.98%	95.04%	
30	Accuracy:	82.94%		Accuracy:	86.64%		Accuracy:	86.36%	
	120	26	82.19%	108	14	88.52%	98	11	89.91%
	3	21	87.50%	15	80	84.21%	25	130	83.87%
	97.56%	44.68%		87.80%	85.11%		79.67%	92.20%	

Figure 86: Matrix of results of x-val on different feature numbers and SMOTE oversampling

The comparison of classification performance of different settings (no. of features & SMOTE oversampling) derived using cross-validation with an ANOVA kernel (orig. parameters) shows interesting results. So the best overall performance (accuracy) is reached not by the expected full feature set with 200% SMOTE oversampling, but by the data set with 30 features and 100% SMOTE oversampling. Surprising is that even a significantly reduced data set (5 features) performs relatively well. The data set with 15 features outperforms the full data set (57 features) in the variation without SMOTE and SMOTE 100%. This confirms that by select-

ing relevant features based on the Guyon et al. (2002) feature selection technique, the process may be described sufficiently to allow for good classification results.

As stated before, the applied SVM based feature selection method utilizes the weight vector $\boldsymbol{w}$ as a ranking criterion. The weight vectors $\boldsymbol{w_i}$ represent a distinct number (small fraction) of vectors of the training set used to construct the hyperplane (decision boundary). As described before (section 5.2.1.1) the hyperplane is constructed to leave the largest (maximum) margin between the two classes. The so-called 'support vectors' (hence the name SVM) are located on the margin and define the hyperplane (decision boundary) (Guyon et al., 2002). The smaller the weight vector $\boldsymbol{w}$ value is, the less relevant is the feature for classification decisions.

Through Recursive Feature Elimination (RFE) based on the weight vector $\boldsymbol{w}$ value, the feature with the lowest $\boldsymbol{w}$ value is eliminated in each run (Guyon et al., 2002). This is different to a ranking based on one run with the complete feature set. The RFE approach takes changing relevancy of features when the number of features I reduced into account. It is possible to eliminate multiple features simultaneously in each run if computing efforts make that necessary. The utilized WEKA function allows to specify the number of (lowest ranking) features to be eliminated in each run. In this dissertation the amount of to-be-eliminated features for each run is set to '1' as originally suggested (Guyon et al., 2002).

As a test of the general accuracy of the feature ranking based on SVM weight vectors $\boldsymbol{w}$, a basic evaluation of the classification performance using a data set with the lowest ranked 10 and 20 features with an ANOVA kernel (orig. parameters) is illustrated in the following Figure 87 a) and b). The results confirm that the feature ranking is working probably as the two resulting confusion matrices show significantly lower classification performance than the ones of the comparable set using the highest ranked 10 and 20 features.

a) TOM(CHEM) (10 lowest ranked features)

Accuracy:	74.12%	
116	37	75.82%
7	10	58.82%
94.31%	21.28%	

b) DICK(CHEM) (20 lowest ranked features)

Accuracy:	76.47%	
113	30	79.02%
10	17	62.96%
91.87%	36.17%	

Figure 87: X-val class. perf. of TOM(CHEM) with lowest ranking 10 & 20 features selected

Finally, the classification performance is tested by again dividing the data sets in learning (70%) and test (30%) set. This way the ability of a trained classification model to identify classes in previously unknown data is evaluated after feature selection and compared to one another (incl. the performance of the complete set). The results are also compared to the previous classification performance as illustrated in Figure 83.

The comparison of results (confusion matrix) of the trained model application (learning 70% & test 30%) in various settings show that the results of data sets with a reduced feature set can outperform the data set with a full feature set. The best accuracy is reached by a data set with 15 features, 100% SMOTE oversampling on the learning set and an ANOVA kernel with original parameter settings almost even meeting the target threshold for cross-validation classification performance. This is a very good result for classification on a new, formerly unknown data set. This confirms that the feature selection is able to identify relevant state drivers.

	Learning (70%) & Test (30%) - random								
	ANOVA (kernel gamma 3; kernel degree 3; C 1; con eps 0.001)						ANOVA (orig. para.)		
	No SMOTE			100%			100%		
57	Accuracy:	71.15%		Accuracy:	78.85%		Accuracy:	78.85%	
	0	15	0.00%	11	4	73.33%	11	4	73.33%
	0	37	100.00%	7	30	81.08%	7	30	81.08%
	#DIV/0!	71.15%		61.11%	88.24%		61.11%	88.24%	
5	Accuracy:	75.00%		Accuracy:	76.92%		Accuracy:	76.92%	
	4	11	26.67%	12	3	80.00%	12	3	80.00%
	2	35	94.59%	9	28	75.68%	9	28	75.68%
	66.67%	76.09%		57.14%	90.32%		57.14%	90.32%	
10	Accuracy:	71.15%		Accuracy:	73.08%		Accuracy:	86.54%	
	5	10	33.33%	6	9	40.00%	12	3	80.00%
	5	32	86.49%	5	32	86.49%	4	33	89.19%
	50.00%	76.19%		54.55%	78.05%		75.00%	91.67%	
15	Accuracy:	78.85%		Accuracy:	82.69%		Accuracy:	88.46%	
	9	6	60.00%	11	4	73.33%	13	2	86.67%
	5	32	86.49%	5	32	86.49%	4	33	89.19%
	64.29%	84.21%		68.75%	88.89%		76.47%	94.29%	
20	Accuracy:	78.85%		Accuracy:	80.77%		Accuracy:	75.00%	
	8	7	53.33%	9	6	60.00%	9	6	60.00%
	4	33	89.19%	4	33	89.19%	7	30	81.08%
	66.67%	82.50%		69.23%	84.62%		56.25%	83.33%	
30	Accuracy:	71.15%		Accuracy:	69.23%		Accuracy:	69.23%	
	0	15	0.00%	7	8	46.67%	7	8	46.67%
	0	37	100.00%	8	29	78.38%	8	29	78.38%
	#DIV/0!	71.15%		46.67%	78.38%		46.67%	78.38%	

Figure 88: Accuracy of SVM classifier models for learning/test set (random) variations with different feature selection variations

The results describing the feature ranking for all processes and process combinations are discussed in detail in section 7. A complete illustration of the detailed ranking results is presented in Table 20 in the Annex.

The effects of feature selection and SMOTE oversampling on the classification performance of formerly unknown data organized in timely sequence is also substantial. Comparing the results presented in Figure 89 with the previous results presented in Figure 83, it shows that the performance of the variation with 15 features and 200% SMOTE oversampling can be considered good and significantly better than the one without this pre-processing steps.

	Learning (70%) & Test (30%) - time sequence								
	ANOVA (kernel gamma 3; kernel degree 3; C 1; con eps 0.001)								
	No SMOTE			100%			200%		
FS10	Accuracy:	56.86%		Accuracy:	72.55%		Accuracy:	72.55%	
	7	18	28.00%	17	8	68.00%	17	8	68.00%
	4	22	84.62%	6	20	76.92%	6	20	76.92%
	63.64%	55.00%		73.91%	71.43%		73.91%	71.43%	
FS15	Accuracy:	64.71%		Accuracy:	70.59%		Accuracy:	74.51%	
	10	15	40.00%	13	12	52.00%	17	8	68.00%
	3	23	88.46%	3	23	88.46%	5	21	80.77%
	76.92%	60.53%		81.25%	65.71%		77.27%	72.41%	
FS20	Accuracy:	62.75%		Accuracy:	64.71%		Accuracy:	72.55%	
	9	16	36.00%	12	13	48.00%	16	9	64.00%
	3	23	88.46%	5	21	80.77%	5	21	80.77%
	75.00%	58.97%		70.59%	61.76%		76.19%	70.00%	

Figure 89: Accuracy of SVM classifier models for learning/test set (time sequence) variations with different feature selection and SMOTE variations

Focusing on the best performing variant with 15 features and SMOTE 200% over-sampling applied on the learning set, a slight optimization is possible by changing the kernel gamma (Figure 90).

Accuracy:	76.47%	
17	8	68.00%
4	22	84.62%
80.95%	73.33%	

Figure 90: Para. optimization for TOM(CHEM) FS15 & SMOTE 200% (kernel gamma: 2)

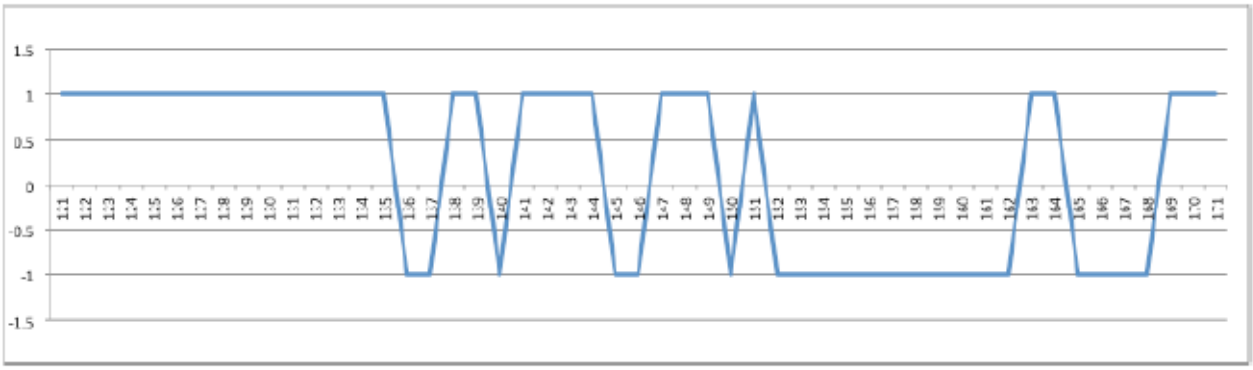

Figure 91: Time plot of predicted state change (‚pass'=1/'fail'=-1) of TOM(CHEM) process

When the predicted 'fail' examples are plotted in timely sequence (see Figure 91), one can observe that the process seems relatively stable at the beginning. However, then the process begins to become more unstable until it produces 11 fail examples in a row (example 152 to example 162).

Comparing this predicted time plot with the original (see Figure 92), it can be observed that the massive disruption between 'example 152' and 'example 162' is correctly predicted by the classifier. Such a prediction, especially with a significantly reduced feature set may allow the process owner to preemptively adjust the process and reduce the risk of such a relatively long period of manufacturing products with not sufficient quality.

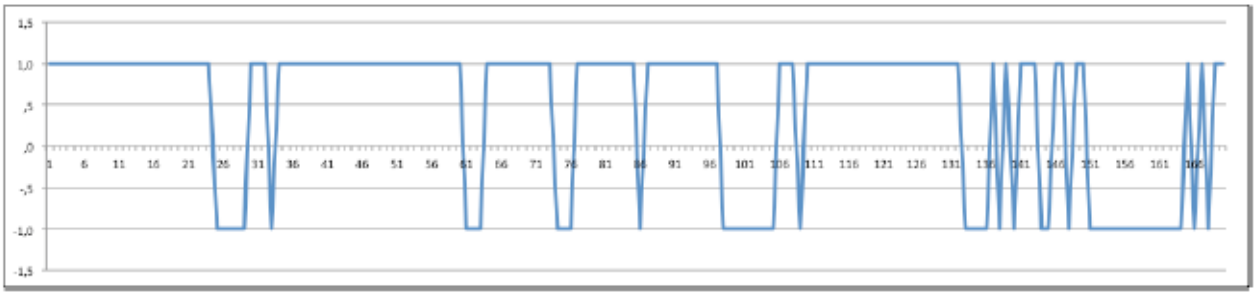

Figure 92: Original time plot of changing state (‚pass'=1/'fail'=-1) of TOM(CHEM) process

In the next sub-section, the presented approach is applied to scenario III, a highly complex, exemplary process from the semiconductor manufacturing domain.

6.4 Scenario III – SECOM

In this section, the SVM algorithm is set up as a classifier and applied on the preprocessed SECOM data set as described within the research plan. At first a suitable configuration of the SVM algorithm, given the specific nature of the SECOM data set, is identified through n-fold cross-validation (n=10) in the following subsection. Subsequently, a learning set is defined before the SVM algorithm is applied on test and evaluation set. The results are presented in section 7.1.

6.4.1 SVM kernel & parameters for hyperplane by x-validation

Before the hyperplane can be constructed, a suitable SVM kernel method and corresponding parameter settings have to be identified. These factors depend strongly on the available data and its structure. The previously described splitting of the data set in training, evaluation and test set is automated by applying n-fold cross-validation (n=10) in RapidMiner (v5.3). This process automatically divides the data set in 9/10 and 1/10 packets n-times for evaluation.

The first step is importing the complete and normalized SECOM data set (approach 2 variant 2) in the process (component "Read Excel") and mark the label ("good"/"bad") and identifier (No. of example) (component "Set Role"). The results show that all 1209 examples and 528 features/attributes are successfully imported and the label and identifier are correctly assigned.

The process component "validation" is available as a pre-set building block in RapidMiner (v5.3). The building block has to be adapted. Therefore the classification algorithm is set to "SVM". The parameters for both, the cross-validation and the SVM classifier can be adjusted. For the cross-validation they are set to n=10. For the SVM the standard parameters for a polynomial kernel are used.

The result of running the previously introduced cross-validation process is a confusion matrix. The confusion matrix illustrates how well the learning data can be separated by a hyperplane using various SVM kernels. Ideally, both classes should reach a result of 80% or above. This indicates that the $\boldsymbol{w}$ values represent a set of good features (state drivers). If the results are significantly below 80%, then the SVM kernel choice does not correspond well with the structure of the data set and needs to be adapted. The same is true for the parameters of the chosen kernel, which may have to be adjusted to create a good fit and thus an assumingly good weight vector $\boldsymbol{w}$.

accuracy: 84.54% +/- 25.16% (micro: 84.53%)

	true negative	true positive	class precision
pred. negative	9	112	7.44%
pred. positive	75	1013	93.11%
class recall	10.71%	90.04%	

Figure 93: SECOM cross-validation with RapidMiner (v5.3) first results

In this case, the first results of the previously presented RapidMiner (v5.3) cross-validation process of the SECOM data set are significantly below the target threshold of 80% over class recall as well as class precision (see Figure 93). The kernel for the SVM was chosen as polynomial with all parameters used as pre-set, except the kernel degree, which was set to 3.0.

This results could not be improved significantly by changing the kernel and/or the SVM parameters. Therefore, a step back to pre-processing has to be taken.

6.4.1.1 Under- and oversampling

In this case, it can be assumed that the bad performance was partly based on the SECOM data set's highly unbalanced (ratio of fails/pass of 6,95%) data (see Figure 94. Unbalanced data is fairly common in many application areas (Provost, 2000; Evgeniou & Pontil, 2001; Li, Hu & Hirasawa, 2008). SVM algorithms may experience problems when it comes to classification tasks of some unbalanced data (Li & Shawe-Tayler, 2003; Li et al., 2008; Tang et al., 2009; Wang & Japkowicz, 2010; Choi, 2010). As can be observed in Figure 94, the negative examples are furthermore not equally distributed of the duration of the monitoring. The concentration of 'fail' examples (minority class) is higher during the early runs of the process.

Figure 94: Time plot of changing state (‚pass'=1/'fail'=-1) of SECOM process

There are different approaches to handle the problem suggested by literature. For example, Veropoulos, Cristianini & Campbell (1999) suggest to introduce weights (penalties) for the misclassification of the underrepresented class, whereas Wang, Liu, Li & Zhou (2004) suggest to handle such classification task with a one-vs.-one classifier rather than a one-vs.-rest. A rather promising approach for handling unbalanced data sets is under- or oversampling of the respective class (Chawla, 2010). However, random under- and oversampling can lead to problems like deleting of important examples (random undersampling) or over-fitting (random oversampling) (Chawla, 2010). Next, first undersampling of majority class within the SECOM data set will be analyzed before oversampling of the minority class is looked into more closely.

6.4.1.1.1 Undersampling

In order to address this issue of unbalanced data by under-sampling, a new component is added to the cross-validation process. It can be seen in the results (see Figure 93), that the classification performance for the underrepresented class ('1'/'fail') are not acceptable. Therefore, a sampling method is installed prior to the cross-validation of the classifier in order to even out the ratio of the learning set. The chosen sampling mechanism is based on the Kennard-Stone algorithm. The Kennard-Stone algorithm allows to identify outliers of the data set, which are assumed to be good representations of the class (De Groot, Postma, Melssen & Buydens, 1999). This sampling method is just applied to the overrepresented class ('-1'/'non fail').

In order to integrate the Kennard-Stone sampling algorithm in the previously introduced cross-validation process in RapidMiner (v5.3), the SECOM data set has to be divided in two. One data set containing all examples of class '1' ('fail'), 84 examples in total and the other all examples of class '-1' (non fail'), 1125 examples in total. As described before, both data sets have to be imported and the respective roles have to be assigned (label & identifier). For the data set representing the over-represented class '-1', the Kennard-Stone sampling is added, allowing to define the number of examples being sampled either absolute or relative. In this case the number is chosen absolute and set to 84, matching the number of examples of the other class. Conducted tests with different variations of the sampling number did not lead to improved classification results. After this process step, the two data

sets are merged again to one complete data set containing both classes (component 'append').

accuracy: 52.34% +/- 4.52% (mikro: 52.38%)

	true positive	true negative	class precision
pred. positive	30	30	50.00%
pred. negative	60	69	53.49%
class recall	33.33%	69.70%	

Figure 95: First results after integrating Kennard-Stone sampling in cross-validation process

After this addition, the cross-validation process was run again, setting the SVM kernel to polynomial and using the pre-set parameters except the kernel degree, which was set to 3.0.

It can be observed that the results are improved but still significantly below the target threshold. However, it was possible through parameter optimization, to reach the threshold. Despite that, during the later model development and evaluation utilizing the learning and test set, the classification performance of this approach was not acceptable. This leads to the application of the oversampling approach which is illustrated in the following subsection.

6.4.1.1.2 Oversampling

Before starting to assess the possible measures for oversampling, the deleted negative examples are assessed again if the feature eliminating conducted in preprocessing may result in "complete" (no missing values) negative examples. The following 15 additional negative examples were identified and added to the existing data set: example 51; example 189; example 236; example 239; example 283; example 322; example 327; example 328; example 345; example 407; example 442; example 602; example 606; example 635; example 710.

a) no oversampling (x-val; orig. para. polyn. kernel)

Accuracy:	92.56%	
1	7	12.50%
83	1118	93.09%
1.19%	99.38%	

b) oversampl. neg. class x3 (x-val; orig. para. polyn. kernel)

Accuracy:	83.88%	
35	5	87.50%
217	1120	83.77%
13.89%	99.56%	

c) oversampl. neg. class x6 (x-val; orig. para. polyn. kernel)

Accuracy:	81.34%	
209	9	95.87%
295	1116	79.09%
41.47%	99.20%	

d) oversampl. neg. class x9 (x-val; orig. para. polyn. kernel

Accuracy:	87.29%	
531	14	97.43%
225	1111	83.16%
70.24%	98.76%	

e) oversampl. neg. class x10 (x-val; orig. para. polyn. kernel)

Accuracy:	90.69%	
667	10	98.52%
173	1115	86.57%
79.40%	99.11%	

Figure 96: Random oversampling of minority class x-val results in RapidMiner (v5.3)

It has been stated that random oversampling may lead to overfitting (Chawla, 2010). In a first attempt, random oversampling was conducted to observe the effect it has on the results. In Figure 96, the confusion matrix of random oversampling of

the minority class is illustrated. With 10-fold oversampling the threshold is almost reached. In this case, the original parameter settings of the Kernel choice in RapidMiner (v5.3) was used to allow a better comparability of the results. With optimized parameters, the results would most likely vary only marginally. Therefore, parameters optimization is applied after the best variation was identified.

In a next step, the parameters of the kernel are adjusted to improve the results and bring them over the threshold of 80%. The results indicate an almost perfect classification performance (see Figure 97). This result already indicates a possible overfitting of the generated model.

accuracy: 99.54% +/- 0.48% (mikro: 99.54%)

	true negative	true postive	class precision
pred. negative	840	9	98.94%
pred. postive	0	1116	100.00%
class recall	100.00%	99.20%	

Figure 97: results x-val of 10-fold random oversampling after parameter adjustment

To test the performance of the random oversampling approach, a learning set (70%) and test set (30%) of the original data set is divided, holding up the original ratio of both classes. The learning set is then pre-processed by randomly oversampling the negative examples (minority class) ten times. Then a model is created with the learning set and the classification performance is evaluated using the test set in RapidMiner (v5.3). The results can be observed in Figure 98.

Accuracy:	92.01%	
1	24	4.00%
5	333	98.52%
16.67%	93.28%	

Figure 98: Classification performance of created model by random oversampling minority class

The results show poor classification performance on previously unknown data. This may support the previous suspicion of a possible overfitting problem by applying random oversampling through multiplication of the minority class or again be a sign for a high diversity within the example population of the minority class. However, instead of random oversampling a rather promising oversampling technique called SMOTE is applied. The advantage of this technique is, as previously stated (section 6.3.1) that it is specifically designed to avoid overfitting when oversampling is used.

SMOTE is an established method to handle unbalanced data sets. The method is implemented as a built in function in the WEKA toolkit. Before applying the SMOTE function on the SECOM data set, an additional 15 minority class exam-

ples are added to the SECOM data set. Those examples are derived after the preprocessing step "attribute deletion" during which attributes containing missing values are eliminated. This step allowed for 15 minority class examples to be subsequently added as they do not contain any missing values. The data set then contains 1224 examples of which 99 examples fall under the 'fail' class.

After applying the SMOTE technique in WEKA, the ratio between the classes is now 34.6%. This is still not a completely balanced data set but the ratio should allow for a better model development and thus classification performance.

A 10-fold cross validation is then conducted on the now complete data set in RapidMiner (v5.3). The results are promising as they are all over or close to the target threshold of 80% in a first run with the original parameters of an AVM algorithm with a DOT kernel (see Figure 99). Similar as before, in this case the original parameter settings of the kernel choice in RapidMiner (v5.3) were applied to allow a better comparability of the results. With optimized parameters, the results would most likely improve further marginally. However, parameter optimization is very individual for every data set. This reasoning is applied in all cases where original parameter settings are used without optimization.

accuracy: 87.79% +/- 3.40% (mikro: 87.78%)

	true negative	true positive	class precision
pred. negative	519	135	79.36%
pred. positive	75	990	92.96%
class recall	87.37%	88.00%	

Figure 99: Results x-val SMOTE oversampling DOT kernel

The results of the DOT kernel are already good and, given that no parameter adjustment was yet conducted significantly better than with simple oversampling by multiplication of the minority class as illustrated previously. This shows that the following feature ranking based on linear SVM is applicable. However, it is possible that this may still represent a similar case of overfitting as the oversampling by multiplication of the existing examples presents (see Figure 96), especially in cases of SMOTE oversampling 500% and higher.

	No SMOTE			200%			500%		
528 (original)	Accuracy:	91.91%		Accuracy:	92.76%		Accuracy:	95.35%	
	0	0		196	2	98.99%	518	4	99.23%
	99	1125	91.91%	101	1123	91.75%	76	1121	93.65%
	0.00%	100.00%		65.99%	99.82%		87.21%	99.64%	
	1000%			1500%					
	Accuracy:	96.43%		Accuracy:	97.05%				
	1019	9	99.12%	1519	15	99.02%			
	70	1116	94.10%	65	1110	94.47%			
	93.57%	99.20%		95.90%	98.67%				

Figure 100: Results of x-val ANOVA kernel and different SMOTE oversampling percentages

In Figure 100 the different results of 10-fold cross-validation with AONVA kernel (no parameter optimization conducted) and different SMOTE percentages are compared. The results indicate that the classification performance improves from 200% SMOTE to 500% SMOTE but then the improvement slows down for 1000% SMOTE and 1500% SMOTE.

In a next step, the kernel parameters are optimized, finding a better setting to improve the classification performance in case of 200% SMOTE oversampling.

accuracy: 94.45% +/- 2.16% (mikro: 94.44%)

	true negative	true positive	class precision
pred. negative	229	11	95.42%
pred. positive	68	1114	94.25%
class recall	77.10%	99.02%	

Figure 101: Results x-val ANOVA kernel and 200% SMOTE oversampling (parameters: kernel gamma: 2.0; kernel degree: 3.0; C: 2.6; convergence epsilon: 0.001)

The classification performance of the cross-validation using SVM algorithm with adjusted parameters shows that the results improve and come close for the SMOTE oversampled data set (200%) which was previously not acceptable. In general, SMOTE oversampling is applicable for larger percentages (e.g., 300% or 500%) as well (Chawla et al., 2002). However, the author decided to only use a smaller percentage, 200% in this case, which allows for good classification results (see Figure 101) when parameter optimization is utilized.

SECOM 528 SMOTE 200%					
Accuracy:	91.85%		Accuracy:	90.22%	
2	28	6.67%	0	30	0.00%
2	336	99.41%	6	332	98.22%
50.00%	92.31%		0.00%	91.71%	
Parameters: (Anova, KG 2, KD 3, C 2.6)			Parameters: (Poly., KD 3, C 2.6, C-eps. 0.001)		

SECOM 528 SMOTE 500%								
Accuracy:	91.03%		Accuracy:	90.76%		Accuracy:	92.12%	
2	28	6.67%	1	29	3.33%	1	29	3.33%
5	333	98.52%	5	333	98.52%	0	338	100.00%
28.57%	92.24%		16.67%	91.99%		100.00%	92.10%	
Parameters: (Anova, KG 3, KD 3, C 2.6)			Parameters: (Poly., KD 3, C 2.6, C-eps. 0.001)			Parameters: (Poly., KD 2, C 2, C-eps. 0.001)		

Figure 102: Classification results of test set applying 200% & 500% SMOTE on learning set

In order to evaluate the applicability of the classification model in a manufacturing environment, e.g., in monitoring (hypothesis 3), the classification performance, model development and subsequent evaluation through learning and test set is conducted with this kernel and parameters. In a first step, the original data set (pre-oversampling) is divided in learning (70%) and test (30%) set, upholding the same

ratio of the two classes. This is done using the random function on two files, each containing one class in RapidMiner (v5.3). The separated class files are then merged to learning and test set. Then the learning set is subject to SMOTE oversampling (200% and 500%) to create the classification model. Finally, the classification performance of this model is tested by applying it on the test set and evaluating the classification results.

It can be seen that even so the cross-validation performance is very good after the application of SMOTE oversampling, the low classification performance on previously unknown data persists. This may again be based on the overfitting problem. As the SMOTE approach is designed to avoid overfitting, the alternative suspicion is that the high diversity within the example population of the minority class is responsible for the poor classification results becomes more likely. The showing classification performance results are not acceptable (see Figure 102) even so slightly better than with random oversampling (see Figure 98). The reason for overfitting may be found in the fact that the overfitting problem is caused by two (independent) characteristics of the data set or the data structure does not allow for the classification of novel examples with a trained model. First, the unbalanced data set concerning the two classes which is enhanced by random oversampling. A possible solution for this is the SMOTE oversampling method. However, there is the second cause of overfitting that may be found in the ratio of vectors (examples) and attributes (features) in the data set (Yu & Liu, 2004). Therefore, in the next sub-section, this challenge is targeted.

1.1.1 Feature ranking using SVM classifier

It was argued in previous sections during the original pre-processing stage that the goal is to keep as many features as possible. The reason is, that in theory, though this should provide more discriminant power (Yu & Liu, 2004) and thus allow for detection of more process intra- and inter-relations between state characteristics. However, in practice, when only a limited amount of training examples (vectors) is available combined with a large amount of features (attributes), overfitting is common when there are irrelevant features in the learning set (Guyon et al., 2002; Yu & Liu, 2004). Even so SVM algorithms are relatively robust towards overfitting, they profit from a feature reduction (Guyon et al., 2002).

At this point, the feature (attribute) size needs to be reduced in order to achieve a better ratio between examples and attributes/features in the data set. At this stage it can be assumed that the important features are selected in accordance with the objective of feature selection as presented by (e.g., Guyon et al., 2002; Yu & Liu, 2004; Chang & Lin, 2008). Feature selection was found to have a positive impact on most learning algorithms (Waikowski & Chen, 2010).

However, this might be problematic in cases where process steps, integrated in an overall manufacturing programme, are concerned. In such cases, the feature selec-

tion might deem features "not important" which may be 'important' (state drivers) within a larger context, e.g., a {process1; process2} combination as shown in scenario I & II. In such a case, it may be possible to apply feature selection on the individual process but then the process is repeated from scratch for the combined process vector (incl. all original features) to identify further 'important features' (state drivers) for the overall manufacturing programme.

In a first step towards eliminating unimportant features for the SECOM process, all the features with all (99.0% - adjustable) similar values or largely varying values are removed using the WEKA function 'RemoveUseless' (Unsupervised; Attribute) (Bhuvaneswari & Dhulipala, 2013). This is due to the assumption that those features do not support classification in the majority of cases. The application of this function removes 116 features (of 528) leaving 412 features for further processing. In order to evaluate the assumption that those features are indeed 'useless' for classification purposes, eliminated features are compared to the later conducted feature ranking by SVM. During the later feature ranking by SVM evaluation, the 116 eliminated features are all deemed the least important by the classifier (see Table 21 & Table 22 in the Annex) which supports the assumption made before.

	No SMOTE			200%			500%			1000%			1500%		
528 (original)	Accuracy:	91.91%		Accuracy:	92.76%		Accuracy:	95.35%		Accuracy:	96.43%		Accuracy:	97.05%	
	0	0	#DIV/0!	196	2	98.99%	518	4	99.23%	1019	9	99.12%	1519	15	99.02%
	99	1125	91.91%	101	1123	91.75%	76	1121	93.65%	70	1116	94.10%	65	1110	94.47%
	0.00%	100.00%		65.99%	99.82%		87.21%	99.64%		93.57%	99.20%		95.90%	98.67%	
412 (no SVM_FS)	Accuracy:	91.91%		Accuracy:	90.15%		Accuracy:	95.52%		Accuracy:	96.66%		Accuracy:	96.83%	
	0	0	#DIV/0!	158	1	99.37%	519	2	99.62%	1025	10	99.03%	1518	20	98.70%
	99	1125	91.91%	139	1124	88.99%	75	1123	93.74%	64	1115	94.57%	66	1105	94.36%
	0.00%	100.00%		53.20%	99.91%		87.37%	99.82%		94.12%	99.11%		95.83%	98.22%	
50	Accuracy:	91.91%		Accuracy:	85.94%		Accuracy:	88.71%		Accuracy:	91.01%		Accuracy:	92.28%	
	0	0	#DIV/0!	119	22	84.40%	488	88	84.72%	1030	140	88.03%	1550	175	89.86%
	99	1125	91.91%	178	1103	86.10%	106	1037	90.73%	59	985	94.35%	34	950	96.54%
	0.00%	100.00%		40.07%	98.04%		82.15%	92.18%		94.58%	87.56%		97.85%	84.44%	
80	Accuracy:	91.91%		Accuracy:	90.37%		Accuracy:	92.32%		Accuracy:	94.49%		Accuracy:	95.04%	
	0	0	#DIV/0!	179	19	90.40%	519	57	90.10%	1040	73	93.44%	1551	77	95.27%
	99	1125	91.91%	118	1106	90.36%	75	1068	93.44%	49	1052	95.55%	33	1048	96.95%
	0.00%	100.00%		60.27%	98.31%		87.37%	94.93%		95.50%	93.51%		97.92%	93.16%	
100	Accuracy:	91.91%		Accuracy:	91.49%		Accuracy:	92.67%		Accuracy:	94.94%		Accuracy:	96.01%	
	0	0	#DIV/0!	190	14	93.14%	514	46	91.79%	1039	62	94.37%	1548	72	95.56%
	99	1125	91.91%	107	1111	91.22%	80	1079	93.10%	50	1063	95.51%	36	1053	96.69%
	0.00%	100.00%		63.97%	98.76%		86.53%	95.91%		95.41%	94.49%		97.73%	93.60%	
120	Accuracy:	91.91%		Accuracy:	91.56%		Accuracy:	93.54%		Accuracy:	96.07%		Accuracy:	96.86%	
	0	0	#DIV/0!	190	13	93.60%	520	37	93.36%	1044	42	96.13%	1542	43	97.29%
	99	1125	91.91%	107	1112	91.22%	74	1088	93.63%	45	1083	96.01%	42	1082	96.26%
	0.00%	100.00%		63.97%	98.84%		87.54%	96.71%		95.87%	96.27%		97.35%	96.18%	

Figure 103: Matrix comparing x-val performance (ANOVA – orig. parameters) of SECOM data in different variations of SMOTE and feature selection parameters

In the following step, the feature selection function based on Guyon et al. (2002) incorporated in WEKA is applied in order to reduce the number of features. The elimination can either be triggered by a fixed threshold given the number of "to-be-maintained" features or a threshold concerning the minimum weight of the weight vector $\boldsymbol{w}$ of the sustaining features. As previously described, the weight vector $\boldsymbol{w}$ represents the major ranking criterion for the SVM feature selection method (Guyon et al., 2002). This was previously described in more detail in section 6.3.4.

In a first analysis, the SECOM data set undergoes feature selection with a hard threshold of 50; 80; 100 and 120 features. Each of those resulting data sets is then supplemented by oversampling using SMOTE 200%, 500%, 1000% and 1500%. Figure 103 summarizes the results of the confusion matrix. In this case the cross-validation is run with a SVM algorithm with an ANOVA kernel (orig. parameters).

The results show that for cases with a lower percentage of SMOTE oversampling (200%) the feature selection seems to improve the overall classification results. For the test runs with higher SMOTE oversampling, even the data set with a feature set reduced to 50 features produces classification examples above the threshold of 80% for class recall and class prediction. One finding of this analysis is that the feature selection allows the reduction of the total amount of features to a set of relevant features, representing state drivers, responsible directly and indirectly for the quality outcome. This confirms the assumption that a set of relevant state characteristics may be selected and used to describe the state comprehensively as it is understood in the *product state concept*.

Optimizing Parameters (FS_50 SMOTE_200)

Accuracy:	94.16%	
241	27	89.93%
56	1098	95.15%
81.14%	97.60%	

Parameters:	
- Kernel	Anova
- Kernel gamma	2.0
- Kernel degree	3.0
- C	2.6

Optimizing Parameters (FS_50 SMOTE_200)

Accuracy:	90.08%	
253	97	72.29%
44	1028	95.90%
85.19%	91.38%	

Parameters:	
- Kernel	Poly.
- Kernel degree	3.0
- C	2.0
- convergence eps	0.001

Optimizing Parameters (FS_80 SMOTE_200)

Accuracy:	93.95%	
234	23	91.05%
63	1102	94.59%
78.79%	97.96%	

Parameters:	
- Kernel	Anova
- Kernel gamma	3.0
- Kernel degree	3.0
- C	2.0
- convergence eps	0.001

Optimizing Parameters (FS_80 SMOTE_200)

Accuracy:	94.30%	
260	44	85.53%
37	1081	96.69%
87.54%	96.09%	

Parameters:	
- Kernel	Poly.
- Kernel degree	3.0
- C	2.0
- convergence eps	0.005

Figure 104: Results x-val after parameter optimization on SECOM after feature selection (50 & 80) and SMOTE oversampling (200%)

One has to bear in mind, that in this compression, no SVM classification parameter optimization was conducted. By adjusting the SVM parameters to the data structure, the results may improve further. With an optimized parameter set, the SECOM data set with 50 and 80 features and 200% SMOTE oversampling meets the target threshold of 80% (see Figure 104). For the data set resembling 50 remaining relevant features, an ANOVA kernel achieves a relatively evenly distributed result over the target threshold of 80%. However, the 80 feature/attribute version achieves better classification performance results with the Polynomial kernel.

The final results with the optimized parameters show a confusion matrix with values significantly higher than the target threshold of 80% for both versions (50 and 80 features/attributes). These results confirm that applying feature selection allows the identification of relevant state drivers by which the quality of a product may be measured even for a challenging data set like the SECOM process.

To evaluate the impact the feature selection has on the classification performance of a previously unknown data set, the previously used randomly selected learning (70%) and test (30%) set (see Figure 102) is reduced to 50 features and 80 features. The obsolete features were deleted manually from the previous data set to ensure the same examples are in the learning and test data set. This increases the comparability of the results. Afterwards, the model is generated by applying SOMTE (200% & 500%) to the learning set with an ANOVA kernel (kernel gamma 2.0; kernel degree 3.0; C 2.6) before the test set is evaluated.

A second analysis concerning the classification results on previous unknown data is conducted. In this case, the learning and test set is selected in timely succession of the process. The first 70% of the examples are used as the learning set and the latter 30% as the test set. This resembles an industrial application scenario as the latter 30% resemble new examples which are to be classified based on historic data. However, in this case, the 'fail' ratio differs minimally between the test and the learning set. The pre-SMOTE learning set has an already low ratio of 8.75% whereas the test set has only 6.54%.

	No SMOTE			200%			500%		
528 (no FS)	Accuracy:	91.85%		Accuracy:	91.85%		Accuracy:	91.03%	
	1	29	3.33%	2	28	6.67%	2	28	6.67%
	1	337	99.70%	2	336	99.41%	5	333	98.52%
	50.00%	92.08%		50.00%	92.31%		28.57%	92.24%	
FS50	Accuracy:	91.85%		Accuracy:	91.85%		Accuracy:	90.49%	
	2	28	6.67%	4	26	13.33%	3	27	10.00%
	2	336	99.41%	4	334	98.82%	8	330	97.63%
	50.00%	92.31%		50.00%	92.78%		27.27%	92.44%	
FS80	Accuracy:	92.12%		Accuracy:	92.12%		Accuracy:	92.12%	
	2	28	6.67%	3	27	10.00%	5	25	16.67%
	1	337	99.70%	2	336	99.41%	4	334	98.82%
	66.67%	92.33%		60.00%	92.56%		55.56%	93.04%	

Figure 105: Comparison matrix of classification results of test set (randomly selected) after feature selection (50 & 80) and SMOTE (200% & 500%) application on learning set

The results of the cross-validation of the SVM classifier using an ANOVA kernel (kernel gamma 2.0; kernel degree 3.0; C 2.6) similar to the previous test are not very promising (Figure 106). The recognition rate for 'fail' examples is even worse than for the randomly divided data set (Figure 105).

As can be observed in Figure 105 and Figure 106, the classification results are not satisfactory for previously unknown data even with sophisticated pre-processing measures like feature selection and SMOTE oversampling. This confirms the suspicion that the poor classification results are most likely based on the high diversity within the example population of the minority class. In the results section (section 7.1.5) this is picked up in more detail and additional evaluation results are provided that confirm the suspicion further.

	no SMOTE			200%			500%		
528 (no FS)	Accuracy:	93.46%		Accuracy:	93.46%		Accuracy:	93.46%	
	0	24	0.00%	0	24	0.00%	0	24	0.00%
	0	343	100.00%	0	343	100.00%	0	343	100.00%
		93.46%			93.46%			93.46%	
FS50	Accuracy:	93.19%		Accuracy:	93.19%		Accuracy:	93.19%	
	0	24	0.00%	0	24	0.00%	1	23	4.17%
	1	342	99.71%	1	342	99.71%	2	341	99.42%
	0.00%	93.44%		0.00%	93.44%		33.33%	93.68%	
FS80	Accuracy:	93.46%		Accuracy:	93.46%		Accuracy:	92.64%	
	0	24	0.00%	0	24	0.00%	1	23	4.17%
	0	343	100.00%	0	343	100.00%	4	339	98.83%
		93.46%			93.46%		20.00%	93.65%	

Figure 106: Comparison matrix of classification results of test set (in timely succession) after feature selection (50 & 80) and SMOTE (200% & 500%) application on learning set

However, the results do show that the classification results after feature selection, especially with 80 features remaining, are slightly better than without. As the difference is marginal, the question remains what conclusion can be drawn from that. This is also discussed in greater detail in the following section.

In the next section, the evaluation results are presented in a structured way and critically discussed. Furthermore, the limitations of the approach and the evaluation are illustrated.

7 Evaluation of the developed approach

The evaluation results derived from the previous application section are presented in a condensed fashion and critically discussed within this section. The critical discussion is roughly structured along the previously presented research hypotheses. Following, the limitations identified during the evaluation and analysis including data pre-processing are highlighted. Within that section the implications of those limitations on the hypotheses and the research results are illustrated.

7.1 Evaluation results

In this section the results of the application and evaluation conducted in the previous section are presented in a condensed fashion. This provides a basis for the following critical discussion in the next sub-section. The results compromise not only results with an impact on the raised research question, but also additional findings that surfaced during the evaluation.

7.1.1 Data pre-processing

All three scenarios include pre-processing the data sets which is detailed in section 9.2 in the Annex. Even so the pre-processing itself is not considered a main part of the dissertation, the subsequent limitations are regarded as highly relevant and thus presented within the main body of the thesis. Scenario I represents a special case in this context as the data was provided by Rolls-Royce in anonymized form and thus already pre-processed to a certain extent (no missing values and normalized). However, the other scenarios illustrate the challenges data pre-processing represents in manufacturing. The CHEM data set, contained relatively few missing values and with only 176 examples and 57 attributes/features can be considered small. This is reflected in the effort needed for preprocessing and handling of missing values. Also the computing requirements are lower for this scenario than for the larger data sets of scenario I & III.

Scenario III provides an example of a very challenging data set when it comes to pre-processing. This is supported by the fact that it was published as part of the 'Causality Challenge' (McCann et al., 2008). The data set with its originally 1567 examples and 591 attributes/features contains a large amount of missing data. Not only is the ratio of missing data high, but the missing values are also distributed over almost all examples and attribute/features, making the handling challenging. The approach on how to handle the data was found to have an impact on the later behavior of the data during evaluation by comparing two of the variants presented in section 9.2.3.

As stated previously, the different data pre-processing approaches regarding the elimination of missing values differentiate themselves in the number of features and examples. 'Approach 1', deleting all features with missing data (see Figure

113) compared to 'Approach 2 Variant 2 (plus 15)', keeping all features with less than 10 missing values show that the ranking of the relevant features (WEKA AttributeEvaluation) show that of the 485 features of the data set pre-processed according to 'Approach 1', 43 are not ranked within the top 485 features of the most relevant features of the data set according to 'Approach 2 Variant 2 (plus 15)'. Comparing the classification performance of the two (see Figure 107), the results are comparable showing a more even distribution for the data set containing more features (528) prior to feature ranking/elimination.

	200%			500%			200%		
Approach 1 FS 50	Accuracy:	87.38%		Accuracy:	90.35%		Accuracy:	94.61%	
	84	4	95.45%	401	46	89.71%	197	15	92.92%
	174	1149	86.85%	115	1107	90.59%	61	1138	94.91%
	32.56%	99.65%		77.71%	96.01%		76.36%	98.70%	
Approach 2 var 2 plus 50 FS 50	Accuracy:	85.94%		Accuracy:	88.71%		Accuracy:	94.16%	
	119	22	84.40%	488	88	84.72%	241	27	89.93%
	178	1103	86.10%	106	1037	90.73%	56	1098	95.15%
	40.07%	98.04%		82.15%	92.18%		81.14%	97.60%	
	Parameters: Orig.			Parameters: Orig.			Parameters: (Anova, KG 2, KD 3, C 2.6)		
Approach 1 FS 80	Accuracy:	89.72%		Accuracy:	93.89%		Accuracy:	95.18%	
	123	10	92.48%	445	31	93.49%	221	31	87.70%
	135	1143	89.44%	71	1122	94.05%	37	1122	96.81%
	47.67%	99.13%		86.24%	97.31%		85.66%	97.31%	
Approach 2 var 2 plus 50 FS 80	Accuracy:	90.37%		Accuracy:	92.32%		Accuracy:	94.30%	
	179	19	90.40%	519	57	90.10%	260	44	85.53%
	118	1106	90.36%	75	1068	93.44%	37	1081	96.69%
	60.27%	98.31%		87.37%	94.93%		87.54%	96.09%	
	Parameters: Orig.			Parameters: Orig.			Parameters: (Poly., KD 3, C 2, co.eps 0.005)		

Figure 107: Comparison of pre-processing approach 1 and approach 2 var. 2 (plus 15) by classification performance after feature selection and SMOTE application

Looking at the comparison of the two variations of approach 2, the following features are eliminated in variant 1 during pre-processing as they contained more than 5 missing values: 20; 85; 156; 220; 291; 358; 429; 492. Looking in the ranking position of those features in the feature ranking during the evaluation of 'Approach 2 variant 1 (plus 15)', it shows the following results: Feature 429 rank 14 (528); Feature 156 rank 59 (528); Feature 492 rank 61 (528); Feature 20 rank 114 (528); Feature 291 rank 127 (528); Feature 85 rank 149 (528); Feature 358 rank 282 (528); Feature 220 rank 358 (528). The relatively high ranking position of feature no. 429 indicates that by deleting features during pre-processing, valuable information (state drivers) might get lost and with them potential knowledge/information about the process and product state development.

These results indicate the existing influence that data pre-processing has on the later application of supervised ML algorithms and its results. In this case, there is no judgment made with regard of one approach being better. The main reason for pre-

senting this result is to highlight the importance of data pre-processing and the possible influence on the results.

7.1.2 Cross-validation performance of SVM classifier

Overall it can be said that all scenarios showed acceptable to good performance in the cross-validation test, partly after significant optimization efforts.

The results of the evaluation of the classification performance through cross-validation for the Rolls-Royce data set (scenario I) show very good results. This stands true for application of a linear kernel (important for the later feature selection approach) and even more so for the optimized algorithm sporting an ANOVA kernel. In this set up, the results are extremely good (see Figure 62). Even so this is promising and shows that the data set's structure/nature allows for classification, this might present a previously induced bias. The data set was provided in anonymized form and by doing so the providing party, Rolls-Royce, applied SMOTE to a) alter the data set so no information can be extracted by competitors and b) to make it more balanced. However, as it is not known by how many percent the minority class was extended, the classification results are to be interpreted with care. When comparing the difference a SMOTE application can make on a previously not ideal performing data set (from a classification perspective) like the SECOM one (see Figure 100), it has to be assumed that the SMOTE application has an influence on the good results of the RR data set as well. Nevertheless, the performance results still confirm that it is possible to identify the quality outcome of the process with a good accuracy using an SVM classifier algorithm.

In the second scenario, a chemical manufacturing process was analyzed. Originally this process was published as a regression data set. However, by selecting a threshold (Yield 39), the data set was transformed in a data set with two classes ('pass' & 'fail'). The classification performance of the original data set was below the target threshold in the cross-validation evaluation, even with optimized parameters. However, by applying the SMOTE method and subsequent parameter optimization, the classification results following were very good and significantly higher than the target threshold. As the percentage of SMOTE enhanced minority class was rather low with 100%, the results are good. The high classification performance results confirm that the CHEM data set, now used for classification, is applicable for this evaluation and the SVM algorithm is able to distinguish the product quality of the process with high accuracy (see Figure 79).

The originally unsatisfactory classification performance results of the SECOM data in the cross-validation test (see Figure 93) could be improved significantly by different measures targeting the identified problematic areas of the data set. The reasons for the poor classification performance where identified to be based on the unbalanced data set. Overall, the minority class ('fail') was underrepresented and additionally, the large feature set (528 features) with only 1224 examples was

problematic. This was approached by under-/oversampling and features selection. The results of these applications are presented in the following sub-section.

As for scenario I and II, which also included synthetic and combined vectors in the evaluation, the classification performance of those is briefly discussed here. For the RR manufacturing programme, the classes were assigned based on a previous conducted cluster analysis. This is different from the more random approach used for the synthetic processes of the CHEM manufacturing programme. Different approaches have been chosen to reduce the possible bias possibly induced by either one. For the RR data set, the original process TOM(RR) shows acceptable classification results whereas the synthetic and combined vectors show very good classification results in cross-validation using a linear kernel (see Figure 63). This was expected as the synthetic processes and combined vectors are designed in such a way with weak inherent clustering and standard deviation. This allows for a good application of the following feature ranking method in the data sets. For the CHEM manufacturing programme the synthetic and combined vectors were evaluated with regard to their classification performance using a linear kernel and the original parameters as well. Similar to the RR manufacturing programme, the reason is to see whether the feature ranking method is applicable to the vectors. With an exception of the TD(CHEM) vector, all other processes and TDH(CHEM) show very good classification results. This stands especially true given that a basic linear kernel has been used. The TD(CHEM) vector stands out as its performance is below the threshold. However, it was decided to not adjust it (e.g., repeat the random selection of class) to reproduce realistic circumstances as much as possible.

Overall, all scenarios and the respective 'real world' and synthetic data sets show at least acceptable classification performance with SVM algorithm classifiers. After some adjustments have been made to improve the performance, mainly targeting the unbalanced nature of two of the three cases, and additional parameter optimization, the majority shows very good classification performance results.

7.1.3 Unbalanced data

Looking at the issue of unbalanced data sets, it shows that when working with 'real world' manufacturing data sets this issue often surfaces. In scenario II & III the unbalanced nature of the data set had to be tackled by appropriate measures. The chosen methods, SMOTE oversampling and feature selection have shown good results (see Figure 86 & Figure 103).

For scenario I however, SMOTE oversampling was applied previously to the provision of the data set from Rolls-Royce with regard to the anonymity issues. Therefore, the received data set does not feature unbalanced ratio of 'fails' and 'pass' examples. However, as SMOTE was applied previously, it suggests that the raw data set prior to anonymizing actions also faced unbalanced ratio of the minority and majority class.

It has to be noted that unbalanced data with a smaller minority class (hence the name) is actually desired in manufacturing even so it makes life harder for model generation. Ideally the ratio is as small as possible as this means the manufacturing programme has very little quality issues (small 'fail' rate). However, this highlights the need to develop appropriate methods to select representative examples for model generation which counter the unbalanced data bias.

7.1.4 Feature selection and feature ranking

In this section the results of the applied feature selection based on the feature ranking following Guyon et al. (2002) of the three scenarios are presented. It has been shown that the classification performance, thus performance of correct judgment of quality in this case, is equally good or even better in some cases when feature ranking and selection is applied.

Looking at the results of the Rolls-Royce process (scenario I) it was found that for the original TOM(RR) process the feature selection has a rather small effect on the classification performance (cross-validation). In this case the feature ranking was conducted by two different programs. One feature ranking of the TOM(RR) process was conducted using the RapidMiner (v5.3) function 'Weight by SVM" which provides the actual weight vector $\boldsymbol{w}$ values (normalized) as an output. The classification performance of the reduced data set can be considered very good for all tested variations (FS10; FS15; FS20; FS30; FS50 & FS57). However, the results seem to show that the more features used, the better the results and the closer to the results of the full feature set. The best results were achieved by the full features data set without feature selection. On the other hand, the classification performance of the reduced feature set, even the smallest one with 10 features, shows good classification performance above the target threshold (see Figure 65).

The second feature ranking variant was conducted with the WEKA function 'SVMAttributeEval', which is designed based on the feature ranking method developed by Guyon et al. (2002). The resulting feature ranking is different from the one obtained by the RapidMiner (v5.3) SVM weight function. The performance of the different variations with different amount of features by cross-validation show overall a better performance of the WEKA based feature selection (see Figure 67).

Looking at a comparison of two variants, one with the top ranked 20 features and the other with the 20 lowest ranked features, the classification performance differs for each programme (RapidMiner & WEKA) (see Figure 68). Whereas the RapidMiner (v5.3) variant surprises with a better classification performance for the lowest ranked features, the WEKA variant shows the expected result of significantly better classification performance for the version with the 20 highest ranking features. Based on these results, the overall better classification performance of variants with feature selection and the better documentation of the WEKA function, the ranking of the RR manufacturing programme is conducted based on the WEKA

feature ranking. In the following, scenarios II & III, the WEKA function is also employed as the function of choice.

Looking at the individual and combined vectors of the manufacturing programme, the results are interesting. Table 6 illustrates the change in rank of certain features along the manufacturing programme in an excerpt of the full feature ranking set up. The full ranking containing all features is provided in the Annex (see Annex Table 19). The RR feature ranking evaluation results are similar to the ones obtained by the analysis of the CHEM scenario which is analyzed after.

Table 6: Feature Ranking RR manufacturing programme (selected)

Rank	TOM(RR)	Rank	DICK(RR)	Rank	HARRY(RR)	Rank	TD(RR)	Rank	TDH(RR)
1	para.51	1	para.DICK.29	1	para.HARRY.43	1	para.DICK.43	1	para.HARRY.67
2	para.21	2	para.DICK.7	2	para.HARRY.33	2	para.DICK.32	2	para.DICK.32
3	para.50	3	para.DICK.41	3	para.HARRY.24	3	para.DICK.47	3	para.HARRY.11
4	para.33	4	para.DICK.12	4	para.HARRY.32	4	para.DICK.42	4	para.DICK.43
5	para.6	5	para.DICK.27	5	para.HARRY.47	5	para.DICK.33	5	para.HARRY.2
6	para.36	6	para.DICK.21	6	para.HARRY.42	6	para.DICK.24	6	para.HARRY.52
7	para.14	7	para.DICK.48	7	para.HARRY.58	7	para.DICK.11	7	para.DICK.33
8	para.9	8	para.DICK.31	8	para.HARRY.30	8	para.DICK.2	8	para.HARRY.36
9	para.47	9	para.DICK.9	9	para.HARRY.67	9	para.DICK.52	9	para.DICK.24
10	para.59	10	para.DICK.56	10	para.HARRY.11	10	para.DICK.36	10	para.HARRY.1
11	para.29	11	para.DICK.17	11	para.HARRY.2	11	para.DICK.1	11	para.77
12	para.55	12	para.DICK.23	12	para.HARRY.52	12	para.26	12	para.DICK.26
13	para.61	13	para.DICK.37	13	para.HARRY.36	13	para.DICK.45	13	para.HARRY.45
14	para.60	14	para.DICK.51	14	para.HARRY.64	14	para.DICK.55	14	para.HARRY.55
15	para.44	15	para.DICK.22	15	para.HARRY.65	15	para.DICK.30	15	para.DICK.30
16	para.45	16	para.DICK.28	16	para.HARRY.39	16	para.DICK.13	16	para.HARRY.60
17	para.34	17	para.DICK.34	17	para.HARRY.45	17	para.DICK.28	17	para.DICK.39
18	para.32	18	para.DICK.13	18	para.HARRY.59	18	para.DICK.37	18	para.HARRY.13
19	para.64	19	para.DICK.5	19	para.HARRY.55	19	para.62	19	para.DICK.44
20	para.2	20	para.DICK.6	20	para.HARRY.60	20	para.6	20	para.HARRY.42
21	para.5	21	para.DICK.55	21	para.HARRY.13	21	para.DICK.23	21	para.DICK.47
22	para.35	22	para.DICK.45	22	para.HARRY.19	22	para.DICK.17	22	para.HARRY.47
23	para.46	23	para.DICK.1	23	para.HARRY.28	23	para.70	23	para.HARRY.32
24	para.48	24	para.DICK.36	24	para.HARRY.50	24	para.DICK.56	24	para.DICK.19
25	para.31	25	para.DICK.52	25	para.HARRY.37	25	para.DICK.9	25	para.HARRY.28
26	para.38	26	para.DICK.2	26	para.HARRY.23	26	para.DICK.39	26	para.DICK.50
27	para.11	27	para.DICK.11	27	para.HARRY.17	27	para.DICK.31	27	para.HARRY.37
28	para.42	28	para.DICK.42	28	para.HARRY.56	28	para.45	28	para.DICK.46
29	para.20	29	para.DICK.47	29	para.HARRY.9	29	para.DICK.19	29	para.24
30	para.16	30	para.DICK.32	30	para.HARRY.15	30	para.DICK.50	30	para.HARRY.43
...		...		...		...		...	
46	para.24	31	para.DICK.43	54	para.HARRY.1	32	para.DICK.29	120	para.6
53	para.26	32	para.DICK.33			85	para.77	136	para.DICK.29
70	para.77					103	para.51	158	para.51
						138	para.24	174	para.26

The interesting development is that certain parameters (features) which are ranked rather high within the individual processes (TOM(RR), DICK(RR) & HARRY(RR)) feature rankings, are often not ranked as that relevant when it comes to the combined vectors TD(RR) and TDH(RR) and vice versa. For example, 'pa-

ra.51', the highest ranked feature of process TOM(RR) (No.1) is ranked number 103 in TD(RR) and even number 158 in TDH(RR). On the other hand, 'para.77', ranked rather low with number 77 in TOM(RR) is ranked number 11 in TDH(RR) but number 85 in TD(RR). Interestingly, for DICK(RR), three parameters which are ranked closely together in the individual ranking (no. 30; no. 31; no. 32) are all ranked within the top ten highest ranked features for both TD(RR) and TDH(RR). However, the top ranked feature 'para.DICK.29' during the individual process DICK(RR) is ranked number 32 in TD(RR) and number 136 in TDH(RR). For process HARRY(RR), two of the top ten ranked features ('para.HARRY.67'; 'para.HARRY.11') are also ranked within the top ten of TDH(RR). One of them, 'para.HARRY.67' ranked the most important feature of TDH(RR). The top ranked feature of the individual process HARRY(RR), 'para.HARRY.43' is ranked lower at position 30 in TDH(RR). A feature, rather lowly ranked with number 54 in the individual process HARRY(RR), 'para.HARRY.1' is ranked the high number 10 in the combined vector TDH(RR).

The feature rankings of the CHEM manufacturing programme are obtained using the SVM evaluation (feature weights). This is applied according to Guyon et al. (2002) by using the WEKA function 'SVMAttributeEval'. The features of the (individual/combined) vector are ranked by the square of the weight assigned by the SVM classifier. The following Table 7 shows an excerpt of the resulting ranking for the different processes, combined processes and the complete manufacturing programme (to analyze cross-process intra-relations). In this table, the 30 features ranked most important are displayed, expanded by selected additional features chosen to illustrate the changing importance over the process. A full ranking containing all features is provided in the Annex (see Annex Table 20).

In Table 7, it becomes apparent that in the combined state vectors TD(CHEM) and TDH(CHEM), different and/or additional state drivers (features) become relevant compared to those identified as relevant by feature ranking in the individual processes. Next, selected examples are discussed to analyze the findings further.

Looking at feature '*BiologicalMaterial01*' of process TOM(CHEM), ranked as most important in the individual process, it can be observed that the importance decreases to rank no. 17 in the combined TD(CHEM) vector. In the ranking of the features for the complete manufacturing programme TDH(CHEM) the feature 'BiologicalMaterial01' is the least important feature of all, ranked as no. 137. This indicates that features which are highly relevant state drivers for individual processes, may have little influence when the whole multi-stage manufacturing programme is concerned.

On the other hand, looking at feature '*ManufacturingProcess42*' (rank no. 36 in TOM(CHEM)), the influence of features in individual processes may be insignificant. However, the same feature is the ranked no. 6 (TD(CHEM)) and no. 8 of the

most influential features for the whole manufacturing programme TDH(CHEM), resembling the highest rank of all features from process TOM(CHEM) in the ranking, surpassing all 35 features ranked higher individually.

Similar examples include '*ManufacturingProcess16*', being ranked no. 51 (TOM(CHEM)) and no. 1 (TD(CHEM)) and '*BiologicalMaterial06*' ranked no. 45 (TOM(CHEM)) and no. 5 (TD(CHEM)). However, both features rank significantly lower when analyzing the complete manufacturing programme TDH(CHEM) with rank no. 88 ('ManufacturingProcess16') and no. 31 ('BiologicalMaterial06'). This indicates that the increase of importance of a feature within a manufacturing programme is not necessarily increasing towards the final state but can have its peak at different checkpoints throughout the manufacturing programme.

Table 7: Feature Ranking CHEM manufacturing programme (selected)

Rank	TOM(CHEM)	Rank	DICK(CHEM)	Rank	HARRY(CHEM)	Rank	TD(CHEM)	Rank	TDH(CHEM)
1	BiologicalMaterial07	1	Parameter 29	1	Parameter29	1	ManufacturingProcess16	1	ParameterII29
2	ManufacturingProcess32	2	Parameter 38	2	Parameter11	2	ManufacturingProcess05	2	ParameterII11
3	ManufacturingProcess09	3	Parameter 11	3	Parameter13	3	ManufacturingProcess06	3	ParameterII13
4	ManufacturingProcess34	4	Parameter 13	4	Parameter32	4	ManufacturingProcess33	4	ParameterII32
5	ManufacturingProcess13	5	Parameter 32	5	Parameter02	5	BiologicalMaterial06	5	ParameterII02
6	ManufacturingProcess19	6	Parameter 02	6	Parameter09	6	ManufacturingProcess42	6	ParameterII09
7	ManufacturingProcess30	7	Parameter 09	7	Parameter28	7	ManufacturingProcess13	7	ParameterII28
8	ManufacturingProcess39	8	Parameter 45	8	Parameter17	8	BiologicalMaterial12	8	ManufacturingProcess42
9	ManufacturingProcess21	9	Parameter 43	9	Parameter31	9	ManufacturingProcess14	9	ParameterII17
10	ManufacturingProcess01	10	Parameter 28	10	Parameter25	10	ManufacturingProcess22	10	ParameterII31
11	BiologicalMaterial05	11	Parameter 34	11	Parameter07	11	ParameterI 38	11	ParameterII25
12	BiologicalMaterial02	12	Parameter 39	12	Parameter26	12	BiologicalMaterial01	12	ParameterII07
13	ManufacturingProcess10	13	Parameter 47	13	Parameter12	13	ManufacturingProcess38	13	ParameterII26
14	BiologicalMaterial09	14	Parameter 17	14	Parameter10	14	BiologicalMaterial08	14	ParameterII12
15	ManufacturingProcess03	15	Parameter 31	15	Parameter03	15	ManufacturingProcess45	15	ParameterII10
16	ManufacturingProcess31	16	Parameter 42	16	Parameter04	16	ManufacturingProcess11	16	ManufacturingProcess22
17	ManufacturingProcess37	17	Parameter 25	17	Parameter22	17	BiologicalMaterial07	17	ParameterII03
18	ManufacturingProcess12	18	Parameter 07	18	Parameter15	18	ManufacturingProcess36	18	ParameterII04
19	BiologicalMaterial10	19	Parameter 37	19	Parameter01	19	ManufacturingProcess43	19	ParameterII22
20	ManufacturingProcess23	20	Parameter 26	20	Parameter06	20	ManufacturingProcess15	20	ParameterII15
21	ManufacturingProcess02	21	Parameter 12	21	Parameter23	21	ManufacturingProcess08	21	ParameterII01
22	ManufacturingProcess15	22	Parameter 10	22	Parameter19	22	ManufacturingProcess12	22	BiologicalMaterial10
23	ManufacturingProcess38	23	Parameter 03	23	Parameter16	23	BiologicalMaterial10	23	BiologicalMaterial01
24	ManufacturingProcess36	24	Parameter 04	24	Parameter05	24	ManufacturingProcess04	24	ParameterII06
25	ManufacturingProcess14	25	Parameter 22	25	Parameter14	25	ManufacturingProcess02	25	ParameterII23
26	ManufacturingProcess40	26	Parameter 15	26	Parameter08	26	ManufacturingProcess07	26	BiologicalMaterial11
27	BiologicalMaterial12	27	Parameter 35	27	Parameter20	27	ManufacturingProcess20	27	ManufacturingProcess04
28	ManufacturingProcess06	28	Parameter 01	28	Parameter24	28	ManufacturingProcess32	28	ParameterII19
29	ManufacturingProcess08	29	Parameter 48	29	Parameter18	29	ManufacturingProcess28	29	ManufacturingProcess34
30	ManufacturingProcess24	30	Parameter 06	30	Parameter21	30	BiologicalMaterial04	30	ManufacturingProcess07
...		...				...		...	
31	ManufacturingProcess22	46	Parameter 46			40	ParameterI 29	31	BiologicalMaterial06
32	BiologicalMaterial11	47	Parameter 27			59	ParameterI 46	55	ParameterI27
36	ManufacturingProcess42					84	ParameterI 27	60	ParameterI46
45	BiologicalMaterial06					86	BiologicalMaterial11	88	ManufacturingProcess16
51	ManufacturingProcess16					105	ManufacturingProcess17	103	ParameterI38
57	ManufacturingProcess17							129	ParameterI29
								136	ManufacturingProcess17
								137	BiologicalMaterial07

On the other hand, looking at the two highest ranking features, '*Parameter 29*' and '*Parameter 38*' of process DICK(CHEM), a steady decrease in importance can be observed. Here formerly highest ranking 'Parameter 29' is showing a larger de-

crease with rank no. 40 (TD(CHEM)) and no. 129 (TDH(CHEM)) than 'Parameter 38' with rank no. 11 (TD(CHEM)) and no. 103 (TDH(CHEM)). However, both indicate that features considered important state drivers for individual processes may have steadily decreasing impact considering the final state.

Another interesting observation of the ranking is that the individual processes seem to have different weights overall when it comes to the combined state vectors. When looking at TD(CHEM), it can be observed that only one feature of process DICK(CHEM) is ranked within the top 30, this being 'Parameter 38'. Looking further, in the complete manufacturing programmes vector TDH(CHEM), it is transparent that no feature representing process DICK(CHEM) is ranked within the top 30 most important features. Furthermore, only 8 features from process TOM(CHEM) are among the top 30, leaving 22 features from process HARRY(CHEM) dominating the feature ranking. Speculations that the importance of the processes increases over the sequence cannot be supported by the data as TOM(CHEM) is the first process before DICK(CHEM) in the manufacturing programme. It can be assumed that the processes have a rather individual influence.

In the third scenario, the SECOM process was analyzed. In this case the evaluation was focusing on an individual process instead of a manufacturing programme with multiple processes. However, feature ranking and selection was applied to the SECOM process and the results show similar results to the ones obtained in the prior scenarios.

Overall the presented findings within the evaluation of feature ranking (feature selection) are considered important for answering the research question and thefore the hypotheses. The findings indicate that parameter relationships may vary considerably through the manufacturing programme's process chain and may illuminate some of the known/unknown process intra- and inter-relations discussed earlier from a theoretical point of view. However, this is discussed in more detail in the following section 7.2.

The classification results of the SECOM data set after feature selection is applied show that the results are equally good or better than with the full feature set (see Figure 103 and Figure 104). This confirms the results of the other two scenarios, that by applying feature selection, relevant state characteristics (state drivers) are selected which allow a judgment of the quality performance (final product state) of the product.

7.1.5 Classification performance on previously unknown data

For the Rolls-Royce data set the test of classification performance has been conducted. The results show very good classification performance on previously unknown data (learning 70% & test 30% split). However, these results might not be transferable/generable and thus evaluation may not be possible as SMOTE was ap-

plied previous to provision by Rolls-Royce. With SMOTE applied, the split of the data set does not guarantee that the test data is not (partly) incorporated in the learning set as SMOTE does create additional examples based on existing ones. Thus a bias is involved and the results are questionable in relevance. Thereafter, this question was evaluated using the SECOM data set and the results seem to confirm the suspicion (see Figure 108).

The CHEM data set in scenario two showed poor classification performance when it comes to previously unknown data. However, after applying parameter optimization of the SVM algorithm, SMOTE oversampling of the minority class for the learning set and subsequent feature selection, the classification results of the minority class are acceptable (see Figure 88). The best results show an overall accuracy of 88.5% with three of the four percentages being significantly over the threshold of 80% and just one being slightly below (76.5%). Even so they are lower than the optimized cross-validation results, which is common and would raise suspicion if not so, the results show that it is possible to reach good classification performance with a trained model on previously unknown data. This is a prerequisite for an application of the approach within an industrial manufacturing environment for e.g., monitoring tasks.

The classification results of previously unknown data in a timely sequence along the process is conducted by using the first 70% of examples as the training set and the latter 30% of examples as the test set. The classification performance is not as good as with the random split (see Figure 83). As the distribution of 'fail' examples in the data set was found to be uneven (see Figure 92) and more dense towards the end of the process, this was to be expected. The random split had a similar 'fail'/'pass' ratio whereas in the timely split, the ratio of the learning set was significantly lower than that of the test set. This indicates that the trained model was not able to take the characteristics of most 'fail' examples into consideration which may explain the lower performance. By applying parameter optimization, feature selection and SMOTE oversampling of the learning set, the classification results improved further and thus resembled the tendency of the results from the random split case (see Figure 90). However, similar to the random split, it can be assumed that the performance will improve even further with the increase of examples available for the model generation. Looking at the resulting time plot of the predicted product states (see Figure 91), the classifier enables early identification of disruptions within the manufacturing process and thus allows the process owner to react. This, combined with the reduction of the feature set and thus identifying relevant state drivers, supports process control within the manufacturing system.

Furthermore, if expert knowledge is available, the examples for the learning set can be selected in a supervised fashion and possibly improve the results even further. In this case, the examples within a class do not have such a wide spread and variety as they do in scenario three, making the task at hand very challenging.

The SECOM data set performs poorly when classification of previously unknown data is concerned based on the split of the data set in a learning (70%) and test (30%) set in two different cases. One case is based on a random split of learning and test set with a similar ratio in both. The second case is based on a split in timely succession, with the first 70% of the examples resembling the learning set and the latter 30% the test set. In the second case, the ratio is a little more uneven as the 'fail' examples are not distributed evenly throughout the process. Even so a slight increase in performance could be observed in the first case (see Figure 105 & Figure 106), the difference is marginal, that it neither confirms nor negates the ability of feature selection to improve the classification of previously unknown data. At this point, the classification of previously unknown data has to be considered not applicable for the current SECOM data set.

In order to show how the increase in examples of the minority group will affect the performance of the classification results on previously unknown data the following Figure 108 depicts the results of a learning (70%) and test (30%) split after SMOTE is applied on the SECOM data set ('Approach 2 Var. 2 plus 15'). The SVM algorithm has an ANOVA kernel, with the same optimized parameters identified previously (kernel gamma 2.0; kernel degree 3.0; C 2.6).

	200%			500%			1000%		
528 (split pre-SMOTE)	Accuracy:	91.85%		Accuracy:	91.03%				
	2	28	6.67%	2	28	6.67%			
	2	336	99.41%	5	333	98.52%			
	50.00%	92.31%		28.57%	92.24%				
FS80 (split pre-SMOTE)	Accuracy:	92.12%		Accuracy:	92.12%				
	3	27	10.00%	5	25	16.67%			
	2	336	99.41%	4	334	98.82%			
	60.00%	92.56%		55.56%	93.04%				
528 (split after SMOTE)	Accuracy:	94.85%		Accuracy:	95.54%		Accuracy:	96.23%	
	74	18	80.43%	160	20	88.89%	305	21	93.56%
	4	331	98.81%	3	333	99.11%	4	334	98.82%
	94.87%	94.84%		98.16%	94.33%		98.71%	94.08%	

Figure 108: Comparison of classification performance results on previously unknown data (split after SMOTE application)

In this case it is assumed, that future minority examples are of similar nature as the previous ones. It has to be noted that the different SMOTE percentages are not completely comparable as the random split of the data was applied after enhancing the data set. Therefore the examples contained in learning and test set may vary. However the general direction of the results show significant improvements of the classification when more examples are available. Even with a small amount of SMOTE (200%) applied, the results meet the target threshold for classification performance of cross-validation. This indicates also that the assumption of not comparing the classification of previously unknown data of the RR data set (split after SMOTE) to the ones achieved by the CHEM and SECOM data sets (split pre-SMOTE). The assumption that the SECOM data set minority class contains exam-

ples of very diverse nature is supported by this result. This may ease the pressure of the previous assumption of overfitting to some extent.

In the following sub-section, the presented results are critically discussed based on the previously raised research hypotheses (see section 5.3).

7.2 Discussion of evaluation results

The critical discussion of the evaluation results is based on the three application scenarios and structured around the three main hypotheses. However, this provides only a rough structure, with possible overlaps in argumentation due to the intertwined nature of the results.

Hypothesis 1

At first, the evaluation results relevant for *hypothesis 1* are presented. The hypothesis states that the ***'Capturing of process intra- and inter-relations by implication through application of SVM'*** is possible. This hypothesis was split in two sub-hypothesis. Hypothesis 1.1 ('Application of SVM allows to identify state drivers of individual processes') focusing on the individual process and hypothesis 1.2 ('Combining different processes allows to identify relevant drivers at different phases of the manufacturing programme') focusing on intra-relations within the manufacturing programme.

The evaluation results show that by selecting relevant information as a representation, in this case features/attributes by their weight vector $\boldsymbol{w}$ allows to describe the description of the product and process state. This is in accordance to the *product state concept*'s main idea to identify a set of relevant information by which the product's state can be described comprehensively. As discussed previously, the challenge was to include the process intra- and inter-relations between the state characteristics, both within a process and across process borders as there exists a knowledge gap.

For the individual process (*hypothesis 1.1* ***'Application of SVM allows the identification of state drivers of individual processes'***), the equally good or improved classification performance of the reduced data set in comparison to the original (full featured) data set in all three scenarios confirms that it is possible to identify relevant state characteristics or state drivers by applying feature ranking on manufacturing process data. The feature ranking method based on Guyon et al. (2002) utilized the process intra- and inter-relations between features (or in this case state characteristics/state drivers), and thus includes those considerations in the selection process. As previously stated, the weight vector $\boldsymbol{w}$ is the basis of the feature selection method by Guyon et al. (2002). This emphasizes the role of the weight value of the support vectors in identifying the relevant state drivers of the processes and combined vectors. However, it has to be noted, that the ranking does only rank all

available features (using RFE) but does not provide information concerning the optimal threshold.

In this case, either expert knowledge needs to be included to select the set of relevant features or the set may be established by further experimentation. In scenario II and III further experimentation was applied to chose a promising amount of features through comparing the classification performance by cross-validation. For scenario I, the results of feature selection based on the feature ranking show that the classification performance of a reduced set is very good but not as good as the full feature set. For scenario I, expert knowledge is available, but due to the confidentiality agreement, the experts were not able to provide detailed feedback of what features are already known as relevant process parameters and what features are potential new ones. In this case, the information regarding to what individual parameters mean, contain or measure is not available to the author. However, the qualitative feedback by experts confirms the accuracy of the identification of relevant state drivers by applying SVM based feature selection on product state data.

Another difference of the feature ranking between scenario I and scenario II & III is that for scenario one it was done utilizing the RapidMiner (v5.3) function 'Weight by SVM' which allows an export of the weight values (normalized) and the WEKA function, whereas in scenario II & III only the 'SVMAttributeEval' function based on Guyon et al. (2002) in WEKA was utilized. The WEKA ranking was showing more consistent results in the evaluation and according to the expert feedback in scenario I, the results are more compliant to the existing knowledge of the processes. Therefore, the WEKA ranking was utilized thereafter. However, both approaches show that good classification of acceptable ('pass') and unacceptable ('fail') is possible within a 'real world' manufacturing process by identifying and using relevant state drivers (features).

After confirming the statement of hypothesis 1.1, the following paragraphs focus on the evaluation results with relevance to *hypothesis 1.2* ***'Combining different processes allows the identification of relevant drivers at different phases of the manufacturing programme'***. Hypothesis 1.2 is focussed on the following: the accumulating (combined) vectors are constructed from the TOM, DICH and HARRY vectors in scenario I & II. At stage 1, post process TOM, the state vector is the single process TOM's vector. For each vector a quality assessment is available in the form of a 'pass' and 'fail' label. Similarly, for the second stage, post-process DICK, the state vector TD will be the concatenated vectors TOM and DICK. This is repeated for the final stage post-operation HARRY. In this way the state vectors increase their dimensionality by the number of features of the last process included for each process stage of the manufacturing programme TDH.

In the previous section, it has been shown (see Table 6 & Table 7) that the feature selection applied to combined vectors show variations within the ranked features.

The combined vectors partly rank previously top ranked features (individual processes) rather low and previously low ranked (unimportant) features rather high. Given the established results of feature selection for individual processes by cross-validation performance, it is confirmed that the ranking reflects the relevancy of features correctly.

As the combined vectors do not reflect the ranking of features of the contained individual processes, the variation can be retraced to the cross-process process intra-relations and their influence on the results. It has to be noted that the ranking can only reflect the information available. Therefore, the combined vector has to include all available features of the original processes pre-elimination. This stands true also for information not available to the classifier. Projecting this on the previously used example, if there is no feature available indicating what clamping method was used, the influence on the heat treatment cannot be identified. Therefore it is utmost important to collect as much information as possible prior to feature selection.

It has been shown in previous applications of feature ranking and selection in different domains and has been confirmed by the three manufacturing evaluation scenarios that the identification of relevant information is possible. In conclusion, it was confirmed by the evaluation results that the stated hypothesis is confirmed given the assumption. The limitations of the evaluation approach that may have an impact on this judgment are presented in the final sub-section of this section.

The identification by feature selection and thus incorporation of implicit process intra- and inter-relations on process and programme level supplements the three areas of relevant information identification of the *product state concept* (section 4.3.5). The ML approach is intended to support experts who design the monitoring system for the manufacturing programme. It allows to benefit from previously unknown process intra- and inter-relations relations between state characteristics within processes/operations and across process-borders. It has to be noted that it is not intended as a standalone and fully automated approach at this stage. The approach integrates not only in the previously introduced *product state concept* but supports the intelligent manufacturing vision.

Hypothesis 2

The discussion of *hypothesis 2* ***'Adaptability to changing conditions through application of SVM'*** is based on the results of the three different manufacturing scenarios. The adaptability may be viewed from two general perspectives, first, considering the application domain and second regarding changes in process and/or environmental factors.

Looking at the adaptability of the proposed concept for different domains, the results of the three scenarios indicate that this is quite high within the overall manu-

facturing domain. The concept was successfully applied to three different manufacturing domains:

- Scenario 1 - mechanical manufacturing (see section 6.2)
- Scenario 2 - chemical manufacturing (see section 6.3)
- Scenario 3 - semiconductor manufacturing (see section 6.4)

The three domains chosen as evaluation scenarios represent a wide variety of manufacturing applications and thus allow the conclusion that the concept is applicable and highly adaptable within the manufacturing domain. Even though the wide spread of the chosen scenarios may indicate that the concept may also be applicable within other domains outside of manufacturing, this can neither be confirmed nor negated based on the conducted evaluations. The adaptability of the concept on other processes/process chains, e.g., in the service domain has to be analyzed in future studies (see section 8.2). Given that Guyon et al. (2002) successfully applied the feature selection method in the medical domain indicates a good chance for a successful transfer of the findings to different domains.

The second perspective, focusing on the adaptability of the concept based on changing process and/or environmental factors, can be confirmed. The concept is able to adapt to changing conditions rather quickly as the basic model generation of the SVM algorithm can be adapted as soon as new learning data is available, which allows a real time adaptation. This is supported by the small amount of time needed to train the classifier model with new training data and apply the new model on new classification tasks. However, in order to do that, expert input is necessary to evaluate when the model needs to be updated by new learning examples (supervised learning). Looking at the results of the SECOM data set (scenario 3), the importance of having a meaningful learning set is eminent especially when the approach is supposed to classify formerly unknown date with high accuracy.

Partly related to the second perspective, the computing efforts can be considered reasonable once a set of suitable parameters for the model generation is established. This stands true even for big data sets like the SECOM or RR data set. The model generation and subsequent application of the trained model takes very little computing resources and effort. With regard to computing time, the training of the classifier model requires seconds/minutes rather than hours and can be considered almost real time. This is a significant advantage when it comes to the application within an industrial manufacturing environment.

Finding suitable parameters through optimization of cross-validation classification performance however may require a certain amount of computing resources and effort. In particular for bigger data sets, an optimization run can easily take (significantly) more than 10 hours in RapidMiner (v5.3), even when the optimizing pa-

rameters are split in different runs[22]. This is due to the exponential increase of to-be-calculated cases with every added optimization parameter. A complete optimization run is not required for every adjustment of the training data, therefore in established scenarios, the optimization runs are more likely to be located within a 10-15 minute timeframe.

Overall, the results confirm that the concept is indeed adaptable to different domains (of manufacturing) and changing conditions of the process and/or environmental factors and thus hypothesis 2 can be regarded as confirmed.

Hypothesis 3

The *third hypothesis* ***'Through application of the SVM approach, defect products can be identified'*** focuses on the evaluation of the ability to integrate the approach in a manufacturing programme in the current form in order to improve product and process quality in the sense of an intelligent manufacturing system.

In this case the overall hypothesis was split in two sub-hypothesis, *hypothesis 3.1* ***'the trained SVM system is able to detect faulty products in the manufacturing programme'*** and *hypothesis 3.2* ***'a connection to the identified state drivers can be established within the set of (within the manufacturing programme) identified defect products'***.

Overall, the issues raised in hypothesis 3 cannot be confirmed at this stage based on the evaluation conducted and the obtained findings. Whereas some findings support the hypothesis others do not. For hypothesis 3.1, the findings of the classification performance on formerly unknown data are highly relevant. The three scenarios present very diverse feedback on this issue.

As stated beforehand, even though tests of the classification performance of previously unknown data have been conducted with the RR data set, these results can neither be applied to confirm nor negate the raised hypothesis. This is due to the fact that SMOTE oversampling has been applied prior to provision of the data set. Thus the results may be biased as described in section 7.1.5.

Initially showing low classification performance when it comes to previously unknown data, the TOM(CHEM) manufacturing process classification results improved during the evaluation. After applying parameter optimization of the SVM algorithm, SMOTE oversampling of the minority class for the learning set and feature selection, the classification results of the minority class can be considered ac-

[22] Specifications of machine used: Processor: 2.6GHz dual-core Intel Core i5 processor (Turbo Boost up to 3.1GHz) with 3MB shared L3 cache (fourth generation Intel Haswell); Ram: 8GB of 1600MHz DDR3; SSD: 512GB PCIe; Graphics: Intel Iris 1024 MB; OS: OS X 10.9.2

ceptable (see Figure 88). The best results show overall accuracy of 88.5% with three of the four percentages being significantly over the threshold of 80% and just one being slightly below (76.5%). Even though they are lower than the optimized cross-validation results, the findings confirm that it is possible to reach good classification performance with a trained model on previously unknown data for the TOM(CHEM) data set. This is a prerequisite for an application of the approach within an industrial manufacturing environment e.g., monitoring tasks.

As stated before, looking at the split based on timely sequence of learning/test set, which is the most relevant for industrial application, the uneven distribution of 'fail'/'pass' examples over the process run made a comprehensive model generation difficult and may explain the classification results. By utilizing pre-processing steps similar to the randomly split variant, feature selection and SMOTE oversampling the classification performance on previously unknown data in a timely sequence split could be significantly improved (see Figure 90). Given the challenging starting position, this can be regarded as a very good result. Looking at the time plot of the predicted product states, it can be observed that disruptions within the manufacturing process can be correctly predicted by the classifier at an early stage, allowing for preemptive measures to bring the process back on track. Such a prediction, especially with a significantly reduced feature set may allow the process owner to preemptively adjust the process and reduce the risk of such a relatively long period of manufacturing products without sufficient quality. This result of the evaluation of scenario II can be considered in favor of hypothesis 3 and especially hypothesis 3.1.

It can be assumed that the performance of classification on previously unknown data will improve further with the increase of examples available for the model generation. Furthermore, if expert knowledge is available, the examples for the learning set can be selected in a supervised fashion and possibly improve the results even further. In this case, the examples within a class have not had such a wide spread and variety as they do in scenario three, making the task at hand very challenging.

As previously stated, the poor classification results of the SECOM data set even after elaborate pre-processing through feature selection and SMOTE oversampling do not support the hypothesis. A possible reason may be that the negative examples (minority class) are very diverse in nature which do not allow the classifier to prepare the model accordingly to successfully classify new negative examples. It may be possible that over time the classification results of previously unknown data improve when a bigger selection of minority examples are available to train the model. However, at this point this is speculation and thus the results of scenario three do not confirm the hypothesis.

Table 8: Summary of the results with regard to postulated hypotheses

Research hypotheses		Result of evaluation
1	'Capturing of process intra- and inter-relations by implication through application of SVM'.	The evaluation results confirm the hypothesis statement that by applying feature ranking based on SVM it is possible to capture process intra- and inter-relations by implication throughout the manufacturing programme.
1.1	'Application of SVM allows the identification of state drivers of individual processes'	The performance of processes with a reduced feature set are equally good or better than the ones using the full feature set. This confirms that by selecting relevant information, the product and process state can be sufficiently described. Hence, hypothesis 1.1 is confirmed.
1.2	'Combining different processes allows the identification of relevant drivers at different phases of the manufacturing programme'	Evaluation results of scenario I & II show that the combined vectors' feature ranking differ from the individual rankings for specific features. This confirms that cross-process process intra-relations are reflected in the results and all relevant state characteristics of a manufacturing programme may be identified by applying the proposed method. This confirms hypothesis 1.2.
2	'Adaptability to changing conditions through application of SVM'	The three scenarios describe very diverse domains and data sets. The approach was shown to be applicable to all three scenarios which shows its broad applicability and adaptability. The relative ease of adapting the learning set in case new examples, expert knowledge and/or attributes are available confirms the adaptability of the approach. Determining the classes of the synthetic processes in two different ways with comparable results, shows additionally the adaptability and broad applicability of the concept. Adaptability of the model based on the learning set and application does not require much computing effort and resources once suitable parameter configuration has been established. This allows for a fast creation and application of model-updates as soon as new examples for the learning set are available e.g., when process and/or environmental factors change over the course of a manufacturing programme. Therefore, hypothesis 2 can be considered confirmed for (manufacturing) processes with a similar structure of the data sets.
3	'Through application of the SVM approach, defect products can be identified'	Hypothesis 3 could not be confirmed nor negated during the evaluation within this dissertation. Whereas some results indicate that the hypothesis is true, others do not entirely support the hypothesis at this stage. To fully elaborate this issue, further data and in process application and evaluation is necessary.
3.1	'Trained SVM system is able to detect faulty products in the manufacturing programme'	Whereas the results of the classification performance on formerly unknown data partly support this sub-hypothesis, they do not confirm it beyond doubt over all evaluated scenarios. In this specific area, a collaborative evaluation in process with an industrial partner is necessary to answer the raised research question.
3.2	'A connection to the identified state drivers can be established within the set of (within the manufacturing programme) identified defect products'	The results confirm that the approach takes implicit process intra- and inter-relations into account and identifies the relevant state drivers. However, at this stage, the results allow no connection of process intra- and inter-relations and individual state drivers to defects. In order to do that, expert input is needed. Furthermore, the SVM approach does take process intra- and inter-relations into consideration but does not allow to extract those. Therefore this sub-hypothesis cannot be confirmed at this stage.

For hypothesis 3.2, the results confirm that the approach takes implicit process intra- and inter-relations into account and identifies the relevant state drivers. However, at this stage, the results allow no connection of process intra- and inter-relations and individual state drivers to defects. In order to do that, expert input is needed. Furthermore, the SVM approach does take process intra- and inter-

relations into consideration but does not allow to extract those. Therefore, this sub-hypothesis cannot be confirmed at this stage.

Overall, it shows that in order to fully explore the questions raised in hypothesis 3, further research, ideally in collaboration with industry directly in the processes, is necessary. At this stage, the results from e.g., the predictive states in time plots of the TOM(CHEM) process indicate an outcome in favor of the raised hypothesis.

After the previous discussion of the evaluation results structured roughly around the raised hypotheses, Table 8 provides a summary of the findings.

7.3 Limitations

In this section, the limitations of the concept are illustrated and discussed.

First of all, the chosen approach proved useful with quantifiable parameters during the evaluation. However, theoretically the state can also be described by using qualitative parameters. Even if those are digitized, it may prove difficult to use qualitative parameters in methods, which depend upon the needs to calculate distances between vectors like the chosen SVM. There are several methods available to transform qualitative measures in quantitative ones which can be utilized with supervised learning methods (e.g., Bratko & Suc, 2003). If such an approach is applicable within the developed concept, it needs to be studied in future research.

The synthetic data sets used for the evaluation in the first and second scenario may not represent a 'real world' manufacturing data set in all nuances. Even though an effort was made to synthesize the data set as close to the realistic manufacturing "master data set" as possible, a complete (100%) accurate simulation of real world data may not be possible. To reflect two different extremes, the synthetic processes and combined vectors of scenario I & II are tailored to show a rather good classification performance (scenario I) and a more challenging classification performance (scenario II). This is only based on the assigned class labels, not on the process data itself. The synthetic processes resemble their 'parent' real world processes by characteristics such as mean and standard deviation. An effort was made to change the definition of class (by cluster in scenario I & random in scenario II) in order to show that the results are comparable in different variations. However, the main findings are mostly of methodological benefit. The results show that such process intra- and inter-relations and driving states can be identified by the approach and how the results look like. Furthermore, the interpretation is a main finding. It does not and is not intended to represent evidence that the same results may be obtained when applying the approach to a 'real world' manufacturing programme. It has been shown that even the three 'real world' data sets behave in different ways when it comes to applying classification algorithms. Nevertheless, the successful application of the approach on three different 'real world' data sets shows comparable and similar results over all scenarios.

The data set on which scenario II is based upon is published as a regression data set. Therefore, there are no two classes defined by design but a quality feature 'Yield'. The approach of selecting a certain threshold (Yield 39) to divide the data set in 'pass' and 'fail' classes may be the reason for the partly low classification accuracy. However, overall the data set behaved rather well and the results are comparable to the ones obtained within the other two scenarios. As 'Yield' is often used as an important quality measure, this is not surprising. Nevertheless, this limitation has to be taken into consideration.

A rather basic limitation is the needed resources for the application of the approach. The resources needed to create a feature ranking may take time and computational effort. This stands especially true for large data sets with high number of examples and features (e.g., scenario III). However, the generation of the classification model and its application does not require significant resources. Therefore, the limitation based on needed (computational) resources does not effect the application of the approach but the efforts beforehand.

Using the built-in feature selection function based on Guyon et al. (2002) in WEKA for scenario II & III does not allow to extract the weight vector $\boldsymbol{w}$ values. This would allow to determine the threshold in a different way and the results could be evaluated based on the classification performance. In these cases, a variety of thresholds was tested and compared to identify a well performing set of features. However another limitation of the method is that it does not indicate an optimal number of features to be selected for the best classification performance. This has to be done by manual experimentation and thus represents a challenge.

Applying another function for SVM feature ranking (RapidMiner (v5.3)) instead of the WEKA function allows to extract the weight vector $\boldsymbol{w}$ values. However, the two functions are not completely comparable. Therefore, the results of scenario I and scenario II & III may be not 100% comparable when it comes to feature ranking. In this case, the option to extract the weight value was regarded more important than complete comparability. In any case, the WEKA feature ranking option is regarded the more applicable one as it is directly based on Guyon et al. (2002)'s approach.

By dividing the data set in learning (70%) and test (30%) set randomly, the comparability between different performance results is not guaranteed 100%. Depending which examples are chosen randomly for each set, the training of the model may vary to some extent as does the to-be-classified test set. Depending on the amount of outliers among the examples within a set (learning or test) this may or may not influence the performance significantly. To reduce this effect, the same version of the split data set (same examples for each set) is utilized in the different evaluations in order to ensure comparability within the analysis.

Missing data or corrupted data represents a significant challenge and limitation to the approach. In 'real world' manufacturing process/product data, missing values are a common problem accompanied by challenges like noise, redundancy and/or inconsistency (Zhang, Jin, Zhu & Zhang, 2009). For the presented approach, missing values (incomplete) within the data set need to be removed/replaced in order to determine a learning set. However, as can be observed in the SECOM data set, missing values are often not distributed equally within a manufacturing data set. Often certain state characteristics (process parameters and features/attributes) contain a significantly higher number of missing values than others. Depending on the strategy to handle missing values within a data set, either information may get lost or a certain bias may be introduced. In the chosen approach to eliminate the missing values applying a 4-stage method (see section 9.2.3), the feature space was reduced to 528 (89,49%) from the original 590 features. A comparison of the results, identified 'state drivers' with the data sets obtained using 'approach 1' and 'approach 2 variant 1' show that when features are eliminated (reduction of dimension) also information is lost. For example feature 165 represents a relevant driver within the data set (Top 50, ranking no. 2, 'Approach 2 Var. 2 plus 15') but is not part of the top 50 ranked relevant features of 'Approach 1' data set due to the elimination process. This highlights the influence of the data pre-processing on the approach when handling 'real world' manufacturing data. This challenge will hopefully be decreasing in importance over time with sensor technology and other data capturing technologies developing at a fast pace and provide data sets with less missing values and noise.

Ideally all identified state characteristics may be included especially in combined vectors (to identify cross-process process intra-relations) as the SVM algorithm is able to handle large dimensionality (1000+). If it becomes necessary to limit dimensionality beforehand, then one will have to start selecting. In such a case, the presented dimension reducing methods can be used based on the feature weight rather than removing variable according to our limited knowledge. However, the effect of a selection of features prior to the feature ranking of combined vectors has to be studied. The evaluation conducted within this dissertation did not include such variations.

A rather important limitation of the proposed concept, being mentioned before, is that if state characteristics (features) are not 'measured' (and thus included in data) they cannot be ranked and identified as important. Furthermore, possible process intra- and inter-relations between state characteristics cannot be taken into consideration and identified. However, this limitations leads to a potentially important benefit of the concept. By incorporating all possible measures, (even those which relations are not known or expected to have no impact) allows to identify relevant ones. This way, even formerly neglected state characteristics may prove important in one way or another. This may present a starting point for further investigations on that particular state driver. This represents a chance at the same time where this

concept may have a significant impact on the understanding and transparency of manufacturing programmes. At the same time it has to be understood that the concept will not automatically identify and extract all important (relevant) features and process intra- and inter-relations for all processes without expert input. It presents a tool to support experts in their work and utilize their knowledge.

Overall, it can be stated that there are considerable limitations to the concept which may influence the applicability in practice to a certain extent. Some of the mentioned limitations may represent starting points for future research efforts.

8 Recapitulation

This section, structured in two subsections summarizes the research work and concludes the findings before giving a short outlook into potential future directions in this research domain.

8.1 Conclusion

As initially stated, the manufacturing domain faces major challenges which may be summarized by increasing complexity and dynamics of products and processes as well as increasing requirements towards quality. The research problem of this thesis is set in multi-stage manufacturing programmes and focuses on the holistic handling of information with the goal of improving product and process quality. Today, existing solutions focus mostly on individual processes instead of the whole manufacturing system and do not incorporate product and process inter- and intra-relations. It was found that these process inter- and intra-relations can have a significant and varying impact on the quality outcome of successive processes and thus on the whole manufacturing programme.

In the dissertation, the *product state concept* has been developed as a method to describe comprehensively a product by its states along a complete manufacturing programme. A core mechanism of this concept is the description of the product state by a set of state characteristics. The fundamental question of how to identify this set of state characteristics to allow a comprehensive description of the products state, set the foundation for the conducted research. A major aspect within the work was found to be process intra- and inter-relations between state characteristics, later referred to as state drivers. Today, most manufacturing programmes lack sufficient knowledge and transparency with regard to process intra- and inter-relations making a complete modeling of the system unrealistic. In order to be able to incorporate this crucial element in the analysis, supervised machine learning was employed in form of SVM based feature ranking to incorporate successfully implicit process intra- and inter-relations of the manufacturing programme.

The evaluation of the research was conducted by using three different scenarios from distinctive manufacturing domains based on 'real world' data sets. The first scenario represented the mechanical manufacturing domain, blade manufacturing, with a case provided by Rolls-Royce. The second scenario focused on the chemical manufacturing domain and the third scenario resembled a semiconductor manufacturing case. The purpose of choosing three different scenarios was to highlight the general applicability of the developed concept. The evaluation confirmed that it is possible to incorporate implicit process intra- and inter-relations on process as well as programme level as required by the *product state concept* through applying SVM based feature ranking. Even so the results confirm that the approach successfully utilizes the implicit process intra- and inter-relations between states and state

characteristics, at this point the relations are not provided as an explicit output of the analysis. However, they are implicitly included within the section or relevant state drivers. In this regard, expert knowledge is still a crucial factor for the successful application of the concept in manufacturing.

Concluding, the presented *product state concept* allows to identify relevant state drivers of complex manufacturing systems. The concept is able to utilize complex, diverse and high-dimensional data sets which often occur in manufacturing applications. This fits nicely with current initiatives like 'Industrie 4.0', 'Cyber Physical Systems' in Europe and the 'Industrial Internet' and 'Advanced Manufacturing Partnership' in the US as well as the growing area of Big Data research. It can be safely said that in the near future, the amount of data derived from manufacturing operations will increase due to these developments. This offers both opportunities and challenges for manufacturing companies and manufacturing research. With the developed concept, the increasing data streams can be analyzed efficiently and applicable results can be derived. The analysis results present a direct benefit in form of the most important process parameters and state characteristics, the state drivers, of the manufacturing system. These can be directly utilized in, e.g., quality monitoring and advanced process control. Additionally, the results represent a first indication of what processes and parameters may benefit from a more in-depth analysis. This way, the *product state concept* indirectly contributes to a sustainable growth of knowledge in manufacturing.

8.2 Outlook and future work

During the research conducted within the framework of this dissertation a variety of topics emerged which may be worthwhile to trigger further investigations. In this section, a short outlook is presented illustrating some of those areas of future research.

One of the bigger aspects of future research is the possibility to apply and evaluate different approaches of feature selection. This includes a combination of other supervised and unsupervised ML methods with the previously applied SVM approach (e.g., random forest). This is expected to strengthen the focus vis-a-vis the importance of the state variables. The necessary tools are overall readily available in the RapidMiner (v5.3), R or WEKA suit. Another interesting aspect of feature ranking and feature selection is to investigate the optimal threshold for a feature set (feature selection). So far the optimal amount of features is not part of the feature selection technique (Guyon et al., 2002).

Going in the same direction, it may be worth investigating to apply a combination of ML algorithms when creating a monitoring model. Recent advances of e.g., RL show promising results in similar application scenarios. A possibility would be to utilize the developed approach to determine relevant state drivers of the manufac-

turing programme and subsequently set up a RL model with those features. This would allow utilizing the advantages of both techniques.

Even though as of today it is not common practice in manufacturing to consider the inner product state characteristics, technological development will provide more and advanced tools, which allow an efficient and economical capturing of more data points. In the wood market it is already possible to scan whole trunks through a computer tomography scanner in order to plan the following processes according to the given (internal) structure of the wood (e.g., knobs and knots). It can be assumed that in the near future it is possible to measure and collect more data that is not only more accurate but also requires lower investments. As stated earlier, the more data is available the more implicit process intra- and inter-relations may be incorporated by applying the proposed concept.

As mentioned before, knowledge about the customer requirements and the degree of fulfillment by the product is important to determine 'good' and 'bad' states within the concept. Looking into the usage (middle-of-life) in order to identify quality problems, which occur after the delivery to the customer, may help to identify additional state drivers and support process and product quality improvement further. The authors developed a supporting concept to derive information about the usage of products during middle- and end-of-life from all stakeholders involved. The so-called product avatar concept may allow to access additional information and knowledge which is not easily accessible for the manufacturers by creating a digital counterpart for interaction purposed between different stakeholders of a product (Wuest, Hribernik & Thoben, 2012a; Wuest, Hribernik & Thoben, 2013a; Wuest et al., 2014a). However the integration of the two concepts has to be evaluated more closely in future research.

Another important aspect related to the presented research is to analyze the transferability of the findings, which are valid only for manufacturing, to other product lifecycle phases e.g., the usage phase (middle-of-life). Here an application in maintenance (health) monitoring could be feasible and has to be investigated further. Furthermore, a possible connection of identified state drivers from the manufacturing phase may prove useful in the design phase, more specifically, the conceptual design phase. Transferability of gained knowledge concerning the relevancy of certain state characteristics into functional requirements and/or design paramters needs to be investigated further.

Studying the transferability of results to other domains than engineering could be beneficial, here especially, the health or education domain seem promising. Also an application within a service ("service state") environment may allow new insights and improvements in the field. However, without further research no statement of transferability can be given at this point. Apart from researching transfer possibilities of the findings to other lifecycle phases, the transfer to other domains

may be beneficial as well. For example, the health care industry might profit from certain findings and ideas of the *product state concept* and an adaption towards a health state concept could be discussed in order to support transparency of health monitoring. One possibile application within this domain might be to interpret different examinations/examination results as processes and combined process vectors to utilize the implicit process intra- and inter-relations. Another lever may be to include environmental factors and different stages in the analysis of genes. However, this needs to be done in close collaboration with experts in the respective field.

A visionary goal may be a self-assessing/analyzing manufacturing system within the product state framework supported by the developments in AI, (supervised and unsupervised) ML and in sensor technology.

References

Abe, S. (2003). On Invariance of Support Vector Machines. In *4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'03)*. Hong Kong, China.

Abe, S. (2010). *Feature Selection and Extraction*. In Abe, S. (eds.) Support Vector Machines for Pattern Classification - Advances in Pattern Recognition (p. 471). London: Springer. doi:10.1007/978-1-84996-098-4

Abowd, J. M., & Lane, J. (2004). *New Approaches to Confidentiality Protection: Synthetic Data , Remote Access and Research Data Centers*. In J. Domingo-Ferrer & V. Torra (Eds.), Privacy in Statistical Databases, Lecture Notes in Computer Science Vol. 3050 (pp. 282–289). Berlin Heidelberg: Springer. doi:10.1007/978-3-540-25955-8_22

Abramovici, M., & Sieg, O. (2001). PDM - Technologie im Wandel - Stand und Entwicklungsperspektiven. Orientierung für die Praxis. *Industrie Management, 5*, 71–75.

Abramovici, M. (2007). Future trends in Product Lifecycle Management (PLM). *In: Krause, F.-L. (2007). The Future of Product Development – Proceedings o the 17th CIRP Design Conference*. Berlin, Heidelberg: Springer, 665-674.

Aggarwal, C. (Ed.). (2013). *Managing and Mining Sensor Data*. New York: Springer. doi 10.1007/978-1-4614-6309-2

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Machine Learning: ECML 2004*.

Albino, V., Pontrandolfo, P. & Scozzi, B. (2002). Analysis of information flows to enhance the coordination of production processes. *International Journal of Production Economics, 75*(2002), 7-19.

Almeida, F. L. F. (2011). Designing and implementation of an intelligent manufacturing system. *Journal of Industrial Engineering and Management, 4*(4), 718–745. doi:10.3926/jiem.371

Alpaydın, E. (2010). *Introduction to Machine Learning (Second Edition)*. Cambridge, USA: The MIT Press.

Alvo, M. & Park, J. (2002). Multivariate non-parametric tests of trend when the data are incomplete. *Statistics & Probability Letters, 57*(3), 281–290. doi:10.1016/S0167-7152(02)00062-7

Anderl, R., Picard, A., & Albrecht, K. (2013). Smart Product Engineering. *In M. Abramovici & R. Stark (Eds.), 23rd CIRP Design Conference* (pp. 1–10). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-30817-8

Angel, E. & Zissimopoulos, V. (1998). On the quality of local search for the quadratic assignment problem. *Discrete Applied Mathematics, 82*(1-3), 15–25. doi:10.1016/S0166-218X(97)00129-7

Apley, D. & Shi, J. (2001). A Factor-Analysis Method for Diagnosing Variability in Mulitvariate Manufacturing Processes. *Technometrics, 43*(1), 84–95.

Apte, C., Weiss, S., & Grout, G. (1993). Predicting Defects in Disk Drive Manufacturing: A Case Study in High Dimensional Classification. *In IEEE Annual Computer Science Conference on Artificial Intelligence in Application, Los Alamitos*, 212–218.

Arbor, A. (2000). Modeling and diagnosis of multistage manufacturing processes – Part I – State space model. *In JAPAN/USA Symposium on Flexible Automation 2000* (p. 8).

Arif, F., Suryana, N., & Hussin, B. (2013). A Data Mining Approach for Developing Quality Prediction Model in Multi-Stage Manufacturing. *International Journal of Computer Applications, 69*(22), 35–40.

Auer, T. (2010). *ABC des Wissensmanagements* (p. 41). Hedingen. Retrieved from http://www.pwm.at/file_upload/km_abc_v3.pdf

Augustin, S. (1990). *Information als Wettbewerbsfaktor: Informationslogistik – Herausforderung an das Management*. Köln: TÜV Media GmbH.

Aytug, H., Bhattacharyya, S., Koehler, G. J., & Snowdon, J. L. (1994). A Review of Machine Learning in Scheduling. *IEEE Transactions on Engineering Management, 41*(2), 165–171.

Aytug, H., Khouja, M., & Vergara, F. E. (2003). Use of genetic algorithms to solve production and operations management problems: A review. *International Journal of Production Research, 41*(17), 3955–4009. doi:10.1080/00207540310001626319

Azadeh, A., Saberi, M., Kazem, A., Ebrahimipour, V., Nourmohammadzadeh, A., & Saberi, Z. (2013). A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization. *Applied Soft Computing, 13*(3), 1478–1485. doi:10.1016/j.asoc.2012.06.020

Badri, M. a., Davis, D. & Davis, D. (1995). A study of measuring the critical factors of quality management. *International Journal of Quality & Reliability Management, 12*(2), 36–53. doi:10.1108/02656719510080604

Baker, A. D. (1988). Complete manufacturing control using a contract net: A simulation study. *In International Conference on Computer Integrated Manufacturing, 1988* (pp. 100–109). doi:10.1109/CIM.1988.5399

Barse, L. E., Kvarnström, H., & Jonsson, E. (2003). Synthesizing test data for fraud detection systems. In *19th Annual Computer Security Applications Conference (ACSAC 2003)* (p. 11).

Batini, C., Ceri, S. & Navathe, S. B. (1992*). Conceptual Database Design. An Entity-Relationship-Approach.* Redwood City: Addison Wesley.

Batini, C. & Scannapieca, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Heidelberg-Berlin: Springer.

Beach, R., Muhlemann, A. P., Price, D. H. R., Paterson, A., & Sharp, J. A. (2000). The selection of information systems for production management: An evolving problem. *International Journal of Production Economics, 64*(1-3), 319–329. doi:10.1016/S0925-5273(99)00069-9

Becker, J. (1998). *Die Grundsätze ordnungsmäßiger Modellierung und ihre Einbettung in ein Vorgehensmodell zur Erstellung betrieblicher Informationsmodelle.* Retrieved April 15th, 2012, from http://www.wi-inf.uni-duisburg-essen.de/MobisPortal/pages/rundbrief/pdf/Beck98.pdf

Becker, J. & Schütte, R. (2004). *Handelsinformationssysteme.* Frankfurt am Main: Redline Wirtschaft.

Becker, T. (2008). *Prozesse in Produktion und Supply Chain optimieren.* (2nd Edition). Berlin Heidelberg: Springer Verlag.

Ben-hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. *In Data Mining Techniques for the Life Sciences. Methods in Molecular Biology Volume 609* (pp. 223–239). Totowa, NJ: Humana Press. doi:10.1007/978-1-60327-241-4_13.

Bhuvaneswari, E., & Dhulipala, V. R. S. (2013). The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset. *Information Engineering, 2*(1), 6–13.

Bi, J., Bennett, K. P., Embrechts, M., Breneman, C. M., & Song, M. (2003). Dimensionality reduction via sparse support vector machines. *The Journal of Machine Learning Research, 3*, 1229–1243.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (p. 749).* New York: Springer.

Borin, A., Ferrão, M. F., Mello, C., Maretto, D. A., & Poppi, R. J. (2006). Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Analytica Chimica Acta, 579*(1), 25–32. doi:10.1016/j.aca.2006.07.008

Borovicka, T., Jirina, J. M., Kordik, P., & Jirina, M. (2012). *Selecting representative data sets.* In Advances in Data Mining Knowledge Discovery and Applications (pp. 43–70).

Borror, C., Montgomery, D. & Runger, G. (1999). Robustness of the EWMA Control Chart to Non-Normality. *Journal of Quality Technology, 31*(3), 309–316.

Bowden, R., & Bullington, S. F. (1996). Development of manufacturing control strategies using unsupervised machine learning. *IIE Transactions, 4*(June 2013), 319–331.

Bradley, P., & Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines. *In J. Shavlik, J. (eds.). Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. San Francisco, California (pp. 82–90).

Branch, J.W., Giannella, C., Szymanski, B., Wolff, R. & Kargupta, H. (2013). In-network outlier detection in wireless sensor networks. *Knowledge and Information Systems, 34*(1), 23-54. doi 10.1007/s10115-011-0474-5.

Bratko, I., & Suc, D. (2003). Qualitative data mining and its applications. *Journal of Computing and Information Technology, 3*, 145–150. doi:10.1109/ITI.2003.1225313

Brecher, C., Müller, S., Breitbach, T., & Lohse, W. (2013). Viable System Model for Manufacturing Execution Systems. *Procedia CIRP, 7*, 461–466.

Brinkheinrich, M. (2008). Transparenz durch Traceability. Rückverfolgbarkeit von Produkten und Produktionsprozessen - warum und wie. *Der Betriebsleiter, 6*(2008), 18-19.

Brinksmeier, E. (1991). *Prozeß- und Werkstückqualität in der Feinbearbeitung.* Fortschritt-Berichte VDI, Reihe 2: Fertigungstechnik (p. 256). Düsseldorf: VDI-Verlag.

Brinksmeier, E. & Brockhoff, T. (1996). Utilization of Grinding Heat as a New Heat Treatment Process. CIRP Annals - Manufacturing Technology, 45(1), 283–286. doi:10.1016/S0007-8506(07)63064-9

Brockhoff, T. (1999). Grind-Hardeing: A Comprehensive View. CIRP Annals - Manufacturing Technology, 48(1), 255–260.

Brun, Y. (2008). Solving NP-complete problems in the tile assembly model. *Theoretical Computer Science, 395*(1), 31–46. doi:10.1016/j.tcs.2007.07.052

Buhr, A., Graf, W., Power, L.M. & Amthauer, K. (2005). *Almatis global product concept for the refractory industry.* Retrieved September 27, 2008, from http://www.almatis.com/download/ technical-papers/UNITECR05-180.pdf

Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry, 26*(1), 5–14.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

Cao, H., & Folan, P. (2012). Product life cycle: the evolution of a paradigm and literature review from 1950 – 2009. *Production Planning & Control, 23*(8), 641–662.

Cassina, J., Cannata, A. & Taisch, M. (2009). Development of an Extended Product Lifecycle Management through Service Oriented Architecture. *In Proceedings of the 1st CIRP Industrial Product-Service Systems (IPS2) Conference* (pp. 118–124). Cranfield.

Cawley, G. C. & Talbot, N. L. C. (2010). Over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079-2107.

Çaydaş, U., & Ekici, S. (2010). Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel. *Journal of Intelligent Manufacturing, 23*(3), 639–650. doi:10.1007/s10845-010-0415-2

CEN. (2003). Manufacturing processes - Terms and definitions, division (DIN EN ISO 8580:2003). Comité Européen de Normalisation (english: European Committee for Standardization).

CEN. (2005). Quality management systems – fundamentals and vocabulary (DIN EN ISO 9000:2005). Comité Européen de Normalisation (english: European Committee for Standardization).

CEN. (2008). Quality management systems – requirements (DIN EN ISO 9001:2008). Comité Européen de Normalisation (english: European Committee for Standardization).

Chand, S., & Davis, J. (2013). What is Smart Manufacturing? *Time.*

Chander, A., Dean, D. & Mitchell, J. C. (2001). A state-transition model of trust management and access control. *In Proceedings of the 14th IEEE Computer Security Foundations Workshop, 2001. (pp. 27–43).* Ieee. doi:10.1109/CSFW.2001.930134

Chang, Y., & Lin, C. (2008). Feature Ranking Using Linear SVM. *In JMLR: Workshop and Conference Proceedings 3* (pp. 53–64).

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Intelligence Research, 16*, 321–357.

Chawla, N. V. (2010). *Data mining for imbalanced datasets: An overview.* In Maimon, O. & Rokach, L. (Eds.). Data Mining and Knowledge Discovery Handbook (pp. 875–886). Springer. doi:10.1007/978-0-387-09823-4_45

Chen, J., Su, H., Wu, C., & Oh, C. H. (2012). *Quantumness of Product States.* eprint arXiv:1204.1798.

Cherkassky, V., & Ma, Y. (2009). Another look at statistical learning theory and regularization. *Neural networks : the official journal of the International Neural Network Society*, *22*(7), 958–69. doi:10.1016/j.neunet.2009.04.005

Chinnam, R. B. (2002). Support vector machines for recognizing shifts in correlated and other manufacturing processes. *International Journal of Production Research, 40*(17), 4449–4466. doi:10.1080/00207540210152920

Chinnam, R. B. & Baruah, P. (2009). Autonomous diagnostics and prognostics in machining processes through competitive learning-driven HMM-based clustering. *International Journal of Production Research, 47*(23), 6739–6758. doi:10.1080/00207540802232930

Choe, J. (2004). The consideration of cultural differences in the design of information systems. *Information & Management, 41*(5), 669–684. doi:10.1016/j.im.2003.08.003

Choi, J. (2010). A selective sampling method for imbalanced data learning on support vector machines. *AAAI'2000 workshop on imbalanced data sets*. Iowa State University.

Chou, Y., Polansky, A. & Mason, R. (1998). Transforming Non-Normal Data to Normality in Statistical Process Control. *Journal of Quality Technology, 30*(2), 133–141.

Choudhary, a. K., Harding, J. a., & Tiwari, M. K. (2009). Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing, 20*(5), 501–521. doi:10.1007/s10845-008-0145-x

Christopher, M. (2005). *Logistics and Supply Chain Management – Creating Value-Adding Networks*. Harlow: FT Prentice Hall.

Chryssolouris, G. & Guillot, M. (1988). An AI approach to the selection of process parameters in intelligent machining. *Proc. of The Winter Annual Meeting of The ASME on Sensors and Controls for Manufacturing,* Chicago, Illinois.

Chryssolouris, G., Mavrikios, D., Papakostas, N., Mourtzis, D., Michalos, G., & Georgoulias, K. (2009). Digital manufacturing: history, perspectives, and outlook. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 223*(5), 451–462. doi:10.1243/09544054JEM1241

Cios, K. J., & Kurgan, L. A. (2002). Trends in Data Mining and Knowledge Discovery. In N. R. Pal & L. Jain (Eds.), *Knowledge discovery in advanced information systems* (pp. 1–26). Heidelberg: Springer.

Ciao, K.J., Pedrycz, W., Swiniarski, R.W., & Kurgan, L.A. (2007). *Data Mining: A Knowledge Discovery Approach*. New York: Springer.

Collins, J. (1980). Integrated manufacturing - the state of the art. *The production engineer,* (June), 41–44.

Cook, S. A. (1971). The Complexity of Theorem-Proving Procedures. *In STOC '71 Proceedings of the third annual ACM symposium on Theory of computing* (pp. 151–158). doi:10.1145/800157.805047

Cooper, R. G. (2008). Perspective: The Stage-Gates Idea-to-Launch Process – Update, What's New and NexGen Systems. *Journal of Product Innovation Management, 25*(3) 213–232.

Cooper, R. G. (2010). *Top oder Flop in der Produktentwicklung*. 2. Auflage. Weinheim.

Corsten, H. & Gössinger, R. (2008). *Lexikon der Betriebswirtschaftslehre*. 5. Aufl., Oldenburg.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*, 273–297.

Crama, Y., & Klundert, J. J. (1997). *Robotic flowshop scheduling is strongly NP-complete.* METEOR, Maastricht research school of Economics of Technology and Organizations.

Cristianini, N. & Shawe-Taylor, J., (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

CRAN-R. (2014). *AppliedPredictiveModeling: Functions and Data Sets for 'Applied Predictive Modeling'*. Retrieved February 16, 2014 from: http://cran.r-project.org/web/packages/AppliedPredictiveModeling/index.html

Dasu, T. & Johnson, T. (2003). Exploratory data mining ad data cleaning. Wiley Series in Probability and Statistics (Book 442). Wiley-Interscience.

Davenport, T. H., De Long, D. W., & Beers, M. C. (1998). Successful knowledge management projects. *Sloan Management Review, 39*(2), 43–57.

De Groot, P. J., Postma, G. J., Melssen, W. J., & Buydens, L. M. C. (1999). Selecting a representative training set for the classification of demolition waste using remote NIR sensing. *Analytica Chimica Acta, 392*(1999), 67–75.

De Weck, O. L., Ross, A. M., & Rhodes, D. H. (2012). Investigating Relationships and Semantic Sets amongst System Lifecycle Properties (Ilities). In *Third International Engineering Systems Symposium CESUN 2012* (pp. 18–20), Delft University of Technology, 18-20 June 2012.

Deja, M., & Siemiatkowski, M. S. (2012). Feature-based generation of machining process plans for optimised parts manufacture. *Journal of Intelligent Manufacturing*. doi:10.1007/s10845-012-0633-x

Denkena, B. & Tönshoff, H. K. (2011). Spanen. Grundlagen. (3rd Edition). Heidelber: Springer. doi: 10.1007/978-3-642-19772-7

Denton, B., Gupta, D. & Jawahir, K. (2003). Managing Increasing Product Variety at Integrated Steel Mills. *Interfaces, 33*(2), 41-53.

Devadason, F.J. & Lingam, P.P. (1997). A Methodology for Indentification of Information Needs of Users. *IFLA Journal, 23*(1), 41-51.

Dhafr, N., Ahmad, M., Burgess, B. & Canagassababady, S. (2006). Improvement of quality performance in manufacturing organizations by minimization of production defects. *Robotics and Computer-Integrated Manufacturing, 22*(5-6), 536–542. doi:10.1016/j.rcim.2005.11.009

Dijkman, M. (2009). *Automated Compensation of Distortion in the Production Process of Bearing Rings*. Dissertation Universität Bremen, 2009. Aachen: Verlagshaus Mainz GmbH.

Ding, Y., Shi, J. & Ceglarek, D. (2002). Diagnosability Analysis of Multi-Station Manufacturing Processes. *Journal of Dynamic Systems, Measurement, and Control, 124*, 1 - 13.

Dingli, D. J. (2012). *The Manufacturing Industry – Coping with Challenges* (Working Paper No . 2012 / 05). Maastricht.

Doltsinis, S., Ferreira, P., & Lohse, N. (2012). Reinforcement Learning for Production Ramp-Up: A Q-Batch Learning Approach. *In 11th International Conference on Machine Learning and Applications (pp. 610–615)*. Ieee. doi:10.1109/ICMLA.2012.113

Du, R., Elbestawi, M. A., & Wu, S. M. (1995). Automated Monitoring of Manufacturing Processes, Part 1: Monitoring Methods. *Journal of Engineering for Industry, 117*(2), 121. doi:10.1115/1.2803286

EC. (2009). *Intelligent Manufacturing Systems – Background*. European Commission. Retrieved from: http:// http://cordis.europa.eu/ims/background_en.html

Eibe, F., Hall, M. & Holland, K. (2014). Class SVMAttributeEval. WEKA package Attribute Selection. Retrieved March 27, 2014 from http://weka.sourceforge.net/doc.stable/weka/attributeSelection/SVMAttributeEval.html

El-naqa, I., Yang, Y., Wernick, M. N., Galatsanos, N. P., & Nishikawa, R. M. (2002). A Support Vector Machine Approach for Detection of Microcalcifications. *IEEE TRANSACTIONS ON MEDICAL IMAGING, 21*(12), 1552–1563.

Ellram, L., & Krause, D. (1994). Supplier partnerships in manufacturing versus non-manufacturing firms. *International Journal of Logistics, 5*(1), 43–53.

ElMaraghy, H. A. (2006). Flexible and reconfigurable manufacturing systems paradigms. *International Journal of Flexible Manufacturing Systems, 17*(4), 261–276. doi:10.1007/s10696-006-9028-7

Elmaraghy, W., Elmaraghy, H., Tomiyama, T., & Monostori, L. (2012). Complexity in engineering design and manufacturing. *CIRP Annals - Manufacturing Technology, 61*, 793–814.

Enderwick, P. (2005). *Globalization and Labor*. NY: Chelsea House Publications.

English, L.P. (1999*). Improving Data Warehouse and Business Information Quality*. New York: Wiley.

Eversheim, W. (1997). *Prozeßorientiertes Qualitätscontrolling*. Berlin: Springer.

Evgeniou, T., & Pontil, M. (2001). *Support Vector Machines: Theory and Applications*. In G. Paliouras, V. Karkaletsis, & C. Spyropoulos (Eds.), ACAI ' 99, LNAI 2049 (pp. 249–257). Berlin Heidelberg: Springer.

Farquad, M. a. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems, 53*(1), 226–233. doi:10.1016/j.dss.2012.01.016

Fasoli, T., Terzi, S., Jantunen, E., Kortelainen, J., Sääski, J. & Salonen, T. (2011). Challenges in Data Management in Product Life Cycle Engineering, 525-530. *In: Hesselbach, J. & Herrmann, C. (2011). Glocalized Solutions for Sustainability in Manufacturing*. Heidelberg, Berlin: Springer Verlag.

Fasser, Y., & Brettner, D. (2002). *Management for quality in high-technology enterprises*. New York: Wiley-Interscience.

Feelders, A. (1999). Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation ? In Zytkow, J. & Rauch, J. (Eds.), *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science Volume 1704* (pp. 329–334). Berlin Heidelberg: Springer. doi:10.1007/978-3-540-48247-5_38

Filipic, B., & Junkar, M. (2000). Using inductive machine learning to support decision making in machining processes. *Computers in Industry, 43*, 31–41.

Filos, E. (2013). Manufacturing Innovation and Horizon 2020. *In Kovacs, G.L. & Kochan, D. (Eds.): NEW PROLAMAT 2013, IFIP AICT 411 IFIP International Federation for Information Processing* (2013), October, 2013, Dresden, Germany.

Fink, A., Schneidereit, G., & Voß, S. (2005). *Grundlagen der Wirtschaftsinformatik*, 2 überarbeitete Auflage. Heidelberg: Physica-Verlag.

Forza, C. & Filippini, R. (1998). TQM impact on quality conformance and customer satisfaction: A causal model. *International Journal of Production Economics,* (55), 1-20.

Fowler, J. W. (2004). Grand Challenges in Modeling and Simulation of Complex Manufacturing Systems. *Simulation, 80*(9), 469–476. doi:10.1177/0037549704044324

Frey, C. (2007). *Rohstoffe als Beitrag zur Portfoliooptimierung*. München: GRIN Verlag.

Fritz, H. & Schulze, G. (2006). *Fertigungstechnik*. 7. neu bearbeitete Auflage. Berlin, Heidelberg: Springer-Verlag.

Fung, G. M., & Mangasarian, O. L. (2004). A Feature Selection Newton Method for Support Vector Machine Classification. *Computational Optimization and Applications, 28*(2), 185–202. doi:10.1023/B:COAP.0000026884.66338.df

Fung, G. M., & Mangasarian, O. L. (2006). Breast Tumor Susceptibility to Chemotherapy via Support Vector Machines. *Computational Management Science, 3*, 103–112.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *BIOINFORMATICS, 16*(10), 906–914.

Ganesan, R., Das, T. K. & Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions, 36*(9), 787–806.

Gardner, R., & Bicker, J. (2000). Using machine learning to solve tough manufacturing problems. *International Journal of Industrial Engineering-Theory Applications and Practice, 7*(4). 359-364.

Garetti, M. & Terzi, S. (2004). Product lifecycle management: definition, trends and open issues. *In: Proceedings at the 3rd international conference on advances in production engineering*, 17–19 June 2004, Warsaw, Poland.

Garvin, D. A. (1984). What Does "Product Quality" Really Mean? *MIT Sloan Management Review, 26*(1).

Ge, Z., Gao, F. & Song, Z. (2011). Batch process monitoring based on support vector data description method. *Journal of Process Control, 21*(6), 949–959. doi:10.1016/j.jprocont.2011.02.004

Ge, Z., Song, Z. & Gao, F. (2013). Review of Recent Research on Data-Based Process Monitoring. *Industrial & Engineering Chemistry Research, 2013*(52), 3543–3562. doi:10.1021/ie302069q

Geller, W. (2006). *Thermodynamik für Maschinenbauer – Grundlagen für die Praxis.* Heidelberg: Springer.

Ghahramani, Z. & Jordan, M. I. (1994). *Learning from incomplete data* (pp. 1–11). Cambridge, USA.

Giffin, M., de Weck, O., Bounova, G., Keller, R., Eckert, C., & Clarkson, P. J. (2009). Change Propagation Analysis in Complex Technical Systems. *Journal of Mechanical Design, 131*(8), 1–14. doi:10.1115/1.3149847

Gimenez, D.M., Vegetti, M., Leone, H.P. & Henning, G. (2008). Product Ontology: defining product-related concepts for logistics planning activities. *Computers in Industry, 2/3*(59), 232-240.

Gogolla, M. & Parisi-Presicce, F. (1998). State diagrams in UML: A formal semantics using graph transformations. *In M. Broy, D. Coleman, T. S. E. Mai-*

baum, and B. Rumpe, editors, Proceedings PSMT'98 Workshop on Precise Semantics for Modeling Techniques. Technische Universität München, TUM-I9803.

Golovatchev, J. D., & Budde, O. (2007). A holistic Product Lifecycle Management framework facing the challenges of 21st century. *In 3rd International Conference on Lifecycle Management*, Zurich (pp. 577–590).

Gordon, J. & Sohal, A.S. (2001). Assessing manufacturing plant competitiveness. *International Journal of Operations & Production Management, 21*(1/2). 233-253.

Goseva-Popstojanova, K., Wang, F., Wang, R., Gong, F., Vaidyanathan, K., Trivedi, K., & Muthusamy, B. (2001). Characterizing intrusion tolerant systems using a state transition model. *In Proceedings of the DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01*. (Vol. 2, pp. 211–221).

Graf, A., & Borer, S. (2001). Normalization in support vector machines. In *B. Radig & S. Florczyk (Eds.), Pattern Recognition. Lecture Notes in Computer Science Volume 2191 (pp. 277–282)*. Munich. doi:10.1007/3-540-45404-7_37

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual review of psychology, 60*, 549–76. doi: 10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing Data - Analysis and Design* (p. 322). NY: Springer.

Grote, K.-H. & Feldhusen, J. (2007). *Dubbel Taschenbuch für den Maschinenbau*, Zweiundzwanzigste, neubearbeitete und erweiterte Auflage. Berlin Heidelberg New York: Springer-Verlag.

Grzymala-Busse, J. W., Grzymala-Busse, W. J., Hippe, Z. S., & Rzasa, W. (2007). A Comparison of Three Approximation Strategies for Incomplete Data Sets. *2007 IEEE International Conference on Granular Computing (GRC 2007),* 301–301. doi:10.1109/GrC.2007.119

Gunasekaran, A., & Ngai, E. W. (2004). Information systems in supply chain integration and management. *European Journal of Operational Research, 159*(2), 269–295. doi:10.1016/j.ejor.2003.08.016

Guo, X., Sun, L., Li, G., & Wang, S. (2008). A hybrid wavelet analysis and support vector machines in forecasting development of manufacturing. *Expert Systems with Applications, 35*(1-2), 415–422. doi:10.1016/j.eswa.2007.07.052

Gutenberg, E. (1970). *Grundlagen der Betriebswirtschaftslehre, Band 1: Die Produktion.* 24. Auflage. Springer. Berlin, Heidelberg.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). A gene selection method for cancer classification using Support Vector Machines. *Machine Learning, 46*, 389–422. doi:10.1155/2012/586246

Haegeman, J., Cirac, J. I., Osborne, T. J., & Verstraete, F. (2012). *Calculus of continuous matrix product states*. eprint arXiv:1211.3935. Retrieved from arXiv:1211.3935

Hamel, C. K., & Prahalad, G. (1990). The Core Competence of the Corporation. *Harvard Business Review, May–June*, 275–292.

Hamel, L. (2009). *Knowledge discovery with support vector machines*. Hoboken, New Jersey: JOHN WILEY & SONS, INC.

Harding, J. A., Shahbaz, M., Srinivas & Kusiak, A. (2006). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering, 128*(4), 969. doi:10.1115/1.2194554

Hathaway, R. J. & Bezdek, J. C. (2002). Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. *Pattern Recognition Letters, 23*(1-3), 151–160. doi:10.1016/S0167-8655(01)00115-5

Hatvany, J. & Nemes, L. (1978). Intelligent manufacturing systems - a tentative forecast, *In: A link between science and applications of automatic control, Proc. of the VIIth IFAC World Congress*, Helsinki, Finland, 2, 895-899.

Hatvany, J. (1983). The efficient use of deficient information, *CIRP Annals, 32*(1), 423-425.

Haun, M. (2002). Handbook Knowledge Management. Basics and Realization, Systems and Praxis Examples. (original German title: Handbuch Wissensmanagement. Grundlagen und Umsetzung, Systeme und Praxisbeispiele) Berlin: Springer Verlag.

He, N., Zhang, D. Z., & Li, Q. (2013). Agent-based hierarchical production planning and scheduling in make-to-order manufacturing system. *International Journal of Production Economics*. doi:10.1016/j.ijpe.2013.08.022

Heinecke, G., Lamparter, S. & Kunz, A. (2011). Process transparency: Effects of a structured read point selection. In: *Proceedings of the 21st International Conference on Production Research, Innovation in Product and Production*. July 31 - August 4, 2011, Stuttgart, Germany.

Heinrich, L. J., Heinzl, A. & Roithmayr F. (2007). *Wirtschaftsinformatik: Einführung und Grundlegung*. München: Oldenbourg Verlag.

Helfert, M. (2002). *Planung und Messung der Datenqualität in Data-Warehouse-Systemen*. Dissertation. Universität St. Gallen.

Helo, P., Suorsa, M., Hao, Y., & Anussornnitisarn, P. (2014). Toward a cloud-based manufacturing execution system for distributed manufacturing. *Computers in Industry, 65*(4), 646–656. doi:10.1016/j.compind.2014.01.015

Herbrich, R., & Graepel, T. (2001). A PAC-Bayesian Margin Bound for Linear Classifiers : Why SVMs work. In *T. K. Leen, T. G. Dietterich, & V. Tresp*

(Eds.), Advances in Neural Information Processing Systems 13 (pp. 224–230). Cambridge, MA: MIT Press.

Hicks, B. J. (2007). Lean information management: Understanding and eliminating waste. *International Journal of Information Management, 27*(4), 233–249. doi:10.1016/j.ijinfomgt.2006.12.001

Hicks, B. J., Culley, S. J., & McMahon, C. a. (2006). A study of issues relating to information management across engineering SMEs. *International Journal of Information Management, 26*(4), 267–289. doi: 10.1016/j.ijinfomgt.2006.03.006

Höpf, M., & Schaeffer, C. F. (1997). *Holonic Manufacturing Systems*. In J. Goossenaerts, F. Kimura, & H. Wortmann (Eds.), Information Infrastructure Systems for Manufacturing IFIP — The International Federation for Information Processing (pp. 431–438). Springer. doi:10.1007/978-0-387-35063-9_37

Hoffmann, F., Keßler, O., Lübben, Th. & Mayr, P. (2002). Distortion Engineering – Verzugsbeherrschung in der Fertigung, *HTM 57*(3), 213-217.

Hoffmann, M., Goesmann, T., & Kienle, A. (2002). *Analyse und Unterstützung von Wissensprozessen als Voraussetzung für erfolgreiches Wissensmanagement.* In A. Abecker, K. Hinkelmann, H. Maus, & H. J. Müller (Eds.), Geschäftsprozessorientiertes Wissensmanagement (pp. 159–181). Berlin: Springer.

Hoke, G. E. J. (2011). SHORING UP Information Governance with GARP. *Information Management Journal, 45*(1), 26–31.

Hoyer, R. (1988). *Organisatorische Voraussetzungen der Büroautomatisation: Rechnergestützte, prozessorientierte Planung von Büroinformations- und Kommunikationssystemen.* Berlin: Erich Schmidt Verlag.

Hoyer, R.W., & Hoyer, B.B.Y. (2001). What is Quality. *Quality Progress,* (34), 53-62.

Hribernik, K. A., Pille, C., Jeken, O., Thoben, K.-D., Windt, K. & Busse, M. (2010). Autonomous Control of Intelligent Products in Beginning of Life Processes. *In International Conference on Product Lifecycle Management.*

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems, 37*(4), 543–558. doi:10.1016/S0167-9236(03)00086-1

Huang, L., Zhang, H. H., Zeng, Z.-B., & Bushel, P. R. (2013). Improved Sparse Multi-Class SVM and Its Application for Gene Selection in Cancer Classification. *Cancer Informatics, 12*, 143–53. doi:10.4137/CIN.S10212

Hussain, M. A. (1999). Review of the applications of neural networks in chemical process control — Simulation and online implementation. *Artificial Intelligence in Engineering, 13*(1), 55–68

Hüttig, G.F. (1943). Zur Systematik der Aggregatzustände. *Colloid & Polymer Science 104*(2-3), 161-167.

Hutton, J., & Denham, J. (2008). *Manufacturing: New Challenges, New Opportunities* (p. 64). London.

Instone, F. & Dale, B. (1989). A case study of the typical issues involved in quality improvement. *International Journal of Operations & Production Management, 9*(1), 15–26.

Jacob, J., & Petrick, K. (2007). *Qualitätsmanagement und Normung,* 101–121. In R. Schmitt & Pfeifer, T. (Eds.), Masing Handbuch Qualitätsmanagement. München: Carl Hanser Verlag.

Jansen-Vullers, M.H., van Drop, C.A. & Beulens, A.J.M. (2003). Managing traceability information in manufacture. *International Journal of Information Management, 23*(2003), 395-413.

Jarke, M. & Jeusfeld, M., Quix, C. & Vassilidis, P. (1999). Architecture and Quality in Data Ware- houses: An Extended Repository Approach. *Information Systems, 3*(24), 229–253.

Jehle, E. (1999). *Produktionswirtschaft*. Heidelberg: Verlag Recht und Wirtschaft.

Jenab, K., & Ahi, P. (2010). Fuzzy quality feature monitoring model. *International Journal of Production Research, 48*(17), 5021–5030. doi: 10.1080/00207540903117907

Jensen, D. (2007). *Proximity 4.3 Tutorial* (p. 174).

Jiang, P., Jia, F., Wang, Y., & Zheng, M. (2012). Real-time quality monitoring and predicting model based on error propagation networks for multistage machining processes. *Journal of Intelligent Manufacturing*, (2012). doi:10.1007/s10845-012-0703-0

Jun, H.-B., Kiritsis, D., & Xirouchakis, P. (2007). Research issues on closed-loop PLM. *Computers in Industry, 58*(8-9), 855–868. doi:10.1016/j.compind.2007.04.001

Kabacoff, R. I. (2011). *Advanced Methods for missing data*. In R. I. Kabacoff (Ed.), R IN ACTION: Data analysis and graphics with R (pp. 352–371). Shelter Island: Manning Publications Co.

Kaiser, M. J. (1998). Generalized zone separation functionals for convex perfect forms and incomplete data sets. *International Journal of Machine Tools and Manufacture, 38*(4), 375–404. doi:10.1016/S0890-6955(97)00042-4

Kalpakjian, S. & Schmid, S.R. (2009). *Manufacturing engineering and technology*. New Jersey: Prentice Hall.

Kamiske, G., & Brauer, J. (2008). *Qualitätsmanagements von A bis Z*. 6., Auflage. München: Carl Hanser-Verlag.

Kang, B. S., Choe, D. H., & Park, S. C. (1999). Intelligent process control in manufacturing industry with sequential processes. *International Journal of Production Economics, 60-61*, 583–590. doi:10.1016/S0925-5273(98)00178-9

Kano, M. & Nakagawa, Y. (2008). Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers & Chemical Engineering, 32*(1-2), 12–24. doi:10.1016/j.compchemeng.2007.07.005

Kaynak, H. (2003). The relationship between total quality management practices and their effects on firm performance. *Journal of Operations Management,* (21), 405-435.

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation, 15*(7), 1667–89. doi:10.1162/089976603321891855

Keferstein, C. (2011). *Fertigungsmesstechnik praxisorientierte Grundlagen, moderne Messverfahren.* Wiesbaden: Vieweg+Teubner-Verlag.

Kent, J.T. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163-173. doi: 10.1093/biomet/70.1.163

Kern, J. (2008). *Ishikawa Diagramme - Ursache-Wirkungs-Diagramme*. München: GRIN Verlag.

Kerdprasop, K., & Kerdprasop, N. (2011). A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process. *International Journal of Mechanics, 5*(4), 336–344.

Khemchandani, R., & Chandra, S. (2009). Knowledge based proximal support vector machines. *European Journal of Operational Research, 195*(3), 914–923. doi:10.1016/j.ejor.2007.11.023

Kim, D., Kang, P., Cho, S., Lee, H., & Doh, S. (2012). Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. *Expert Systems with Applications, 39*(4), 4075–4083. doi:10.1016/j.eswa.2011.09.088

Kimemia, J. G., & Gershwin, S. B. (1981). *An algorithm for the computer control of production in a flexible manufacturing system*. In Decision and Control including the Symposium on Adaptive Processes, 1981 20th IEEE Conference on (Volume:20) (pp. 628 – 633). doi:10.1109/CDC.1981.269285

Kiritsis, D., Bufardi, A., & Xirouchakis, P. (2003). Research issues on product lifecycle management and information tracking using smart embedded systems. *Advanced Engineering Informatics, 17*(3-4), 189–202. doi:10.1016/j.aei.2004.09.005

Klein, D., Thoben, K.-D. & Nowak, L. (2005). *Using Indicators to Describe Distortion Along a Process Chain.* In Zoch, H.-W. ; Lübben, Th. (Eds.): Proc. 1st Int. Conf. on Distortion Engineering, 14-16.09.2005 in Bremen, Germany, 31-36.

Knoke, B., Wuest, T. & Thoben, K.-D. (2012). Understanding Product State Relations within Manufacturing Processes. *In C. Emmanouilidis, M. Taisch, & D. Kiritsis (Eds.), International Conference of Advances in Production Management Systems (APMS 2012) - Competitive Manufacturing for Innovative Products and Services.* Berlin Heidelberg: Springer.

Kobler, M. (2010). *Qualität von Prozessmodellen: Kennzahlen zur analytischen Qualitätssicherung bei der Prozessmodellierung.* Berlin: Logos Verlag.

Köksal, G., Batmaz, İ. & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications, 38*(10), 13448–13467. doi:10.1016/j.eswa.2011.04.063

König, W. & Klocke, F. (2008). *Fertigungsverfahren Drehen, Fräsen, Bohren* 8., neu bearbeitete Auflage. Berlin-Heidelberg: Springer-Verlag.

Koether, R. & Rau, W. (2008). *Fertigungstechnik für Wirtschaftsingenieure.* München: Carl Hanser-Verlag.

Kopacek, P. (1999). Intelligent Manufacturing: Present State and Future Trends. *Journal of Intelligent and Robotic Systems, 26*, 217–229.

Koren, Y., Hu, S. J., & Weber, T. W. (1998). Impact of Manufacturing System Configuration on Performance. *Annals of the CIRP, 47*(1), 369–372.

Korndörfer, W. (2003). *Allgemeine Betriebswirtschaftslehre.* 12. Auflage. Gabler. Wiesbaden.

Kotler, P., Armstrong, G., Saunders, J. & Wong, V. (2011). *Grundlagen des Marketing*. 5. Auflage. Pearson Studium. München.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica, 31*, 249–268.

Kovačič, M., & Šarler, B. (2009). Application of the Genetic Programming for Increasing the Soft Annealing Productivity in Steel Industry. *Materials and Manufacturing Processes, 24*(3), 369–374. doi:10.1080/10426910802679634

Krallmann, H., Schönherr, M. & Trier, M. (2007). *Systemanalyse im Unternehmen: Prozessorientierte Methoden der Wirtschaftsinformatik.* München, Wien: Oldenbourg Wissenschaftsverlag.

Krcmar, H. (2005). *Informationsmanagement.* Berlin Heidelberg New York: Springer Verlag.

Kreutzberg, J. (2000). *Qualitätsmanagement auf dem Prüfstand, Analyse des Qualitätsmanagements von Informationssystemen.* Dissertation. University of Zurich, Zurich, Switzerland.

Küll, U. (2013). *Im Rausch der Geschwindigkeit: Big Data und Echtzeitanalysen revolutionieren Business Intelligence*. Retrieved May 22, 2013 from: http://www.heise.de/microsites/bigdata-grosse-datenmengen-beherrschen-und-analysieren/big-data-und-echtzeitanalysenrevolutionieren-business-intelligence/150/379/1142/1.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. doi:10.1007/978-1-4614-6849-3

Kumar, S. (2002). *Intelligent manufacturing systems* (pp. 1–20). Ranchi. Retrieved from http://pchats.tripod.com/int_manu.pdf

Kwak, D.-S. & Kim, K.-J. (2012). A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Systems with Applications, 39*(3), 2590–2596. doi:10.1016/j.eswa.2011.08.114

Laili, Y., Tao, F., Zhang, L. & Ren, L. (2011). The optimal allocation model of computing resources in cloud manufacturing system. *Seventh International Conference on Natural Computation*, 2322–2326. doi:10.1109/ICNC.2011.6022564

Lang, S. (2007). *Durchgängige Mitarbeiterinformation zur Steigerung von Effizienz und Prozesssicherheit in der Produktion*. Dissertation. Universität Erlangen-Nürnberg. Bamberg: Meisenbach Verlag.

Lange, K.M. (2007). *Duden Wirtschaft von A bis Z: Grundlagenwissen für Schule und Studium, Beruf und Alltag* (Gebundene Ausgabe). Mannheim: Bibliographisches Institut.

Larose, D. (2005). *Discovering Knowledge in Data - An Introduction to Data Mining*. Hoboken: Wiley.

Lee, J., & Ha, S. (2009). Recognizing yield patterns through hybrid applications of machine learning techniques. *Information Sciences, 179*(6), 844–850. doi:10.1016/j.ins.2008.11.008

Lee, J.-M., Yoo, C. & Lee, I.-B. (2004). Statistical process monitoring with independent component analysis. *Journal of Process Control, 14*(5), 467–485. doi:10.1016/j.jprocont.2003.09.004

Leong, K.K., Yu, K.M. & Lee, W.B. (2002). Product data allocation for distributed product data management systems. *Computers in Industry, 47*(2002), 289-298.

Lessmann, S., Sung, M.-C., & Johnson, J. E. V. (2009). Identifying winners of competitive events: A SVM-based classification model for horserace prediction. *European Journal of Operational Research, 196*(2), 569–577. doi:10.1016/j.ejor.2008.03.018

Levitt, T. (1993). The globalization of markets. *Harvard Business Review, May-June,* pp. 92–102.

Lewis, F. L., Horne, B. G. & Abdallah, C. T. (1996). *On the computational complexity of the manufacturing job shop and reentrant flow time* (pp. 1–27). Retrieved July 10th, 2013 from http://ise.unm.edu/controls/papers/on-the-computational-complexity.pdf

Li, B., Hu, J., & Hirasawa, K. (2008). Support Vector Machine Classifier with WHM Offset for Unbalanced Data. *Journal of Advanced Computational Intelligence and Intelligent Infomatics, 12*(1), 94–101.

Li, T.-S. & Huang, C.-L. (2009). Defect spatial pattern recognition using a hybrid SOM–SVM approach in semiconductor manufacturing. *Expert Systems with Applications, 36*(1), 374–385. doi:10.1016/j.eswa.2007.09.023

Li, H., Liang, Y., & Xu, Q. (2009). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, *95*(2), 188–198. doi:10.1016/j.chemolab.2008.10.007

Li, Y., & Shawe-Taylor, J. (2003). The SVM With Uneven Margins And Chinese Document Categorisation. In *17th Pacific Asia Conference on Language Information and Computation (PACLIC17*), October (pp. 216–227).

Linß, G. (2002). *Qualitätsmanagement für Ingenieure*. München/Wien: Hanser Verlag.

Liu, T.D. & Xu, W. (2001). A review of web-based product data management systems. *Computers in Industry, 44*(2001), 251-262.

Liu, X., Zhang, W. J., & Venuvinod, P. K. (1997). Intelligent Manufacturing Systems in Global Manufacturing Paradigm: a critical review and new research issues. *In CIRP International Symposium - Advanced Design and Manufacture in the Global Manufacturing Era*, August 21-22, 1997, Hong Kong.

Lödding, H. (2013). *Handbook of Manufacturing Control - Fundamentals, description, configuration*. Heidelberg New York: Springer.

Löhr-Richter, P. (1993). Zur Diskussion: Methodologie - Methodik - Methode. Was steckt dahinter? *Gesellschaft für Informatik e.V. - Lecture Notes in Informatics, 1*(1993), 39-41.

Lohr, S. (2012). The Age of Big Data. *New York Times, February 11*, 2012.

Lu, S.C.-Y. (1990). Machine learning approaches to knowledge synthesis and integration tasks for advanced engineering automation. *Computers in Industry, 15*(1990), 105-120.

Lu, S. C.-Y., & Suh, N.-P. (2009). Complexity in design of technical systems. *CIRP Annals - Manufacturing Technology, 58*(1), 157–160. doi:10.1016/j.cirp.2009.03.067

Lundin, E., Kvarnström, H., & Jonsson, E. (2002). *A Synthetic Fraud Data Generation Methodology*. In R. Deng, F. Bao, Z. Jianying, & Q. Sihan (Eds.), Information and Communications Security, Lecture Notes in Computer Science, Vol. 2513 (pp. 265–277). Berlin Heidelberg: Springer.

Lutz, M., Boucher, X., & Roustant, O. (2012). Information Technologies capacity planning in manufacturing systems: Proposition for a modelling process and application in the semiconductor industry. *Computers in Industry, 63*(7), 659–668. doi:10.1016/j.compind.2012.03.003

Maddern, H., Smart, P. A., Maull, R. S., & Childe, S. (2013). End-to-end process management: implications for theory and practice. *Production Planning & Control,* (November 2013), 1–19. doi:10.1080/09537287.2013.832821

Magee, C. L., & de Weck, O. L. (2004). Complex System Classification. In *Fourteenth Annual International Symposium of the International Council on Systems Engineering (INCOSE)*, Toulouse, France, June 20-24, 2004.

Mangasarian, O. L., & Wild, E. W. (2007). Feature Selection for Nonlinear Kernel Support Vector Machines. *In Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007) (pp. 231–236).* Ieee. doi:10.1109/ICDMW.2007.30

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge, UK: Cambridge University Press.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476. doi:10.1016/j.ejor.2006.04.051

Masing, J. (2007). *Handbuch Qualität: Grundlagen und Elemente des Qualitätsmanagement: Systeme-Perspektiven*. München: Carl Hanser Verlag.

Maull, H.W. (1988). *Strategische Rohstoffe – Risiken für die Wirtschaftliche Sicherheit des Westens*. München: Oldenbourg Verlag.

May, G. S., & Spanos, C. J. (2006). *FUNDAMENTALS OF SEMICONDUCTOR MANUFACTURING AND PROCESS CONTROL* (p. 480). Hoboken, New Jersey: Wiley-Interscience.

Mayer-Bachmann, R. (2007). *Integratives Anforderungsmanagement – Konzept und Anforderungsmodell am Beispiel der Fahrzeugentwicklung*. Dissertation Universität Karlsruhe (TH). Karlsruhe: Universitätsverlag Karlsruhe.

Mazumder, J. (2008). Intelligent manufacturing: Role of lasers and optics. *LIA today*, p. 6.

McCann, M. & Johnston, A. (2008). SECOM data set. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. Irvine, CA: University of California, School of Information and Computer Science.

McCann, M., Li, Y., Maquire, L., & Johnston, A. (2010). Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 6*, 277–288.

McFarlane, D. C., & Bussmann, S. (2003). Holonic Manufacturing Control: Rationales , Developments and Open Issues. *In S. M. Deen (Ed.), Agent-Based Manufacturing, Advances in the Holonic Approach (303–326)*. Springer Berlin Heidelberg. doi:10.1007/978-3-662-05624-0_13

McFarlane, D., Sarma, S., Chirn, J. L., Wong, C. ., & Ashton, K. (2003). Auto ID systems and intelligent manufacturing control. *Engineering Applications of Artificial Intelligence, 16*(4), 365–376. doi:10.1016/S0952-1976(03)00077-0

Mei, D. C., Xie, C. W., & Zhang, L. (2004). The stationary properties and the state transition of the tumor cell growth mode. *The European Physical Journal B, 41*(1), 107–112. doi:10.1140/epjb/e2004-00300-1

Meffert, H., Burmann, C., & Kirchgeorg, M. (2008). *Marketing - Grundlagen marktorientierter Unternehmensführung* (p. 924). Gabler.

Mekid, S., Pruschek, P., & Hernandez, J. (2009). Beyond intelligent manufacturing: new generation of flexible intelligent NC machines. *Mechanism and Machine Theory, 44*(1), 466-476.

Merali, Y., & Bennet, Z. (2011).*Web 2.0 and Network Intelligence*. In P.Warren, J. Davies,&E. Simperl (Eds.), Context and semantics for knowledge management (pp. 11–26). Heidelberg: Springer.

Mertins, K., & Seidel, H. (2009). *Wissensmanagement im Mittelstand*. Grundlagen - Lšsungen - Praxisbeispiele. Berlin: Springer Verlag.

Mönch, L., Zimmermann, J., & Otto, P. (2006). Machine learning techniques for scheduling jobs with incompatible families and unequal ready times on parallel batch machines. *Engineering Applications of Artificial Intelligence, 19*(3), 235–245. doi:10.1016/j.engappai.2005.10.001

Mohanty, P. P. (2004). An agent-oriented approach to resolve the production planning complexities for a modern steel manufacturing system. *The International Journal of Advanced Manufacturing Technology, 24*(3-4), 199–205. doi:10.1007/s00170-003-1673-3

Monostori, L. (2002). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *In 15th Triennial World Congress*, Barcelona, Spain (p. 12).)

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering Applications of Artificial Intelligence, 16*(4), 277–291. doi:10.1016/S0952-1976(03)00078-2

Monostori, L., Hornyák, J., Egresits, C. & Viharos, Z. J. (1998). Soft computing and hybrid AI approaches to intelligent manufacturing. *In Tasks and Methods in Applied Artificial Intelligence Lecture Notes in Computer Science Volume 1416* (pp. 765–774). doi:10.1007/3-540-64574-8_463

Monostori, L., Márkus, A., Van Brussel, H. & Westkämper, E. (1996). Machine learning approaches to manufacturing, *CIRP Annals, 45*(2), 675-712.

Monostori, L., Váncza, J., & Kumara, S. R. T. (2006). Agent-Based Systems for Manufacturing. *CIRP Annals - Manufacturing Technology, 55*(2), 697–720.

Montgomery, D. (2005). *Introduction to statistical quality control.* Hoboken, NJ: John Wiley.

Moorthy, A., & Vivekanand, S. (2007). Integration of PLM with other concepts for empowering business environments. *In M. Garetti, S. Terzi, P. D. Ball, & S. Han (Eds.), Product Lifecycle Management Special Publication 3* (p. 93). Bergamo/Milano.

Morris, B., & Johnston, R. (1987). Dealing with Inherent Variability: The Difference Between Manufacturing and Service? *International Journal of Operations & Production Management, 7*(4), 13–22. doi:10.1108/eb054796

Musa, A., Gunasekaran, A., & Yusuf, Y. (2013). Supply chain product visibility: Methods, systems and impacts. *Expert Systems with Applications*. doi:10.1016/j.eswa.2013.07.020

Nagy, D., Jering, D., Strasser, T., Martel, A., Garello, P., Filios, E. (Eds.). (2005). *Intelligent Manufacturing Systems - Impact Report* (p. 48). Washington, D.C. Retrieved from ftp://ftp.cordis.europa.eu/pub/ims/docs/ims_impact_report_final.pdf

Naumann, F. (2007). Datenqualität. *Informatik-Spektrum, 30*(1), 27–31. doi:10.1007/s00287-006-0125-5

Nearchou, A. C. (2011). Maximizing production rate and workload smoothing in assembly lines using particle swarm optimization. *International Journal of Production Economics, 129*(2), 242–250. doi:10.1016/j.ijpe.2010.10.016

Nebl, T. (2007). *Produktionswirtschaft*. 6. Auflage. München: Oldenbourg Verlag.

Negnevitsky, M. (2005). *Artificial Intelligence: A Guide to Intelligent Systems*. Essex, UK: Addison Wesley.

Nilsson, N. J. (2005). *Introduction to machine learning* (p. 188). Stanford, USA.

N.N. (2006). The problem with solid engineering. *The Economist, 379*(8478), 71–73.

Nonaka, I., & Takeuchi, H. (1997). *Die Organisation des Wissens - Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen* (p. 299). Frankfurt/New York: Campus Verlag.

Nonnemaker, J., & Baird, H. S. (2009). Using Synthetic Data Safely in Classification. In *IS&T/SPIE Con. on Document Recognition and Retrieval (DRR 2009),* San Jose, CA, January 28 - February 1.

North, K., & Güldenberg, S. (2008). *Produktive Wissensarbeit(er) – Antworten auf die Managementherausforderungen des 21. Jahrhunderts*. Wiesbaden: Gabler Verlag.

Olbertz, J.-H. & Otto, H.-U. (Eds.). (2001). *Qualität von Bildung. Vier Perspektiven (Arbeitsberichte 2'01).* Hrsg. von HoF Wittenberg - Institut für Hochschulforschung an der Martin-Luther Universität Halle-Wittenberg. Wittenberg, (127). ISBN 3-9806701-4-7. ISSN 1436-3550.

OMG. (2010). *Business Process Model and Notation (BPMN).* Object Management Group. Retrieved October 4h, 2012, from http://www.omg.org/spec/BPMN/2.0/PDF

Oztemel, E. (2010). *Intelligent Manufacturing Systems*. In L. Benyoucef & B. Grabot (Eds.), Artificial Intelligence Techniques for Networked Manufacturing Enterprises Management (pp. 1–39). London Dordrecht Heidelberg New York: Springer. doi:10.1007/978-1-84996-119-6

Paul, G., & Paul, R. (2008). Engineering Data Management and Product Data Management : Roles and Prospects. *In International Scientific Conference Computer Science*, 614–619.

Pearl, J. (2003). CAUSALITY: MODELS, REASONING, AND INFERENCE. *Econometric Theory, 19*, 675–685. DOI: 10+10170S0266466603004109

Peltonen, H., Pitkänen, O., & Sulonen, R. (1996). Process-based view of product data management. *Computers in Industry, 31*(3), 195–203. doi:10.1016/S0166-3615(96)00051-6

Peng, Y. (2004). Intelligent condition monitoring using fuzzy inductive learning. *Journal of Intelligent Manufacturing, 2004*(15), 373–380.

Pham, D. T. & Afify, A. A. (2005). Machine-learning techniques and their applications in manufacturing. Proceedings of the Institution of Mechanical Engineers, Part B: *Journal of Engineering Manufacture, 219*(5), 395–412. doi:10.1243/095440505X32274

Piddington, C. & Pegram, M. (1993). An IMS test case – global manufacturing. *In: Proceedings of the IFIP TC5/WG5.7 5th Int. Conf. on Advances in Production management Systems*, 28-30 September, 1993, Athens, Greece.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211. doi:10.1145/505248.506010

Pokharel, S., & Mutha, A. (2009). Perspectives in reverse logistics: A review. *Resources, Conservation and Recycling, 53*(4), 175–182. doi:10.1016/j.resconrec.2008.11.006

Polanyi, M. (1962). *Personal Knowledge. Towards a Post-Critical Philosophy*. Routledge & Kegan Paul Ltd., London

Ponsignon, T. & Mönch, L. (2012). Heuristic approaches for master planning in semiconductor manufacturing. *Computers & Operations Research, 39*(3), 479–491. doi:10.1016/j.cor.2011.06.009

Porter, M. (1998). *Competitive advantage: creating and sustaining superior performance: with a new introduction*. New York: The Free Press.

Porter, M. E. (2008). *On Competition*. Boston: Harvard Business School Publishing.

Priore, P., De La Fuente, D., Gomez, A., & Puente, J. (2001). A review of machine learning in dynamic scheduling of flexible manufacturing systems. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 15*(3), 251–263. doi:10.1017/S0890060401153059

Priore, P., de la Fuente, D., Puente, J., & Parreño, J. (2006). A comparison of machine-learning algorithms for dynamic scheduling of flexible manufacturing systems. *Engineering Applications of Artificial Intelligence, 19*(3), 247–255. doi:10.1016/j.engappai.2005.09.009

Probst, G. J. B., Raub, S., & Romhardt, K. (2006). *Wissen managen* (p. 307). Wiesbaden: Gabler. doi:10.1007/978-3-8349-9343-4

Provost, S. B. (1990). Estimators for the parameters of a multivariate normal random vector with incomplete data on two subvectors and test of independence. *Computational Statistics & Data Analysis, 9*(1), 37–46. doi:10.1016/0167-9473(90)90069-T

Provost, F. (2000). Machine Learning from Imbalanced Data Sets 101. In *AAAI'2000 workshop on imbalanced data sets*.

Puzzanghera, J. (2013). Manufacturing shows surprising strength. *Los Angeles Times, October 2*(2013), B4.

Qin, S.J., Cherry, G., Good, R., Wang, J., & Harrison, C.A. (2006). Semiconductor manufacturing process control and monitoring: A fab-wide framework. *Journal of Process Control, 16*(3), 179–191. doi:10.1016/j.jprocont.2005.06.002

Reiter, J. P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology, 30*, 235–242.

Reiter, J. P., & Raghunathan, T. E. (2007). The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association, 102*(480), 1462–1471. doi:10.1198/016214507000000932

Rejani, Y. I. A., & Selvi, S. T. (2009). Early detection of breast cancer using SVM classifier technique. *International Journal on Computer Science and Engineering, 1*(3), 127–130.

Reuter, M. (2007). *Methodik der Werkstoffauswahl, Der systematische Weg zum richtigen Material.* München: Carl Hanser-Verlag.

Revilla, J. & Cadena, M. (2008). Trends in intelligent manufacturing systems. *In: Proceedings of the World Congress on Engineering*, London, UK, 1257-1262.

Ribeiro, B. (2005). Support Vector Machines for Quality Monitoring in a Plastic Injection Molding Process. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 35*(3), 401–410. doi:10.1109/TSMCC.2004.843228

Robinson, C.J. & Malhotra, M.K. (2005). Defining the concept of supply chain quality management and its relevance to academic and industrial practice. *International Journal of Production Economics,* (96), 315-337.

Robles, N. & Roy, U. (2004). Optimal tolerance allocation and process-sequence selection incorporating manufacturing capacities and quality issues. *Journal of Manufacturing Systems, 23*(2), 127–133. doi:10.1016/S0278-6125(05)00002-6

Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A. & Schmid, J. (2011*). Informationsqualität - Definitionen, Dimensionen und Begriffe*. In K. Hildebrand, M. Gebauer, H. Hinrichs, & M. Mielke (Eds.), Daten- und Informationsqualität. (pp. 25–45). Wiesbaden: Vieweg+Teubner. doi:10.1007/978-3-8348-9953-8

Rosemann, M. & Schütte, R. (1997). *Grundsätze ordnungsmäßiger Referenzmodellierung.* In: Becker, J.; Rosemann, M.; Schütte, R. (Hrsg.): Entwicklungsstand und Entwicklungsperspektiven der Referenzmodellierung. Arbeitsberichte des Instituts für Wirtschaftsinformatik, 16–33.

Saaksvuori, A. & Immonen, A. (2004). *Product Lifecycle Management*. Berlin, Heidelberg, NY: Springer.

Salahshoor, K., Kordestani, M., & Khoshro, M. S. (2010). Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers. *Energy, 35*(12), 5472–5482. doi:10.1016/j.energy.2010.06.001

Samson, D. & Terziovski M. (1999). The relationship between total quality management practices and operational performance. *Journal of Operations Management,* (17), 393-409.

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal 3*(3), 210–229.

Sánchez A. V. D. (2003). Advanced support vector machines and kernel methods. *Neurocomputing, 55*(1-2), 5–20. doi:10.1016/S0925-2312(03)00373-4.

Sarich, M., Schutte, C. & Vanden-Eijden, E. (2010). Optimal Fuzzy Aggregation of Networks. *Multiscale Modeling and Simulation, 8*(4), 1535-1561.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. doi:10.1037//1082-989X.7.2.147

Schatt, W. & Worch, H. (2003). *Werkstoffwissenschaft 9., Auflage.* Weinheim: Wiley-VCH-Verlag.

Scheidat, T., Leich, M., Alexander, M., & Vielhauer, C. (2009). Support Vector Machines for Dynamic Biometric Handwriting Classification. *In Proceedings of AIAI Workshops* (pp. 118–125).

Schiersch, A. (2009). *Inefficiency in the German Mechanical Engineering Sector (pp. 1–29).* Berlin. Retrieved from http://ssrn.com/abstract=1514278

Schmachtenberg, E. (2000). *Vom Material zum Produkt – Der Prozess der Werkstoffop- timierung*. In: Essener Unikate 13/2000 S.112-121. Heine. Essen.

Schöning, U. (2001). *Theoretische Informatik – kurzgefasst.* Heidelberg; Berlin: Spektrum, Akademischer Verlag.

Seidel, W. W. & Hahn, F. (2010). *Werkstofftechnik 8., neu bearbeitete Auflage.* München: Carl Hanser-Verlag.

Seifert, M. (2007). *Unterstützung der Konsortialbildung in Virtuellen Organisationen durch prospektives Performance Measurement*. Dissertation. Universität Bremen. Bremen, Germany.

Sendler, U. (2009). *Das PLM-Kompendium. Referenzbuch des Produkt-Lebenszyklus- Managements.* Springer, Heidelberg.

Shen, W., Hao, Q., Yoon, H., & Norrie, D. (2006). Applications of agent-based systems in intelligent manufacturing: an update review. *Advanced Engineering Informatics, 20*(4), 415-431. http://dx.doi.org/10.1016/j.aei.2006.05.004

Shetwan, A. G., Vitanov, V. I. & Tjahjono, B. (2011). Allocation of quality control stations in multistage manufacturing systems. *Computers & Industrial Engineering, 60*(4), 473–484. doi:10.1016/j.cie.2010.12.022

Shiang, L.E. & Nagaraj, S., (2011). Impediments to innovation: evidence from Malaysian manufacturing firms. *Asia Pacific Business Review, 17*(2), 209–223.

Silva, R. G. (2009). Condition monitoring of the cutting process using a self-organizing spiking neural network map. Journal *of Intelligent Manufacturing, 21*(6), 823–829. doi:10.1007/s10845-009-0258-x

Simão, J. M., Stadzisz, P. C., & Morel, G. (2006). Manufacturing execution systems for customized production. *Journal of Materials Processing Technology, 179*(1-3), 268–275. doi:10.1016/j.jmatprotec.2006.03.064

Simon, H.A. (1983). *Why should machines learn?* In: Michalski, R., Carbonell, J. & Mitchell, T. (eds). Machine Learning: An Artificial Intelligence Approach. Charlotte: Tioga Press, 25-38

Sitek, P., Seifert, M., & Thoben, K.-D. (2010). Towards an inter-organisational perspective for managing quality in virtual organisations. *International Journal of Quality & Reliability Management, 27*(2), 231–246. doi:10.1108/02656711011014339

Sitek, P. (2012). *Quality management to support single companies in colloborative enterprise networks. Dissertation*. University of Bremen, Bremen, Germany.

Skitt, P. J. C., Javed, M. A., Sanders, S. A., & Higginson, A. M. (1993). Process monitoring using auto-associative, feed-forward artificial neural networks. *Journal of Intelligent Manufacturing, 1993*(4), 79–94.

Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to machine learning*. Cambridge, UK: Cambridge University Press.

Söhner, J. (2003). *Beitrag zur Simulation zerspanungstechnologischer Vorgänge mit Hilfe der Finite-Elemente-Methode.* Universität Karlsruhe: Dissertation

Sölter, J., & Brinksmeier, E. (2008). Modelling and simulation of ring deformation due to clamping. *In International Conference on Distortion Engineering. September 17-19, 2008. Bremen. Germany* (pp. 143–151).

Sölter, J. (2010). *Ursachen und Wirkmechanismen der Entstehung von Verzug infolge spanender Bearbeitung.* Dissertation Universität Bremen, 2010. Aachen: Shaker Verlag.

Sohal, A. S. & Terziovski, M. (2000). TQM in Australian manufacturing: factors critical to success. *International Journal of Quality & Reliability Management, 17*(2), 158–167.

Sonnenberg, F. a., & Beck, J. R. (1993). Markov Models in Medical Decision Making: A Practical Guide. *Medical Decision Making, 13*(4), 322–338. doi:10.1177/0272989X9301300409

Specht, D. & Braunisch, D. (2008). Sekundärrohstofflogistik – Konzepte und Anwendungen. *Zeitschrift für wirtschaftlichen Fabrikbetrieb, 103*(12), 875 - 879.

Spoerre, J., & Ben Wang, H.-P. (1995). Machine performance monitoring and fault classification using an exponentially weighted moving average scheme. *International Journal of Production Research, 33*(2), 445–463.

Srdoč, A., Bratko, I., & Sluga, A. (2007). Machine learning applied to quality management—A study in ship repair domain. *Computers in Industry, 58*(5), 464–473. doi:10.1016/j.compind.2006.09.013

Stark, J. (2011). *Product Lifecycle Management, 21st Century Paradigm for Product Realization*. Heidelberg: Springer Verlag. doi:10.1007/978-0-85729-546-0

Steffelbauer-Meuche, G. (2004). *Qualitätsmanagement in der Internen Revision,* 1. Aufl., Sternenfels: Wissenschaft und Praxis.

Sterman, J. D. (1992). *System Dynamics Modeling for Project Management*. Retrieved October 26th, 2013, from http://web.mit.edu/jsterman/www/SDG/project.pdf

Steven, M. (2007*). Handbuch Produktion, Theorie-Management-Logistik- Controlling*. Stuttgart: Kohlhammer-Verlag

Storey, V. C., Dewan, R. M., & Freimer, M. (2012). Data quality: Setting organizational policies. *Decision Support Systems, 54*(1), 434–442. doi:10.1016/j.dss.2012.06.004

Stoumbos, Z. & Sullivan, J. (2002). Robustness to Non-Normality of the Multivariate EWMA Control Chart. *Journal of Quality Technology, 34*(3), 260–276.

Stavropoulos, P., Chantzis, D., Doukas, C., Papacharalampopoulos, A. & Chryssolouris, G. (2013). Monitoring and Control of Manufacturing Processes: A Review. *Procedia CIRP, 8*, 421–425. doi:10.1016/j.procir.2013.06.127

Suh, N. P. (2005). Complexity in Engineering. *CIRP Annals - Manufacturing Technology, 54*(2), 46–63. doi:10.1016/S0007-8506(07)60019-5

Sukchotrat, T., Kim, S. B., & Tsung, F. (2009). One-class classification-based control charts for multivariate process monitoring. *IIE Transactions, 42*(2), 107–120. doi:10.1080/07408170903019150

Sun, J., Rahman, M., Wong, Y., & Hong, G. (2004). Multiclassification of tool wear with support vector machine by manufacturing loss consideration. *International Journal of Machine Tools and Manufacture, 44*(11), 1179–1187. doi:10.1016/j.ijmachtools.2004.04.003

Sundin, E. (2009*). Life-cycle perspectives of product/service-wSystems: in design theory*. In M. L. Sakao (Ed.), Introduction to Product/Service-system Design (pp. 31–49). London: Springer-Verlag.

Surm, H. (2011). Identifikation der verzugsbestimmenden Einflussgrößen beim Austenitisieren am Beispiel von Ringen aus dem Wälzlagerstahl 100Cr6. Dissertation. University of Bremen, 2011.

Surm, H. & Rath, J. (2012). Mechanismen der Verzugsentstehung bei Wälzlagerringen aus 100Cr6 (Distortion Mechanisms in the Process Chain Bearing Ring). *Journal of Heat Treatment Materials, 67*(5), 1–13.

Sutton, R. S., & Barto, A. G. (2012). *Reinforcement Learning: An Introduction (Second.)*. Cambridge, USA: The MIT Press.

Taguchi, G. (1989). *Introduction to quality engineering* (p. 263). New York: Kraus International Publications.

Taisch, M., Cammarino, B. P., & Cassina, J. (2011). Life cycle data management: first step towards a new product lifecycle management standard. International *Journal of Computer Integrated Manufacturing, 24*(12), 1117–1135. doi:10.1080/0951192X.2011.608719

Tang, Y., Zhang, Y.-Q., Chawla, N. V, & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society, 39*(1), 281–8. doi:10.1109/TSMCB.2008.2002909

Tay, F. E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing, 48*, 847–861.

Terzi, S., Panetto, H., Morel, G., & Garetti, M. (2007). A holonic metamodel for product traceabily in PLM. *International Journal of Product Lifecycle Management, 2*(3), 253–289. doi:10.1504/IJPLM.2007.016292

Thesing, G., Randow, J., Kirchfeld, A., Berberich, S., & Webb, A. (2010). New Rules And Old Companies (how the Mittelstand company approach in Germany encourages and benefits from a strong sense of social responsibility). *Bloomberg Businessweek, 4-10 Oct.*(4198), 72–75.

Thomas, A. J., Byard, P., & Evans, R. (2012). Identifying the UK's manufacturing challenges as a benchmark for future growth. *Journal of Manufacturing Technology Management, 23*(2), 142–156. doi:10.1108/17410381211202160

Tietjen, T., Decker, A. & Müller, D. H. (2011). *FMEA Praxis - Das Komplettpaket für Training und Anwendung*. München: Carl Hanser Verlag.

Tilson, H.A. (1998). Developmental Neurotoxicology of Endocrine Disruptors and Pesticides: Identification of Information Gaps and Research Needs. *Environmental Health Perspective 3*(106), 807-811.

Tiwari, V., Patterson, J. H. & Mabert, V. a. (2009). Scheduling projects with heterogeneous resources to meet time and quality objectives. *European Journal of Operational Research, 193*(3), 780–790. doi:10.1016/j.ejor.2007.11.005

Tönshoff, H. K., & Denkena, B. (2013). *Basics of Cutting and Abrasive Processes.* Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-33257-9

Tönshoff, H.K., Wulsberg, J.P., Kals, H.J.J., König, W. & Van Luttervelt, C.A. (1988). Developments and trends in monitoring and control of machining processes, *CIRP Annals, 37*(2), 611-622.

Udo, G. J. (1992). Neural networks applications in manufacturing processes. *Computers and Industrial Engineering, 23*(1-4), 97–100.

Ulaş, A., Yıldız, O. T., & Alpaydın, E. (2012). Cost-conscious comparison of supervised learning algorithms over multiple data sets. *Pattern Recognition, 45*(4), 1772–1781. doi:10.1016/j.patcog.2011.10.005

Universität Bremen. (2007). Promotionsordnung der Universität Bremen für die mathematischen, natur- und ingenieurwissenschaftlichen Fachbereiche vom 14. März 2007. Retrieved August, 13 2013 from http://www.math.uni-bremen.de/cms/media.php/59/PromO%20FB%202-5%20_14%203%2007_.6442.pdf

Van Dorp, K.-J. (2002). Tracking and tracing: a structure for development and contemporary practices. *Logistics Information Management, 15*(1), 24–33. doi:10.1108/09576050210412648

Van Luttervelt, C.A. Childs, T.H.C., Jawahir, I.S., Klocke, F. & Venuvinod, P.K. (1998). Present situation and future trends in modelling of machining operations. *Annals of the CIRP, 47*(2), 587-626.

Vapnik, V, (1995). *The Nature of Statistical Learning Theory*. NY: Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. Hoboken: Wiley.

Verma, R. (2012). Data quality and clinical audit. *Anaesthesia & Intensive Care Medicine, 13*(8), 397–399. doi:10.1016/j.mpaic.2012.05.009

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). The Application of Support Vector Machines to Medical Decision Support: A Case Study. In *ECCAI Advanced Course in Artificial Intelligence, Chania, Greece (ACAI99)* (pp. 17–21).

Verstraete, F., Wolf, M. M., & Cirac, J. I. (2007). *Matrix product state representations*. Retrieved from http://arxiv.org/pdf/quant-ph/0608197.pdf

Viharos, Z. J. & Monostori, L. (1999). Intelligent, quality-oriented supervisory control of manufacturing processes and process chains. *In DYCOMANS Workshop*, Bled-Slovenia, 12-14 May, 1999 (pp. 129–134).

Viharos, Z., Monostori, L., & Vincze, T. (2002). Training and application of artificial neural networks with incomplete data. In *Lecture Notes of Artificial Intelligence, LNAI 2358, The Fifteenth International Conference on Industrial & Engineering Application of Artificial Intelligence & Expert Systems*, 17-20 June 2002 (pp. 649–659). Cairns, Australia: Springer Berlin Heidelberg.

Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of Management Journal, 15*(4), 407-426.

Wallace, E. & Riddick, F. (2013). *Panel on Enabling Smart Manufacturing (presentation)*. APMS 2013, September 11, 2013, State College, USA.

Wang, B., & Japkowicz, N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems, 25*(1), 1–20.

Wang, L.-H., Liu, J., Li, Y.-F., & Zhou, H.-B. (2004). Predicting protein secondary structure by a support vector machine based on a new coding scheme. Genome Informatics. *International Conference on Genome Informatics, 15*(2), 181–90.

Wang, R.Y. & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 4*(12), 5–33.

Wang, K., & Tsung, F. (2007). Run-to-Run Process Adjustment Using Categorical Observations. *Journal of Quality Technology, 39*(4), 312–325.

Wang, H., & Wang, S. (2009). Towards optimal use of incomplete classification data. *Computers & Operations Research, 36*(4), 1221–1230. doi:10.1016/j.cor.2008.01.005

Wannenwetsch, H. (2010). *Integrierte Materialwirtschaft und Logistik Beschaffung, Logistik, Materialwirtschaft und Produktion*. Berlin: Springer.

Wasikowski, M., & Chen, X. (2010). Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 22*(10), 1388–1400.

Westkämper, E. & Warnecke, H. (2010). *Einführung in die Fertigungstechnik.* (8th Edition). Wiesbaden: Vieweg+Teubner-Verlag.

Weisstein, E. W. (2011). *Bipartite Graph.* Retrieved May 10th, 2013 from http://mathworld.wolfram.com/BipartiteGraph.html.

Whitehall, B. L., Lu, S. C.-Y., & Stepp, R. E. (1990). CAQ: A machine learning tool for engineering. *Artificial Intelligence in Engineering, 5*(4), 189–198. doi:10.1016/0954-1810(90)90020-5

Wiers, V. C.S. (2002). A case study on the integration of APS and ERP in a steel processing plant. *Production Planning & Control, 13*(6), 552-560.

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing, 21*(6), 2560–2574. doi:10.1016/j.ymssp.2006.12.007

Wiendahl, H.-P. & P. Scholtissek (1994). Management and control of complexity in manufacturing, *CIRP Annals, 43*(2), 533-540.

Wiering, M. & Van Otterlo, M. (2012). *Reinforcement Learning: State-Of-The-Art.* New York: Springer.

Wiig, K. M. (1998). Perspectives on introducing enterprise knowledge management. *In Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)*, 29-30 Oct. 1998. Basel, Switzerland.

Williams, D., Liao, X., Xue, Y., Carin, L. & Krishnapuram, B. (2007). On classification with incomplete data. *IEEE transactions on pattern analysis and machine intelligence, 29*(3), 427–36. doi:10.1109/TPAMI.2007.52

Winkler, W. E. (2004). Methods for evaluating and creating data quality. *Information Systems, 29*(7), 531–550. doi:10.1016/j.is.2003.12.003

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (p. 665)*. Burlington: Elsevier.

Wu, Q. (2010). Product demand forecasts using wavelet kernel support vector machine and particle swarm optimization in manufacture system. *Journal of Computational and Applied Mathematics, 233*(10), 2481–2491. doi:10.1016/j.cam.2009.10.030

Wuest, T., Hribernik, K., & Thoben, K. (2012a). Can a Product Have a Facebook ? A New Perspective on Product Avatars in Product Lifecycle Management. *In L. RIVEST, A. Bouraz, & B. Louhichi (Eds.), Product Lifecycle Management: Towards Knowledge-Rich Enterprises. Proceedings of the 9th International Conference on Product Lifecycle Management.* Montréal, Canada.

Wuest, T., Hribernik, K., & Thoben, K. (2013a). Digital Representations of Intelligent Products: Product Avatar 2.0. *In M. Abramovici & R. Stark (Eds.), M. Abramovici and R. Stark (Eds.): Smart Product Engineering, LNPE (pp. 675–684).* Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-30817-8

Wuest, T., Hribernik, K. & Thoben, K.-D. (2014a). Accessing servitization potential of PLM data by applying the product avatar concept. *Production Planning and Control, XX*(XX), XX-XX *(accepted)*

Wuest, T., Irgens, C. & Thoben, K.-D. (2012b). Analysis of Manufacturing Process Sequences, using Machine Learning on Intermediate Product States (as Process Proxy Data). *In: Emmanouilidis, C., Taisch, M., & Kiritsis, D. (Eds.): APMS 2012, Part II, IFIP AICT 398, Heidelberg Berlin: Springer, 1-8.*

Wuest, T., Irgens, C., & Thoben, K.-D. (2013b). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing, 25*(5), 1167-1180.

Wuest, T., Klein D., Seifert, M. & Thoben, K.-D. (2011a). Challenges for grown engineering SMEs with a diverse product portfolio on information management and product tracking and tracing. *In: Frick, J. (Eds.) (2011). Value Networks: Innovation, Technologies and Management. Proceedings of the APMS 2011 Int. Conference of Advances in Production Management Systems*, September 26 - 28, 2011, Stavanger, Norway. ISBN 978-82-7644-461-2.

Wuest, T., Klein, D., Seifert, M. & Thoben, K.-D. (2012c). Method to describe interdependencies of state characteristics related to distortion. *Materialwissenschaft und Werkstofftechnik, 43*(1-2), 186–191. doi:10.1002/mawe.201100908

Wuest, T., Klein, D. & Thoben, K.-D. (2011b). State of steel products in industrial production processes. *Procedia Engineering, 10*(2011), 2220-2225.

Wuest, T., Knoke, B. & Thoben, K.-D. (2014b). Applying graph theory and the product state concept in manufacturing. *Procedia Technology, 15*(2014), 349-358. doi: 10.1016/j.protcy.2014.09.089.

Wuest, T., Liu, A., Lu, S. C.-Y. & Thoben, K.-D. (2014c). Application of the stage gate model in production supporting quality management. *Procedia CIRP, 17*, 32-37. doi: 10.1016/j.procir.2014.01.071.

Wuest, T. & Thoben, K.-D. (2012). Exploitation of Material Property Potentials to Reduce Rare Raw Material Waste - A Product State Based Concept for Manufacturing Process Improvement. *Journal of Mining World Express (MWE), 1*(1), 13-20.

Wuest, T., Tinscher, R., Porzel, R. & Thoben, K.-D. (2014). Experimental research data quality in materials science. *International Journal of Advanced Information Technology, 4*(6).

Wuest, T., Werthmann, D. & Thoben, K. (2013c). *Towards an Approach to Identify the Optimal Instant of Time for Information Capturing in Supply Chains*. In Prabhu, V., Taisch, M. & Kiritsis, D. (Eds.): APMS 2013, Part I, IFIP AICT 414, IFIP International Federation for Information Processing (pp. 3–12).

Yang, K., & Trewn, J. (2004). *Multivariate statistical methods in quality management*. New York: McGraw-Hill.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003*), Washington DC (p. 8).

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research, 5*, 1205–1224.

Yul, T. & Wang, G. (2009). The Process Quality Control of Single-Piece and Small-Batch Products in Advanced Manufacturing Environment. *In 16th International Conference on Industrial Engineering and Engineering Management, 2009. IE&EM '09*. (pp. 306–310). Beijing.

Yun, Z., Nan, M. A., Da, R., & Bing, A. N. (2011). An Effective Over-sampling Method for Imbalanced Data Sets Classification. *Chinese Journal of Electronics, 20*(3), 2–7.

Zantek, P. F., Wright, G. P. & Plante, R. D. (2006). A self-starting procedure for monitoring process quality in multistage manufacturing systems. *IIE Transactions, 38*(4), 293–308. doi:10.1080/07408170500208354

Zhang, T. (2002). On the Dual Formulation of Regularized Linear Systems with Convex Risks. *Machine Learning, 46*, 91–129.

Zhang, S., Jin, Z., Zhu, X., & Zhang, J. (2009). Missing Data Analysis: A Kernel-Based Multi-Imputation. In M. L. Gavrilova & C. J. K. Tan (Eds.), *Trans. on Comput. Sci. III, LNCS 5300* (pp. 122–142). Berlin Heidelberg: Springer.

Zhang, Y., Jiang, P., Huang, G., Qu, T., Zhou, G., & Hong, J. (2010). RFID-enabled real-time manufacturing information tracking infrastructure for extended enterprises. *Journal of Intelligent Manufacturing, 23*(6), 2357–2366. doi:10.1007/s10845-010-0475-3

Zhang, J., & Wang, H. (2009). A minimized zero mean entropy approach to networked control systems. *In Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference (pp. 6876–6881). Shanghai, China*: IEEE. doi:10.1109/CDC.2009.5400679.

Zingel, H. (2009). *Qualitätsmanagement und die ISO 9000er Normenfamilie.* Online-resource: http://hprichter.de/download/Qualitaetsmanagement.pdf. Retrieved 03.08.2011

Zobel, R., & Filos, E. (2006). Technology Management with a Global Perspective: The Case of IMS. *In PICMET'06 Proceedings*, 9-13 July 2006, Istanbul, Turkey (pp. 9–13).

Zoch, H.-W. (2009). Distortion engineering: vision or ready to application. *Mat.-wiss. u. Werkstofftech. 40*(5-6), 342-348.

Zoch, H.-W. (2012). Distortion engineering - interim results after one decade research within the Collaborative Research Center. *Materialwissenschaft und Werkstofftechnik, 43*(1-2), 9–15. doi:10.1002/mawe.201100881

Zoch, H.-W. & Lübben, Th. (2011). Verzugsbeherrschung - Systemorientierter Ansatz als wesentliche Voraussetzung für den Erfolg. Stahl Strukturen. In Bleck, W. (Ed.) (2011). *Industrie-, Forschungs-, Mikro- und Bauteilstrukturen - Tagungsband zum 26. Aachener Stahlkolloquium,* 19./20.05.2011, Verlagshaus Mainz, ISBN 3-86130-258-6.

Zorriassatine, F., & Tannock, J. D. T. (1998). A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing, 9*(3), 209–224

Appreciation of student contribution

This work contains results which have been achieved through supervision of student academic works:

- Dennis Zeglin, “Entwicklung einer Quality-Gate-basierten Methodik zur Steigerung der Prozessqualität in der Produktion“ (“Development of a Quality-Gate-based methodology to increase process quality in production”) (*master thesis*)

- Benjamin Knoke, “Analyse und Darstellung der Abhängigkeiten von Zustandseigenschaften in Produktionsprozessen“ (“Analysis and illustration of dependencies of product state characteristics in manufacturing processes“) (*master thesis*)

- Bernat Fabregà Creixell, “Development of a method to illustrate interdependencies between state characteristics of a manufacturing programme as of the customer's perspective” (*master thesis*)

- Jakub Mak-Dadanski, “Konzeption und Entwurf eines Qualitätsmanagementhandbuchs zur Archivierung und Bereitstellung von werkstoffwissenschaftlichen Forschungsdaten“ (“Conception and design of a Quality Management Manual for archiving and provision of materials science research data”) (*bachelor thesis*)

- Mohamed Yassin Affoun, “Methode zur Identifikation von relevanten Zustandseigenschaften eines Produktes am Beispiel der Zerspanung“ (“Method for identification of relevant product state characteristics evaluated in the machining process“) (*diploma thesis*)

- Sebastian Till Alexander Schäfer, “Konzept zur Einführung von Wissensmanagement in kleinen organisatorischen Einheiten“ (“Concept for implementation of knowledge management in small organisational units“) (*diploma thesis*)

- Johannes Kuder, Atilla Odabasi, Benedikt Rolfes, Jan Schönmann, Henning Sebastian Voeltz, Christian Wolff, Kamran Yazdian and Dennis Zeglin, “Informationsmanagement zur Verfolgung des Produktzustands entlang einer Fertigungskette“ (“Tracing of product states along the production process through information management”) (*student project*)

List of figures

List of abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
APS	Advanced Planning Systems
BOL	Beginning of Life
BN	Bayesian Networks
C	Soft margin parameter of SVM
CAD	Computer-aided Design
CAE	Computer-aided Engineering
CAPP	Computer-aided Process Planning
CAQ	Computer-aided Quality
CBR	Case-based reasoning
CHEM	Chemical Manufacturing
CIM	Computer Integrated Manufacturing
CSV	Comma-Separated Values
DM	Data Mining
DT	Decision Trees
EC	European Commission
EOL	End of Life
ERP	Enterprise Resource Planning
FDC	Fault Detection and Classification
FMEA	Failure Mode and Effects Analysis
FPR	False Positive Rate
FS	Feature selection
GA	Genetic Algorithm
HMS	Holonic Manufacturing System
IM	Information Management
IMS	Intelligent Manufacturing System
IBL	Instance-Based Learning
KD	Knowledge Discovery
KDD	Knowledge Discovery from Databases
KSF	Key Success Factor
KM	Knowledge Management
MAR	Missing at Random
MBR	Memory-Based Reasoning
MCAR	Missing Completely At Random

MES	Manufactruing Execution System
MI	Multiple Imputation
ML	Machine Learning
MLH	Maximum Likelihood
MNAR	Missing Not At Random
MOL	Middle of Life
MRP	Manufacturing Resource Planning
MS	Manufacturing System
NaN	Not a Number
NB	Naïve Bayesian Networks
NN	Neural Networks
NP	nondeterministic polynomial (time)
PCA	Principal Component Analysis
PDM	Product Data Management
PEID	Product Embedded Information Device
PLM	Product Lifecycle Management
PLS	Partial Least Squares
PR	Pattern Recognition
QM	Quality Management
RFE	Recursive Feature Elimination
RFID	Radio Frequency Identification
RL	Reinforcement Learning
RR	Rolls-Royce
SCM	Supply Chain Management
SECOM	Semiconductor Manufacturing
SLT	Statistical Learning Theory
SME	Small and Medium sized Enterprise
SMOTE	Synthetic Minority Oversampling TEchnique
SQC	Statistical Quality Control
STEP	Standard for the Exchange of Product model data
SVM	Support Vector Machine
TPR	True Positive Rate
TQC	Total Quality Control
TQM	Total Quality Management
Universal Plug and Play	UPnP

9 Annex

In the Annex, additional content is presented to extend the previously presented research and results. There are two main parts, first the pre-processing of the evaluation data sets is described in detail. Not only the approach on how to deal with missing values is explained but also how the synthetic processes for scenario I & II are created is shown in detail. A special focus in this part is laid on missing data (missing values) as it presents a common obstacle in data based manufacturing operations. How to handle missing values my influence the outcome of further analysis conducted with the respective data sets to a large extent. Therefore, the first subsection gives an introduction into theory on how to handle missing data. Following, the conducted data pre-processing of the three evaluation scenarios is presented. The respective references used in this section are included in the previous reference list. Later different additional tables and figures are sub summarized under the section miscellaneous.

9.1 Theoretical elaboration on missing data

As Kabacoff (2011) stated, "[…] in the real world, missing data are ubiquitous". In research as well as in application the approach how to handle missing data and information represents an important issue. In this subsection the terms and different kinds of missing data are described and established techniques to handle missing data and information are introduced.

It is important to understand that certain domains have to deal with different challenges when it comes to missing data. Where a lot of empirical research is conducted, e.g., business studies or psychology, missing data can be, among other things, unanswered questions in a questionnaire (Graham, 2009). In more experimental and observation oriented domains, e.g., engineering or environmental studies, missing data can be, among other things, based on technical failure, like failing recording equipment, bad connectivity or miscoded data (Alvo & Park, 2002; Zhang et al., 2011). There are even reasons unknown to the stakeholders why certain data is missing. It can be assumed that in the future, regarding the fast pace development of sensor and communication technology, missing data will remain an important research area (Williams, Liao, Xue, Carin & Krishnapuram, 2007).

The challenge, of how to analyze a data set with missing values depends on various factors. For example, Alvo & Park (2002) point out that missing data in multivariate data sets presents a different challenge than non-multivariate incomplete data sets and needs to be handled in a different way. In this research, the missing data problem can be considered is one of missing values in experimental and observing oriented domains, handling mostly technical reasons for missing values. Additionally, in ML, an important part of this research, data sets are often of high-dimensional and multivariate nature with complex patterns of missing values

(Ghahramani & Jordan, 1994). Therefore, the elaboration will exclusively focus on issue inherent to these problems, leaving aspects from other domains out.

Most analytical and statistical methods work under the assumption, that the data set the analysis is based on is complete. Some, like Neural Networks are known for their capability to handle data sets, which are noisy, imprecise and incomplete to a certain extent (Li & Huang, 2009). It is important to understand the nature of the missing data and the preconditions of the method, which will be applied in order to be able to estimate the impact on the ability to answer the substantive research questions (Kabacoff, 2011). In order to do so, a set of questions should be asked:

- "What percentage of the data is missing?
- Is it concentrated in a few variables, or widely distributed?
- Does it appear to be random?
- Does the covariation of missing data with each other or with observed data suggest a possible mechanism that's producing the missing values?" (Kabacoff, 2011)

Looking at the question whereas the missing data appears to be random or not, this is important for three central concepts in missing data theory: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) (Graham, 2012). This is important for an informed decision to delete or replace the data and if replace, what method is applicable.

Missing data and information can have various effects on analysis based on the data set in the focus, e.g., generate bias, affecting quality of (supervised) learning methods or classification (Zhang et al., 2009). Schafer & Graham (2002) highlight, that ad-hoc editing of data sets leading to the appearance of completeness may do more harm than good. Among the effects of missing data and/or ineffective editing are the following: findings can be biased, e.g., indicate different problems, inefficient and unreliable (Schafer & Graham, 2002).

Missing data can be generally divided in two categories (Graham, 2012):

- *item nonresponse*: describes the case when some, but not all data from the measurement session is available. Reasons can be e.g., loss during collection or storage, equipment malfunction
- *wave nonresponse*: describes the case when repeated measures are taken over time and all data for some sessions (measures at a point in time) are missing completely.

Graham (2012) states that item nonresponse missing data can be handled reasonably well with available methods whereas wave nonresponse missing data is more challenging. Kabacoff (2011) postulates a generic, three step approach researchers should follow when dealing with incomplete data sets:

"1. Identify the missing data.

2. Examine the causes of the missing data.
3. Delete the cases containing missing data or replace (impute) the missing values with reasonable alternative data values."

Of the three steps, only the identification of missing values is considered unambiguous. Even step two can proof hard to elaborate in some cases, as it requires in depth knowledge of the process, e.g., manufacturing programme, and the technique and method used to capture the data. Step three raises a fundamental question of two general options of handling missing values that come to mind (Kabacoff, 2011), both leading to a complete data set:

- Remove the measurements containing missing data from the data set
- Replace (complete) measurements with missing values with reasonable substitute values.

Removing the missing cases seems to be the logical action. However, it is not always the best choice. For one, there are cases, especially dealing with data sets of high-dimensionality, when removing cases with missing values eliminates a significant amount of the data available. This is when the missing values are spread over a large number of cases instead of multiple missing values per case. Another factor can be that valuable information can get lost when the cases with missing values are removed or a bias can be inserted in the data set (Wang & Wang, 2009).

For the other alternative action, replacing the missing values, there are a large number of methods available to complete data sets. Among those are e.g., triangle inequality; complete-case analysis (listwise deletion); Multiple Imputation (MI) and Maximum Likelihood (MLH) (Hathaway & Bezdek, 2002; Schafer & Graham, 2002, Kabacoff, 2011). Under the assumption of MAR, MI and MLH are presenting the state of the art today (Schafer & Graham, 2002; Graham, 2012). However, these approaches of replacing missing values to complete the data set have also certain risks of introducing bias, distorted, and unreliable conclusions, etc. (Feelders, 1999; Dasu & Johnson, 2002; Wang & Wang, 2009; Kwak & Kim, 2012). It is very important to decide on the right method for the available data, taking the product and process into account, and the analysis technique into account (Viharos et al., 2002; Wang & Wang, 2009).

This section focuses on information and data issues in manufacturing. It is split in two factions. The first focuses on the information quality issue, which is relevant to every approach based on manufacturing information like the *product state concept*. The *product state concept* has to comply with the information quality dimensions and incorporate the principles presented here. The second fraction is looking at the common problem of missing data, which is omnipresent in all industrial applications where data capturing is involved. The *product state concept* is dependent on product and process data and thus, missing values occur and have to be handled according to existing standards. In the evaluation section (section 6.1) a real data set with missing data values is used and the above stated principles applied.

In the following subsection, the evaluation scenarios I to III are introduced individually and the data pre-processing steps performed are described.

9.2 Pre-processing of data sets for evaluation scenarios

In this section the three evaluation scenarios are introduced in detail and the available data for each scenario is presented and analyzed. After the three processes and accompanying data sets are presented, individually necessary pre-processing steps are described in detail. The pre-processing entails among other things, replacing missing values (scenario II & III) and the generation of additional data (scenario I & II). The result of this section are three data sets ready for the application of SVM algorithms in order to identify state drivers. The three data sets complement each other in terms of the evaluation focus areas and goals.

9.2.1 Rolls-Royce (RR) - data set (scenario I)

In this section a data set resembling a manufacturing process of a highly stressed product from the aviation domain provided by Rolls-Royce (RR) is introduced. Due to confidentiality requirements by Rolls-Royce, the data set is made anonymous and the tangible product manufactured and observed cannot be disclosed.

The major advantage of this data set is the access to expert knowledge about the process. This allows to specifically choose suitable examples for the learning/training data set for the model generation of the classifier. This is assumed to be highly beneficial for the performance of the approach. This will be explained in more detail in section 6.2.

9.2.1.1 Pre-processing of RR data set

The process described by the data set is named Tom(RR) and consists of a set of 85 features (attributes) and 4195 examples (vectors). The parameters are labeled *para.2*, *para.3*,…,*para.n* and the real names and contexts are not provided to the researcher. The values of the different features are normalized between [-1;1] and the actual original values are not disclosed to the researcher. There are no known missing values within the provided data set. The data set ratio of ‘pass’ and ‘fail’ examples is balanced (50,0%) with 2098 ‘fail’ and 2097 ‘pass’. To achieve this balance and to support the non-identifiable nature of the data set, the data sets minority class was enhanced by applying the SMOTE method (unknown percentage) by the providing agency. The examples in this scenario are also not in timely sequence but in random order. As the data set is pre-processed by Rolls-Royce, there is no need for further pre-processing within the setting of this dissertation.

9.2.1.2 Structure of RR manufacturing programme

As the data set is pre-processed by Rolls-Royce, the need for further pre-processing is minimal. The provided data set describes an individual process and not a whole manufacturing programme with several process instances. Therefore, additional process instances are added by generating synthetic data sets based on the specific characteristics of the original RR data set.

In order to simplify the illustrative nature of the approach it was decided to limit the resulting manufacturing programme (TDH(RR)) to three linked manufacturing processes. The processes are called 'Tom(RR)', 'Dick(RR)' and 'Harry(RR)' (TOM(RR), DICK(RR) and HARRY(RR)) and they form a simple sequence as defined below in Figure 109.

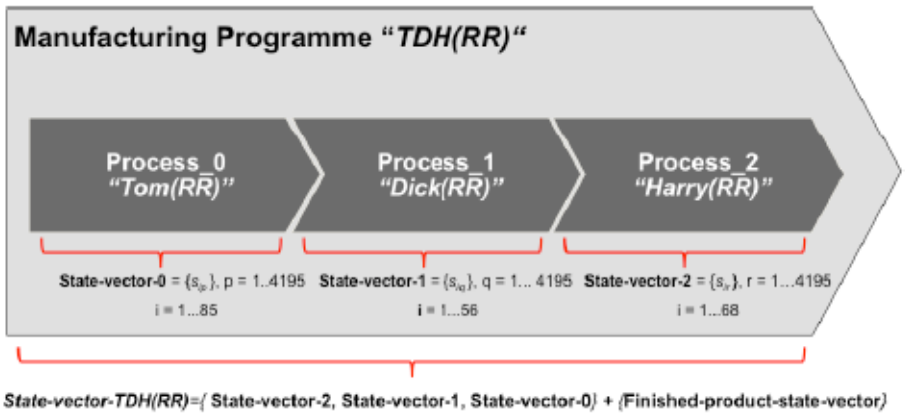

Figure 109: Manufacturing programme TDH(RR) and its three processes

Creating this three process manufacturing programme allows to focus on evaluating hypothesis 1.2, by applying the approach on different combinations of subsequent processes in order to identify state drivers within the programme. In the following paragraphs, the rationale and the generation of the data set is presented in more detail before the complete data set is presented.

The manufacturing programme thus represent a combination of real world process data and generic processes (synthetic data) onto which any specific set of manufacturing processes may be mapped. In the above Figure 109 the variables *k, m* and *n* indicate the numbers of quality observations (products) at each completed process stage (-> product state). In this case the number of examples for each inspection is identical for all processes and set to 4195 as provided by the initial RR data set. While *i* defines the number of process variables (-> state characteristics), in this case set to 85 for TOM(RR), 56 for DICK(RR) and 68 for HARRY(RR).

For clarity the following terminology will be used. The sequence of TOM, DICK and HARRY is the manufacturing programme TDH(RR) while TOM(RR), DICK(RR) and HARRY(RR) are the processes of that programme. As was previ-

ously stated, the (complete) manufacturing programme TDH(RR) and also each individual process (TOM(RR), DICK(RR) or HARRY(RR)) (in Figure 59 only process T is highlighted in red to avoid confusion) may be compared to the manufacturing programme utilized in scenario II and III, as they all have a final product as an outcome. Additionally, the combined state vector TD(RR), containing the processes TOM(RR) and DICK(RR) is also analysed.

In the following sub-section the generation of synthetic instances for the RR manufacturing programme is illustrated.

9.2.1.3 Generation of synthetic processes

The use of synthetic data sets is common in the area of statistical learning and data mining applications, etc. (Lundin, Kvarnström & Jonsson, 2002; Reiter, 2004; Nonnemaker & Baird, 2009). The reasons are manifold, among others the need for confidentiality, testing (Reiter, 2004; Abowd & Lane, 2004; Reiter & Raghunathan, 2007) and/or comparability (Lundin et al., 2002) purposes are arguments for the use of synthetic data. There are concerns that synthetic data produces different results as 'real data' (Abowd & Lane, 2004). However, multiple studies show that synthetic data has provided results and performed well in application (Nonnemaker & Baird, 2009)

There are several ways to create synthetic data and several variations of synthetic data composition (Lundin et al., 2002; Abowd & Lane, 2004; Reiter & Raghunathan, 2007; Jensen, 2007; Nonnemaker & Baird, 2009). Synthetic data has been successfully used in supervised classification, which is similar to the approach utilized in this research (Nonnemaker & Baird, 2009). Synthetic data can represent a large variety of processes, from processes involving heavy human interaction to fully automated ones (Barse, Kvarnström & Jonsson, 2003). A major advantage of synthetic data is that it can be used to demonstrate certain properties of a system (Barse et al., 2003).

In order to create a synthetic data set that replicates an existing authentic 'real world' data set, in this case the RR manufacturing process, certain (statistical) characteristics need to be derived. These characteristics provide the basis for the data generation.

In this case the process of generating complementary synthetic process data is designed as follows:

- Analyze *standard deviation* and *mean* for each feature (attribute) of the original data set (using Excel functions 'AVERAGE' and 'STDEV')
- Create a *probability* for each vector (using Excel function 'RAND' for a Gaussian normal distribution)

- Create a *probability* for each attribute (using Excel function 'RAND' for a Gaussian normal distribution)
- Both probabilities are summarized and divided by 2 giving a unique probability for each vector/attribute combination
- Replace each value by the inverse normal cumulative distribution (using Excel function 'NORMINV('combined probability of vector/attribute', 'mean for feature', 'standard deviation of feature'))
- The number of vectors is kept as it is given by the original authentic data set
- The number of features is varied, but the total number is always lower than the one in the original.
- The new data set is normalized using Range Transformation [-1;1]
- Additionally the class label is defined by identifying extreme clusters and determine the two which are furthest apart as the poles (determining the class) and serve as the learning set. The SVM model based on this learning set is used to assign the classes to the synthetic processes.

9.2.1.4 Cluster analysis to assign classes to synthetic processes

In this section the process of assigning classes ('fail' and 'pass') to the synthetic processes is presented. In this case the labeling is based on a cluster analysis approach whereas in the later described scenario 2, a random approach is utilized (see section 9.2.2.3). This way the complementary processes resemble the main characteristics of manufacturing processes within the specific domain but also distinguish themselves enough from the original to constitute new and stand alone processes.

In a first step a cluster analysis done to identify the extreme clusters which resemble the learning set. This is done using the programme Cluster3.0[23]. The programme analyses the data set using Euclidean distance with an average linkage clustering method (see Figure 124).

The resulting 'dendogram' is then exported and analyzed using Java Treeview (v.1.1.6r4)[24]. This programme allows to identify the extreme clusters of a data set including their correlation (see Figure 110). The software also allows exporting the examples contained within a cluster. By exporting the two most extreme clusters and assigning the label 'fail' to one and 'pass' to the other, a learning set is created. This learning set is then used to construct a hyperplane (train a SVM model). The labels for the complete synthetic process data set are then assigned by applying the created SVM model.

[23] http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
[24] http://jtreeview.sourceforge.net

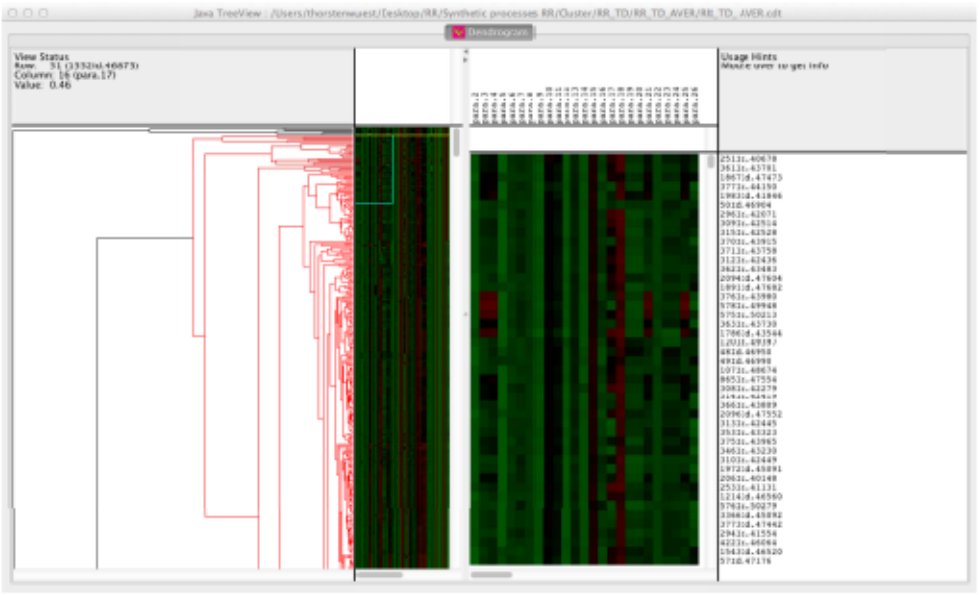

Figure 110: Screenshot of Java Treeview to analyze cluster 'dendogram' and extract examples

This way the resulting synthetic data set including labels are expected to show rather good classification performance in cross-validation. In this case this is welcome as it represents a different extreme to the second set of generated synthetic processes in scenario 2. This allows to test the concept under different circumstances.

For the first synthetic process, *DICK(RR),* the cluster analysis shows two extreme clusters:

- Cluster 1 has a correlation factor of 0.9730 (653 examples)
- Cluster 2 has a correlation factor of 0.9938 (332 examples)

These two clusters are the furthest apart and represent the cluster on the extreme left (cluster 2) and extreme right (cluster 1). The classes are assigned as follows: cluster 1 is considered 'fail' and cluster 2 is considered 'pass'. A cross-validation (DOT kernel, orig. parameters) of the learning set prior to model generation shows a perfect classification result with no misclassified examples indicating that the clusters allow a good separation by the constructed hyperplane.

The same procedure is applied to synthetic process *HARRY(RR),* identifying the two extreme clusters as follows:

- Cluster 1 has a correlation factor of 0.9958 (265 examples)
- Cluster 2 has a correlation factor of 0.9941 (330 examples)

These two clusters are the furthest apart and represent the cluster on the extreme left (cluster 2) and extreme right (cluster 1). The classes are assigned as follows: cluster 1 is considered 'fail' and cluster 2 is considered 'pass'. A cross-validation (DOT kernel, orig. parameters) of the learning set prior to model generation shows

a perfect classification result with no misclassified examples indicating that the clusters allow a good separation by the constructed hyperplane.

The cluster analysis for the combined vector *TD(RR)* leads to three identified extreme clusters:

- Cluster 1 has a correlation factor of 0.5323 (1889 examples)
- Cluster 1.5 has a correlation factor of 0.3991 (4 examples)
- Cluster 2 has a correlation factor of 0.8933 (796 examples)

In this case, cluster 1.5 is the extreme cluster on the right side. However, due to the small size of cluster 1.5 it is combined with cluster 1, which is the next bigger cluster on the right hand side and assigned the label 'fail'. Cluster 2 represent the cluster on the extreme left and is given the label 'pass'. A cross-validation (DOT kernel, orig. parameters) of the learning set prior to model generation shows a perfect classification result with no misclassified examples indicating that the clusters allow a good separation by the constructed hyperplane.

The cluster analysis of *TDH(RR)* shows that 3 distinct extreme clusters can be identified:

- Cluster 1 has a correlation factor of 0.8693 (215 examples)
- Cluster 1.5 has a correlation factor of 0.3804 (3 examples)
- Cluster 2 has a correlation factor of 0.8666 (366 examples)

In this case, cluster 1.5 is the extreme cluster on the right side. However, due to the small size of cluster 1.5 it is combined with cluster 1, which is the next bigger cluster on the right hand side and assigned the label 'fail'. Cluster 2 represent the cluster on the extreme left and is given the label 'pass'. A cross-validation (DOT kernel, orig. parameters) of the learning set prior to model generation shows a perfect classification result with no misclassified examples indicating that the clusters allow a good separation by the constructed hyperplane.

The perfect classification results are to be expected as the classes were assigned based on previously developed SVM model. If the same results would show in a 'real world' data set, this could resemble a case of serious overfitting. In the case of scenario I, the very good classification results are desired to provide an example case. The following manufacturing programme presented in scenario II does not sport such a perfect classification result to represent cases with a more challenging data basis.

9.2.1.5 Complete RR data set

The resulting process vectors (TOM(RR), DICK(RR) & HARRY(RR)), combined state vector (TD(RR)) and the manufacturing programme state vector (TDH(RR)) are presented in this sub-section.

The original process vector TOM(RR) entails 4195 examples with 85 descriptive features/attributes. The data set is balanced with 2098 'fail' and 2097 'pass' examples (50.0%).

The synthetic process vector DICK(RR) comprises 4195 examples as well, each described by 56 descriptive features/attributes. The data set can still be considered rather balanced with 1500 'fail' and 2695 'pass' examples (35.8%).

The last synthetic process vector HARRY(RR) contains 4195 examples with 68 features/attributes. The ratio of 44.5% makes the data set slightly less balanced than the original (2328 'pass' & 1867 'fail').

The combined state vector TD(RR) with its 4195 examples and 141 features/attributes has a ratio of 55.1% making it also slightly less balanced than TOM(RR) (1885 'pass' & 2310 'fail').

The manufacturing programme vector TDH(RR) consists of 4195 examples with a total of 209 features/attributes. With 2019 'fail' examples and 2176 'pass" examples, the ratio is 48.1% and almost as balanced as the original data set.

In the following section, the data pre-processing of the second scenario, resembling a chemical manufacturing process is presented.

9.2.2 CHEM - data set (scenario II)

In this section, the second scenario based on a chemical manufacturing programme supplemented by synthetic processes, similar to the above example, is presented. The scenario and its defining characteristics and principles are illustrated below.

9.2.2.1 Structure of CHEM manufacturing programme

The data set describes a chemical manufacturing process is publically available as part of the 'Applied Predictive Modeling' package (CRAN-R, 2014; Kuhn & Johnson, 2013). In the manufacturing process described, a raw material is going through a sequence of 27 operations to manufacture a pharmaceutical product.

The data set consist of 176 examples (vectors) with 57 attributes (features). Of those 57 attributes, 12 represent measures of the raw material (input product state) and 45 measures of the manufacturing process. The measures of the manufacturing process include but are not limited to: temperature, drying time, washing time, and concentrations of by-products during different operations. The vectors are not all independent as some resemble form a batch of the raw material (Kuhn & Johnson, 2013). The quality criteria in this case are the values of the 'Yield' attributes, measured at the end of the process. As this data set is also used for regression analysis, for this application the yield threshold dividing the data set in 'pass'/'fail' ex-

amples is set at 39. All examples with a yield of equal or greater 39 are considered 'pass' (good state) and all examples with a yield of lower than 39 are considered 'fail' (bad state). Choosing this threshold allows for a realistic and not too unbalanced data set considering both classes.

The chemical manufacturing data set is available in an R data format (*.RData). This is extracted and saved as a *.csv file and imported to Excel. The first analysis using RapidMiner (v5.3) shows the following characteristics:

- 127 examples are of 'pass' quality (72.2%)
- 27 (15.3%) of all examples contain missing values
- 109 values are missing from all attributes (1.1% of all values)
- No missing values in the classification attribute ('Yield')

9.2.2.2 Pre-processing of CHEM data set

The first step in pre-processing of the CHEM data set is to align the seemingly chaotic usage of commas and dots for decimal points. This is done by importing and exporting the data set in RapidMiner (v5.3).

The second step in pre-processing the CHEM data set is to focus on the missing values. Like most 'real' manufacturing process data (Kabacoff, 2011), the CHEM data set contains also a certain amount of missing data (null values) (Mccann et al., 2010). Being a key step of every ML or DM approach, data pre-processing can make up for a significant amount of the overall effort (Cios & Kurgan, 2002). However, the CHEM data set can be considered relatively clean as the total amount of missing values makes up for 1% of all values and is limited to 27 (out of 176) examples. The SECOM data set, illustrated in the following section, presents a more challenging example for a data set in need of significant pre-processing.

In this case, the choice is to eliminate 6 examples who contain more than 10 missing values and replace the missing values of the remaining examples instead of eliminating all examples which contain missing values (see Table 14). The reasons are two-fold: firstly, the total amount of examples in this data set is already limited with 176, further elimination would create an even larger discrepancy (ratio) between the number of examples and the number of attributes. Secondly, the number of missing values is considerably small with 1,1%of all values and 0.4% of remaining values after elimination of the 6 examples containing more than 10 missing values. The remaining data set contains 123 'pass' (72.4%) and 47 'fail' (27.6%) examples.

In order to replace the missing values, first, all 'Na' are replaced by empty cells in excel. This way the missing values named 'Na' and the missing values already represented by empty cells are equalized. RapidMiner (v5.3) contains different func-

tions for replacing missing values in data sets. The most commonly used one is using the average of an attribute. In a majority of cases the average function is applied. The respective attributes resembling the cases are listed in brackets.

- ***Case 1:*** Attributes where missing values are replaced using the *average* function as the values are diverse and resemble the wide variety of possible values (BiologicalMaterial01; ManufacturingProcess03; ManufacturingProcess06; ManufacturingProcess14).
- ***Case 2:*** Attribute where missing values are replaced using the *MinMax* function as within this process, the values alternate between extremes. In such a case, using the average function may introduce some previously not used values and thus a bias (ManufacturingProcess02).
- ***Case 3:*** Attribute where missing values are replaced by a manually defined value using the *Value* function as in this case the variation between very high and very small numbers jumped between extremes (ManufacturingProcess10; ManufacturingProcess11).

With eliminating or replacing missing values, the data pre-processing is not yet completed for the following application of the SVM algorithm. The next step is a normalization (also known as 'standardization') process which has to be executed in order to standardize the data set. This means ensuring the values within the different features/attributes are made comparable by adjusting the scale. As in "many applications, the available features are continuous values, where each feature is measured in a different scale and has a different range of possible values. In such cases, it is often beneficial to scale all features to a common range" (Ben-hur & Weston, 2010). Normalization plays an important role in the preparation of data sets prior to many data analysis and ML algorithms (Herbrich & Graepel, 2001). For SVM application it has been found that pre-processing, especially normalization of the input space is of great importance (Graf & Borer, 2001). The accuracy of SVM can suffer if no normalization step is executed within the data pre-processing stage (Ben-hur & Weston, 2010). Graf & Borer (2001) show that it is possible to apply normalization within the feature space through normalized kernel functions. However, in this research the input will be normalized.

For the data pre-processing before application of SVM algorithms, the use of the normalization method 'range transformation' is widely accepted (Graf & Borer, 2001; Abe, 2003). The range may be set to [0;1] or [-1.0;1.0] without having an effect on the performance of results of the SVM analysis (Abe, 2003). This normalization method will be applied to the CHEM data set, which was previously cleaned from missing values. The normalization is executed by utilizing RapidMiner (v5.3). It allows for an easy design of data processing processes for various purposes. In this case the process for normalizing the CHEM data sets contained three main elements:

- ***Data input:*** The process element reads the source file, in this case a Microsoft Excel *.xlsx file
- ***Normalization:*** The process elements normalizes the data file according to certain parameters
- ***Data output:*** The process element provides an output of the normalized data set, in this case in from of a Microsoft Excel *.xlsx file.

The parameter settings for the normalization task of the designed process are set to the method of range transformation [min;max] [0;1] for all features/attributes in the data set. This method normalizes all values of the selected features/attributes within the specified range.

After the data set is normalized, the labels 'SVM' and 'Identifier' are added. 'SVM' adds the classes with value 'positive' for each example with Yield equal or over 39 and 'negative' for Yield lower than 39. 'Identifier' replaces the heading for the numbering of the examples (original 'A'). At the same time the Yield attribute is eliminated.

After assigning classes to the examples different analyses can be conducted. Looking at the distribution of ‚fail' examples over time by organizing the examples in timely succession (see Figure 92) it can be seen that they are not equally distributed. This may indicate a change of raw material during the end of the succession or wear of machines which influences the quality. Even though, this is not directly relevant for the following analysis, it indicates the importance of the research as the diagram shows a certain timely accumulation of failures within the process. If an early identification of problematic states can be utilized, the parameters may be adjusted to prevent the following failures. However, at this point that has to be considered speculative.

With this last step the data set pre-processing of the CHEM manufacturing process is finalized. In the next sub-section the generation of synthetic complementary processes and their combination to a manufacturing programme is illustrated.

9.2.2.3 Structure of complete CHEM manufacturing programme and generation of synthetic processes

Similar to the generation of synthetic processes in the previously described RR scenario, the CHEM process is supplemented by two additional synthetic processes based on the characteristics of the original real world process. As the process was already described in detail in the sections 9.2.1.2 and 9.2.1.2, in this section just the main parameters are presented. As can be observed in Figure 111, the manufacturing programme consists of three processes with 170 examples (vectors) each and different number of attributes (Tom(CHEM) = 57; Dick(CHEM) = 48; Harry(CHEM) = 32).

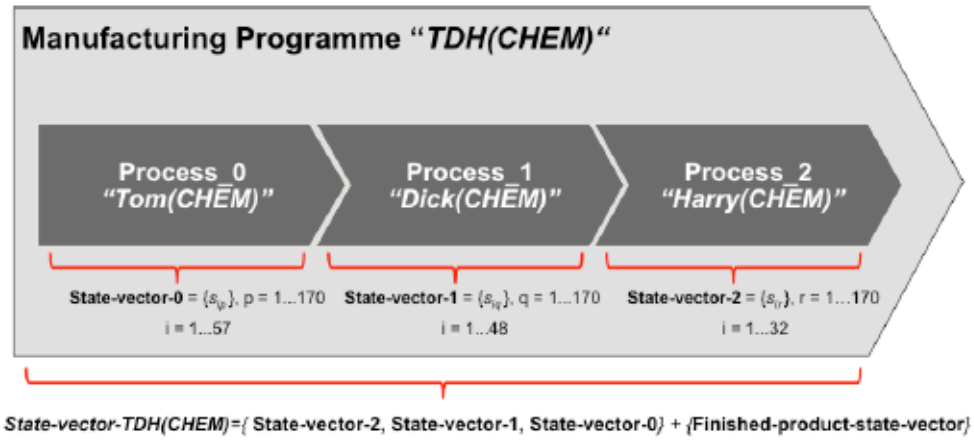

Figure 111: Manufacturing programme TDH(CHEM) and its three processes

The created synthetic process *Dick(CHEM)*, the second process within the manufacturing programme, contains 28 (of 170) (16.5%) negative examples and 48 attributes (features) plus the classifying attribute ('positive'/'negative') based on 'yield' and the Identifier.

The third process, *Harry(CHEM)* contains 22 (of 170) (12.9%) negative values and 32 attributes (features) plus the classifying attribute ('positive'/'negative') based on 'yield' and the Identifier.

The combined state vector TD(CHEM), containing the processes TOM(CHEM) and DICK(CHEM) resembles 170 examples with 105 attributes. Of those 170 examples, 63 are negative (37.1%).

Thus, the overall manufacturing programme *TDH(CHEM)* 170 examples and 137 attributes (features) plus the classifying attribute ('positive'/'negative') based on 'yield' and the Identifier. The final ratio is made up from 30 negative examples at the final quality control (17.6%).

The results of the cross-validation of the different processes and combined vectors are summarized in the later section 6.3 (see Figure 71). The classification performance is significantly lower for the CHEM data than it was for the RR data. For the synthetic and combined vectors that is due to the different approaches in designing the data sets and thus the desired outcome to evaluate different examples during the evaluation.

In the next section, the SECOM manufacturing programme is introduced and analyzed. This data set is considered very challenging given its nature. This corresponds with it being posted as part of the 'Causality Challenge" (McCann et al., 2010). The goal is to show that the approach is also applicable in challenging real world manufacturing examples.

9.2.3 SECOM - data set (scenario III)

In this section the third scenario, also based on a 'real world' data set, utilized to evaluate the hypotheses of this dissertation is introduced. The data set resembles a manufacturing programme from the semiconductor industry (McCann et al. 2010) available in the UCI ML repository called SECOM (McCann & Johnston, 2008). The evaluation with the SECOM data set has three main purposes: a) show if main results of the theoretical (synthetic) data set, introduced in the following sub-section, are comparable; b) test the applicability of the developed approach in a 'real world' problem, represented in the 'real' manufacturing data set and c) show that the approach is indeed able to handle high-dimensional data.

Semiconductor manufacturing involves a multi-stage, highly complex manufacturing programme with high quality requirements and advanced monitoring is often in place for fault detection and semiconductor yield improvement purposes (Harding et al., 2006; Li & Huang, 2009; Kim, Kang, Cho, Lee & Doh, 2012; Arif, Suryana & Hussin, 2013). The complexity and quality requirements are expected to increase further as e.g., device dimensions continue to shrink and the number of chips per wafer is expected to increase as well (May & Spanos, 2006). In semiconductor manufacturing it is possible to have a large number of operations within a process, often >500 which leads to large amounts of monitoring data (McCann et al., 2010). This factor presents an interesting option for the question raised in hypothesis 1.2, as it allows for various options to combine different operations to accumulated state vectors.

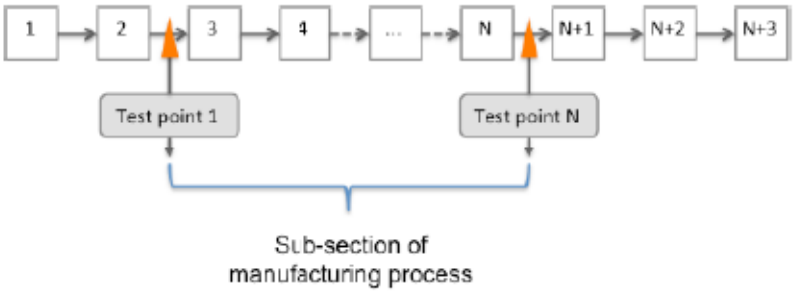

Figure 112: Production cycle (based on McCann et al., 2010)

The SECOM data set consists of 1567 example products with 590 features (591 incl. the quality assessment). All of these examples are additionally tagged with a time stamp and a quality assessment (ok/not-ok). Of these 1567 examples, 104 are not-ok (represented by '1') which means they fail to meet the quality requirements (6.64% failure rate). That means the data set is unbalanced. However, that is common for 'real world' data sets. The products passing the quality examination are marked with '-1' in the data set. The features represent process data (measurements) taken form the manufacturing programme (McCann et al., 2010). Each feature is understood as being a 'state characteristic' or potential state 'driver' during

the manufacturing programme. Information about the checkpoints capturing the process measurements/features and thus determining different processes/operations (product states) along the manufacturing programme are not available for this data set due to privacy concerns of the issuing company (see Figure 112).

Therefore, just the complete manufacturing programme can be analyzed applying the presented method. However, given the assumption that the data set could constitute a process with various operations the results can be interpreted as those of an individual process. This is comparable to the first (individual process) and last (complete manufacturing programme) step of the three process manufacturing programmes TDH(RR) (with its processes 'Tom(RR)', 'Dick(RR)' and 'Harry(RR)') and TDH(CHEM) (with its processes 'Tom(CHEM)', 'Dick(CHEM)' and 'Harry(CHEM)') analysis in the previous sections. Analyzing different subsequent states along the developing manufacturing programme is conducted, given the objective of presenting a challenging real world data set and how to pre-process such.

The SECOM data set, like many 'real' manufacturing process data (Kabacoff, 2011), contains also a certain amount of missing data (null values) (Mccann et al., 2010). Being a key step of every ML or DM approach, data pre-processing can make up for a significant amount of the overall effort (Cios & Kurgan, 2002). Ciao et al. (2007) summarize in their review the estimated effort for data preparation to 45-60% of the total effort in knowledge discovery. This reflects the fact that real world data, especially manufacturing data is often incomplete, redundant, inconsistent, and/or noisy (Zhang et al., 2009). In the following section, the missing data challenge of the SECOM data set is discussed in more detail.

9.2.3.1 Pre-processing of SECOM data set

In this section the SECOM data set is analyzed for its missing values to understand the challenge presented. In case of the SECOM data set, missing data is represented by 'NaN' as per MatLab (McCann & Johnston, 2008). This has to be considered for the data pre-processing and in the discussion of the results as it might affect the outcome depending on the algorithm used. Following, appropriate measures are discussed and applied to create a data set ready for the proposed application and evaluation approach.

Analyzing the SECOM data set and its missing data shows the following numbers:

- In total 41951 data points or 4.54% of the data is missing ('NaN').
- 0 examples (0%) are missing the quality assessment results.
- 104 examples (6.64%) do not pass the quality assessment ('1').
- 1567 examples (100%) contain missing values.
- 0 examples (0%) contain more than 50% missing data.
- 3 examples (0.19%) contain more than 20% missing data.

- 34 examples (2.17%) contain more than 10% missing data.
- 328 examples (20.93%) contain more than 6% missing data.
- 473 examples (30.19%) contain more than 5% missing data.
- 1205 examples (76.90%) contain more than 3% missing data.
- 1558 examples (99.46%) contain more than 1% missing data.
- 538 features (91.19%) contain missing values.
- 28 features (4.75%) contain more than 50% missing data.
- 32 features (5.42%) contain more than 20% missing data.
- 52 features (8.81%) contain more than 10% missing data.
- 52 features (8.81%) contain more than 5% missing data.
- 60 features (10.17%) contain more than 2% missing data.
- 103 features (17.46%) contain more than 1% missing data.
- 0 features (0%) show identical values for all examples (e.g., 0, 100, etc.).
- 117 features (19.83%) show identical values for all examples (e.g., 0, 100, etc.) incl. missing data.
- 126 features (21.36%) show identical values for 98% of the examples (e.g., 0, 100, etc.) incl. missing data.

In the following subsection, the challenge which missing values within a data set present, is elaborated and a method to create a complete data set is chosen.

9.2.3.2 Adding missing values

In a first step of data pre-processing, all examples containing more than 6% of missing values are removed in order to minimize the risk of inflicting a possible bias through replacing missing values. This reduces the data set to 1239 examples/vectors. The reduced data set contains 86 (6.94%) examples/vectors which do not pass the quality assessment ('1'). Compared to the 6.64% of non-pass examples in the original data set 'pre-reduction', this distribution is acceptable and indicates no direct bias being introduced by deleting examples/vectors with more than 6% of missing values. The eliminated examples are summarized in in Table 16. This is a common process when handling data sets with large amounts of missing data. In this case the choice was to first reduce the amount of vectors rather than replacing features directly. By first eliminating features as e.g., Kerdprasop & Kerdprasop (2011) do in their study, the dimensionality would be reduced and this may alter the characteristics of the manufacturing programme and/or its processes more than it is absolutely necessary. Especially considering the feature selection that is applied within the evaluation sections. This is considered too important within this research and thus the feature space is reduced a selection of vectors containing too much missing values is eliminated. Next, the process of eliminating missing values is described in more detail.

To further reduce the number of missing values in the data set, for the features, showing missing values and are part of the 117 features (19.83%) showing identical values for all examples (e.g., 0, 100, etc.) incl. missing data the missing values are replaced by the same value as the remaining identical values for the other examples show (Feature No.: 6; 142; 179; 277; 314; 315; 415; 450; 451).

After this first measures to reduce missing values, the data set contains 3.70% (27057 data points) of missing values ('NaN') compared to 4.54% before. Furthermore, the previous number of 538 features (91.19%) containing missing values is reduced to only 95 features (17.79%) containing missing values.

After theses previous measures reduced the missing values, now the features containing missing values are targeted. Basically two approaches are applicable. First, all features containing missing values may be eliminated. Or secondly, all features containing more than 5 (variant 1) or 10 (variant 2) missing values are eliminated and the missing values of the remaining features (with less than 5 or 10 missing values each) are eliminated by identifying and eliminating the examples still containing missing values (see Figure 113).

The second approach was selected over the possibility to replacing missing values through existing data replacement methods (e.g., Provost, 1990; Schafer & Graham, 2002; Williams et al., 2007; Grzymala-Busse, Grzymala-Busse, Hippe & Rzasa, 2007; Li & Huang, 2009; Kabacoff, 2011; Graham, 2012) as it is less likely to induce a bias to the data set. Furthermore does the application of methods to replace missing values require in depth knowledge (e.g., by engineers) of the manufacturing process and the application of statistical tools (Kwak & Kim, 2012). As the SECOM data set is provided with limited information concerning the process layout, in depth knowledge is not available. Hence, applying advanced data replacement methods like MAR is not possible without a considerable risk of inducing a bias and altering the results.

The first approach applied to the data set leads to 1239 examples (79.06% of the original data set) with 485 features (82.2% of the original data set) without any missing values. The ratio 'pass' to 'non-pass' stays the same with 6.95%. However, by deleting 95 features (for a list of deleted features see Table 17) in the process, this may have a significant impact on the results as some of these features may be relevant for the identification of state drivers. Basically by deleting examples important support-vectors may be deleted and thus the whole knowledge picture be altered.

The second approach, eliminating all features that contain *more than 5 missing values* ***(variant 1)*** leads to 1239 examples with 520 features. A list of the eliminated features is illustrated in Table 18. This brings the ratio of missing values ('NaN') to 0.010% (67 data points) with a total of 14 examples still containing missing values whereas 1225 examples are now complete data vectors. By elimi-

nating the aforementioned 14 examples still containing missing data, the resultant complete data set consists therefore of 1225 examples (78.17% of the original data set) with 520 features (88.14% of the original data set). The ratio of 'pass' / 'non-pass' examples in this complete data set is 6.85% (85 examples with 'non-pass' / '1'). Compared to the first stage reduced data set (containing 86 (6.94%) examples/vectors which do not pass the quality assessment ('1')) and the 6.64% of non-pass examples in the original data set 'pre-reduction', this distribution is acceptable and indicates no direct bias. The additionally eliminated examples are: Example No. 22; 67; 118; 232; 570; 726; 886; 1085; 1219; 1223; 1305; 1321; 1373; 1461.

Applying the second approach with a slightly adjusted parameter of eliminating features with *more than 10 missing values* ***(variant 2)***, 1239 examples and 528 features contain 131 missing values (ratio: 0.020%). The list of the eliminated features of variant 1 illustrated in Table 18 is still valid for variant 2 except features No. 20; 85; 156; 220; 291; 358; 429; 492 are not eliminated initially. Of those 1239 examples, 1209 contain no missing values in this scenario. After eliminating the 30 examples still containing missing values that results in a complete data set containing 1209 examples (77.15% of the original data set) with 528 features (89.49% of the original data set) and a 'pass' / 'non-pass' ratio of 6.95 % (85 examples with 'non-pass' / '1'). The additionally eliminated examples are: Example No. 22; 65; 67; 103; 106; 108; 112; 118; 121; 124; 153; 192; 232; 390; 426; 483; 570; 625; 693; 726; 886; 1085; 1165; 1196; 1219; 1223; 1305; 1321; 1373; 1461.

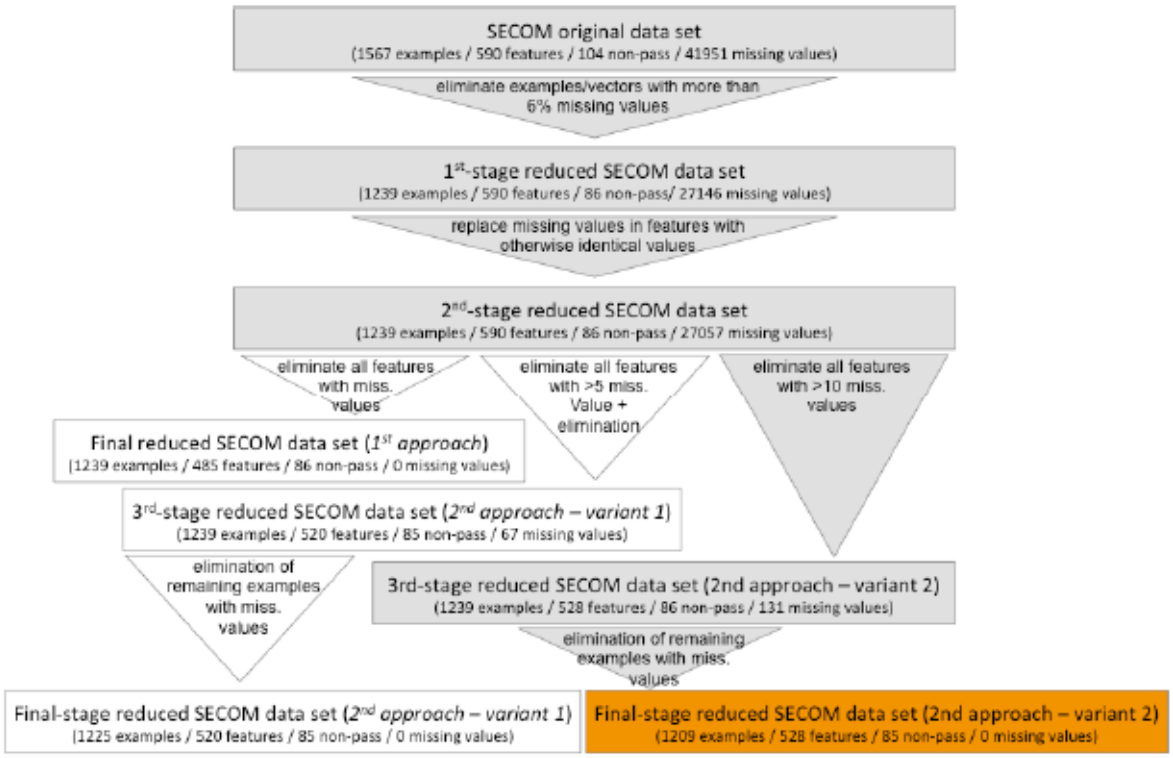

Figure 113: Schematic illustration of chosen approach to generate a complete data set (SECOM)

When comparing the two variations of the second approach, the resulting data sets are considered equally distributed with regard to 'pass' / 'non-pass' ratio. The difference lies in the number of examples and features. Whereas variant 1 contains a larger amount of examples/vectors (1225 vs. 1209), the second variant manages to keep more features (520 vs. 528). As the features represent so to speak state characteristics or potential state drivers, the benefit of having a larger number of features succeeds over a slight increase in examples/vectors for the purpose of this analysis. Therefore, the data set of ***variant 2*** is selected as the primary data set for the first scenario (see Figure 113).

After the challenge of handling missing values within the SECOM data set is successfully taken care of with the creation of a complete data set in different variations, necessary further pre-processing steps are described in the next section.

9.2.3.3 Further pre-processing measures (SECOM)

The normalization of the completed SECOM data sets ('approach 1'; 'approach 2 var. 1'; 'approach 2 var. 2') is executed by utilizing Rapidminer (v5.3) as described in the previous scenarios. The parameter settings for the normalization task of the designed process are set to the method of range transformation [min;max] [-1.0;1.0] for all features/attributes in the data set. This method normalizes all values of the selected features/attributes within the specified range.

Table 9: Summary of SECOM data sets after pre-processing

Name	No. of examples/ vectors	No. of features/ attributes	Normalization method [range]	No. of missing values
Approach 1	1239	485	Range transformation [-1;1]	0
Approach 2 var. 1	1225	520	Range transformation [-1;1]	0
Approach 2 var. 2	1209	528	Range transformation [-1;1]	0
(Approach 2 var. 2 plus 15)	(1224)	(528)	(Range transformation [-1;1])	(0)

The resultant complete SECOM data sets have the following specifications after all data pre-processing steps are finalized (incl. the later added appr. 1 var.2 plus 15) (see Table 9).

It has to be noted that the SECOM data set represents a very challenging data set. It was published as part of the "causality challenge" which suggests that classification will not be an easy task. The baseline results of classification performance published in accordance with the SECOM data set by McCann et al. (2010) show the difficulty of achieving good classification results with this data set. This corresponds with the findings of Kerdprasop & Kerdprasop (2011) investigating classification performance of the same data set.

9.3 Miscellaneous

9.3.1 Principles of modeling

The six main principles of modeling are:

- ***correctness***: in order to comply to the principle of correctness, the model needs to be following syntactic rules of modeling annotations (Rosemann & Schütte, 1997) and additionally be semantically correct. However, the semantically correctness is often not formally provable (Becker, 1998).
- ***relevance***: just elements which are of relevance for the modeling goal are included (Becker, 1998). A model is minimal, when no more elements can be eliminated without losing information important for the goal (Batini et al., 1992).
- ***economic efficiency:*** connected to the principle of relevance is the economic efficiency, which has an effect on all other principles (Rosemann & Schütte, 1997). The detail and effort has to be judged also by economic parameters. It may prove reasonable to use reference models in order to reduce the economic effort to create a model.
- ***clarity:*** this principle is targeting the understandability or comprehensibility. The goal is to create a clear arrangement and an intuitive illustration (Becker & Schütte, 2004).
- ***comparability:*** is targeting the compatibility of models created with different tools. This is especially important when reference models are utilized (Rosemann & Schütte, 1997). Generally it is advised to use as little as possible different tools and modeling annotations for process illustrations in order to reduce comparability problems (Becker, 1998; Becker & Schütte, 2004).
- ***systematic composition:*** is focusing on the structural consistency of the model (Becker & Schütte, 1997). This means that in case a modeling annotations allows different perspectives, the one is to be chosen which represents the to be modeled context. In case more than one perspective is chosen, a meta-model is to be utilized to ensure the abstract overview (Becker, 1998; Becker & Schütte, 2004).

9.3.2 Visualization models of product state concept

9.3.2.1 State model (layer 2)

The goal of the sub model, called ***state-model (layer 2)***, is to illustrate all process intra- and inter-relations of an individual state within a manufacturing programme and characterize these process intra- and inter-relations through a transfer function.

The transfer function or its placeholder shall be integrated in the modeling notation.

In respect of the chosen symbols of the state-model, the elements of the meta-model are applied as well (see Figure 114). In order to increase the readability, the visualization of the process structure and process parameters can be excluded from the model. However, if the visualization of the processes, process parameters and their influence on state characteristics desired, the visualization principles and symbols of the meta-model are to be applied in the state-model.

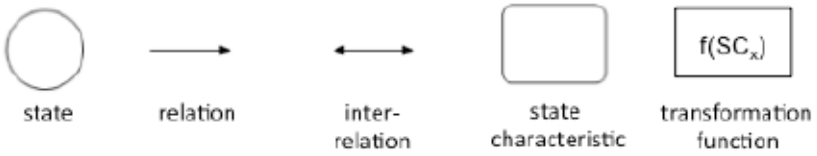

Figure 114: Symbols used in state-model

Within the state-model, the states are separated from the manufacturing programme structure and positioned next to each other. Within this model, only one state is in the focus at a time. This focus state should be highlighted within the model by choosing a deviant color for the color contour (see $state_{(z)}$ in Figure 115). Different from the meta-model visualization, only the state characteristics connected through relations to the focus state (through its state characteristics) are included in the visualization.

For describing the relations, rectangular boxes are placed on the edges, representing relations. This is similar to captions in BONAPART notations (Krallmann, Schönherr & Trier 2007; Hoyer 1988). Besides the exclusion of processes/process parameters, the focus on one state at a time and a reduced number of state characteristics, the state-model differentiates itself from the meta-model through the visualization of overlaying, independent relations. Whereas in the meta model a number indicated the number of independent relations represented, in the state model, the different transfer functions are highlighted by individual boxes on top of the edge (see $SC_{(x1)}$ and $SC_{(y1)}$ to $SC_{(z1)}$ in Figure 115).

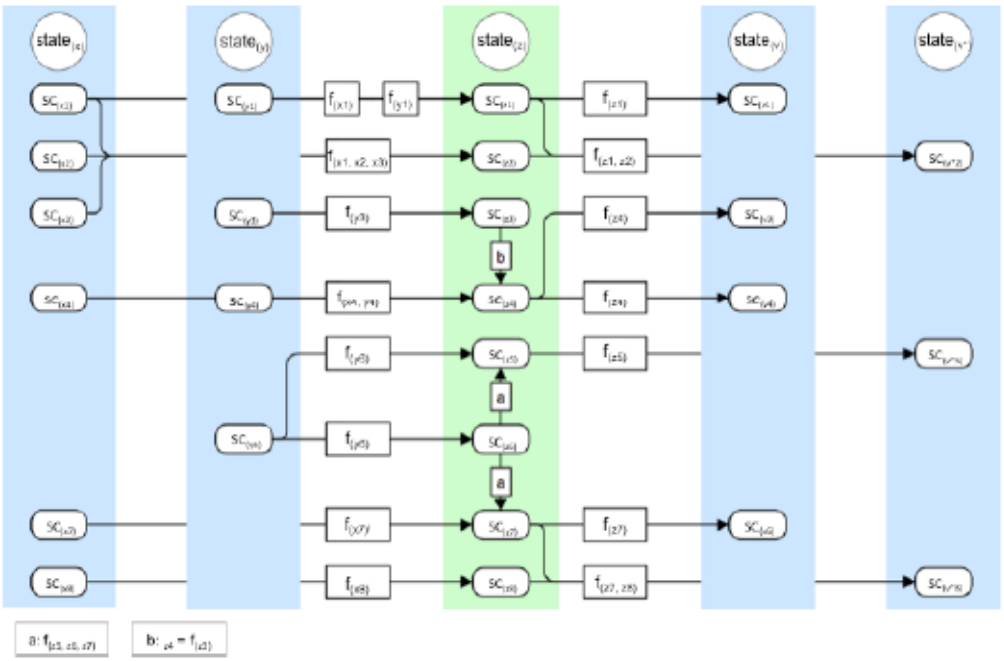

Figure 115: Exemplary illustration of state-model layer (layer 2)

Transfer functions can be of various natures in the state-model, e.g., a mathematical equation like the mass of a cylinder ($m = \rho * l * d^2 * \pi$). More complex transfer functions may be describable as well, be it through a differential equation or even in written form, e.g., the failure rate is in average 5% in the morning shift and 7% in the late shift. Such complex transfer functions shall be replaced by a placeholder in the state-model and connected to the relation underneath the model (see (a) & (b) in Figure 115).

In Figure 115 an exemplary application of the state-model including randomly selected relations and inter-/intra-relations. All previously described varieties of process intra- and inter-relations (see Figure 46 and Figure 47) are included in order to present the example in a comprehensive way. The focus state ($state_{(z)}$) is highlighted by a green contour in the middle of the figure. The exemplary application shows that transfer-functions within the state-model have to be replaced by placeholders even at this low level of inherit complexity the model represents. It might be necessary to split the model in parts, when a large number of state characteristics have to be summarized under the states and if a gap between parts (no existing process intra- and inter-relations) is given (see between $SC_{(z4)}$ and $SC_{(z5)}$ in Figure 115).

9.3.2.2 State characteristic model (layer 3)

Looking at individual state characteristics, there may be a need to reduce the complexity of the model even further under some circumstances. In order to comply with these requirements, the ***state characteristics model (layer 3)*** is developed. Next, the model and the rationale behind its development is introduced.

The state characteristics model has a single state characteristic in the focus and illustrates all existing process intra- and inter-relations of that individual state characteristic with other state characteristics of the manufacturing programme. The goal is to visualize all existing process intra- and inter-relations comprehensively in a clear well-arranged way. It is further important to be able to distinguish relations and inter-/intra-relations within the state characteristics model.

In order to achieve the above stated goals and requirements, the model is developed on the basis of an adapted cause-effect diagram (also known as Ishikawa-diagram) (Kern, 2008). The cause-effect diagram is not known originally as a process-modeling notation but as a tool for failure analysis. The model represents various influences of a problem. Those influences are structures in main causes and sub-causes, which leads to a so-called 'fishbone' structure (Kamiske & Brauer, 2008). The clear structure presents a chance to clearly visualize the stated goal of all process intra- and inter-relations of a single state characteristic once adapted from the original description. The adaptation starts by replacing the main focus from the 'problem' (original) to the 'focus state characteristic' (see SC_{z2} in green in Figure 116). 'Influences' (original) are replaced by process intra- and inter-relations respectively their transfer functions or placeholders in the state characteristics model (see blue boxes in Figure 116). The state characteristics with an influence on the transfer function are replacing the 'main causes' (original) in the fishbone structure (see Figure 116). It has to be distinguished between:

- Relations of state characteristics of current or previous states, which have an influence on the state characteristic in the focus. These relations are represented by an arrow coming form the left side towards the focus state characteristic.
- Inter-/intra-relations with other state characteristics, which occur necessarily within the same state, are represented by a two-headed arrow, positioned above and/or below the state characteristic in the focus.
- Relations to other state characteristics from the state characteristic in the focus are represented by an arrow heading to the right from the focus state characteristic. It is important that state characteristics, which have a combined relation with the state characteristic in the focus on another state characteristic are illustrated by a line without an arrowhead. Arrows heading away represent state characteristics being influenced.

Shall elements of lined dependencies be illustrated without direct connection to the focus state characteristic, this relation needs to be addressed in the transfer function. At the same time, such associated influences shall be represented as state characteristics or process parameters in form of replacing 'sub-causes' (original) in the diagram.

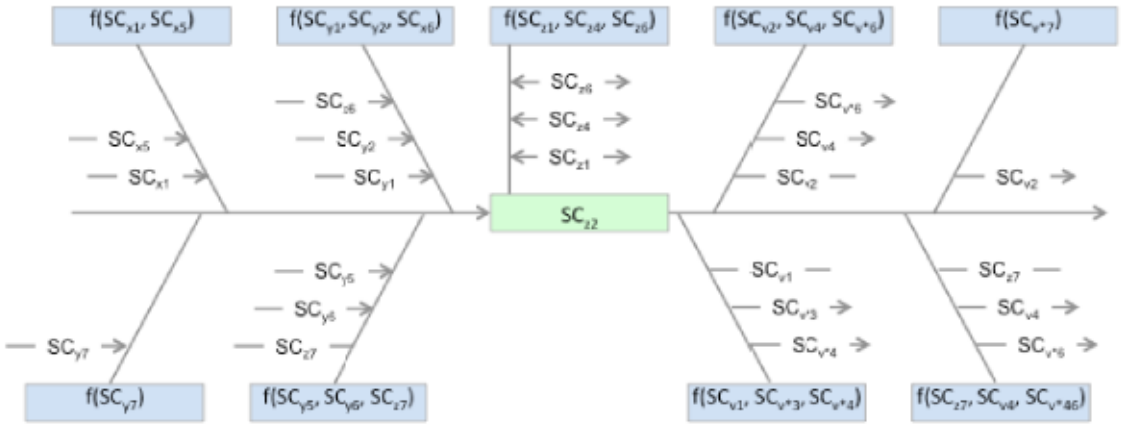

Figure 116: Exemplary illustration of state characteristic model (layer 3)

The exemplary illustration of the state characteristics model in Figure 116 gives an impression of the visualization. Depending on the complexity and possible analytical description of the transfer functions, they may need to be substituted by placeholders and described in detail at another location. In case the influences are quantifiable and the state characteristic in the focus is a stock figure with its value being time dependent, the application of the System Dynamics modeling notation (Sterman, 1992; Krallmann, Schönherr & Trier, 2007) might be appropriate. As this represents a special case and the methodology is adaptable to a large extent, this will not be elaborated in detail within this thesis.

9.3.3 Application of SMOTE in WEKA

A first step in applying the SMOTE technique is loading the SECOM data set in the WEKA software. As WEKA only accepts *.arff format, RapidMiner (v5.3) is used to convert the *.xlsx file to an *.arff file. Once the file is loaded into WEKA, the two classes can be observed (see Figure 117).

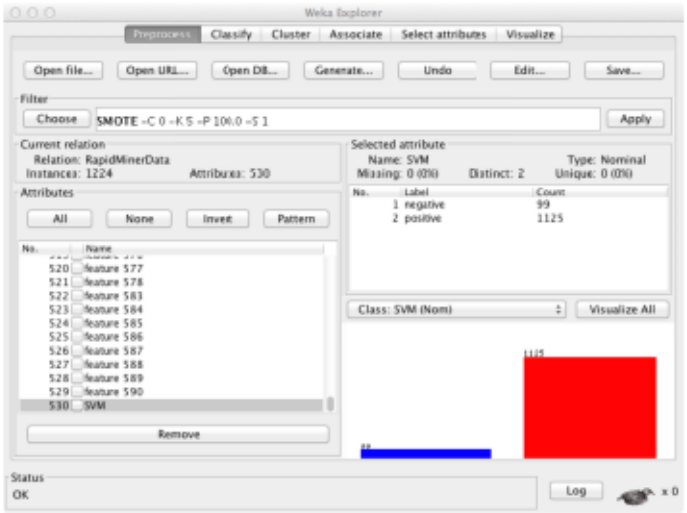

Figure 117: SECOM data set before applying SMOTE oversampling technique in WEKA

In a next step the to be applied SMOTE technique is chosen as a filter and the parameter adjusted (see Figure 118). In this case the classValue=0 defines the minority class to be oversampled. The percentage (in this case 500%) predefines the resulting number of examples in the minority class after the SMOTE application.

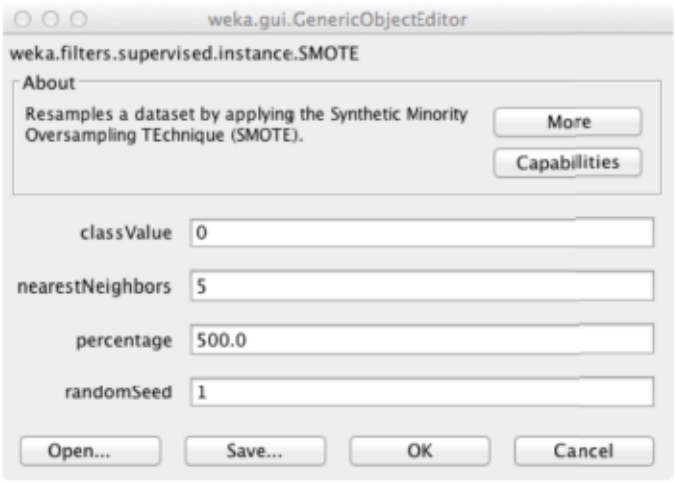

Figure 118: Parameter of SMOTE filter in WEKA

After applying SMOTE in WEKA, a few more steps are necessary (see Figure 119). At first the numbering of the additional examples of the minority class has to be adjusted. WEKA does not continue the numbering of the existing examples but uses existing numbers. As WEKA adds the additional examples to the end of the data set, this is done manual in Excel after converting the *.arff file to an *.xlsx file. The numbering of the additional examples is chosen to start at 'example 1600'.

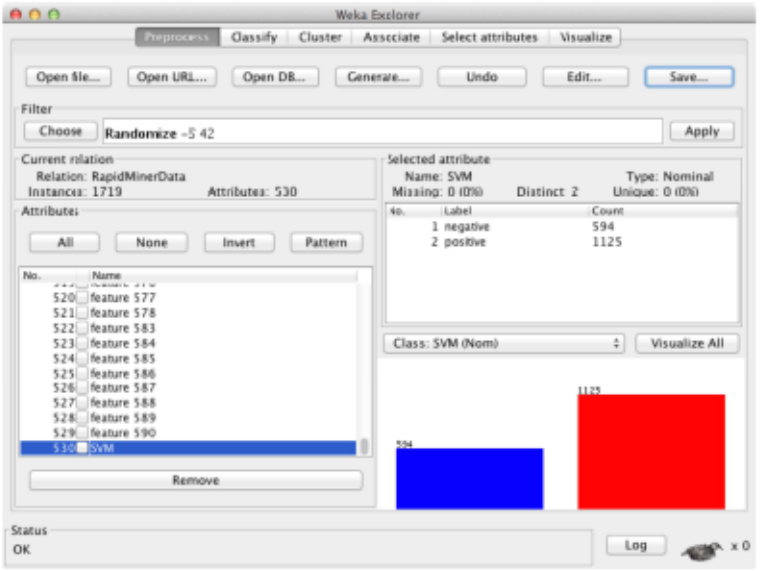

Figure 119: Randomizing the SECOM data set after SMOTE oversampling in WEKA

9.3.4 Tables and graphs

Table 10: 15 information quality dimensions and their influence on the product state concept

IQ dimension	Addressed by product state concept
Accessibility	The product state information and data represents product inherited information and data in universally accepted formats. This way the information can be processed, exchanged and transferred as needed.
Ease of manipulation	The product state information and data is structured according to the product and process requirements and can be used for various purposes despite the goals of the *product state concept*.
Reputation	Different systems can be used to derive the needed information and data. The reputation is based on the used system and it has to be ensured that the chosen system complies with this IQ dimension.
Free of error	This is out of the scope of the *product state concept* but methods and tools to cope with eventual errors in information and data are available.
Objectivity	Product state information and data represents not-interpreted information and (raw) data with in the defined set of relevant information.
Believability	Depending of the company and manufacturing programme, quality certifications can be used to ensure the processes.
Understandability	The product state information and data is directly understandable as it is connected to processes and products of a specific manufacturing programme.
Concise representation	Product state information & data is stored in universally accepted formats.
Consistent representation	The product state information and data is stored in a structured way based on the product and processes of the manufacturing programme.
Interpretability	Product state information and data represents raw data, which is descriptive to product and process and thus inherits a high interpretability.
Timeliness	The product state information and data properties are accurately stored and uniquely identifiable to an individual product through the checkpoint system and mapping.
Value-added	The goal of the *product state concept* is to derive new knowledge and support the increase of transparency through the manufacturing chain in order to support process and product quality improvements.
Completeness	The product state information and data should be stored as complete as possible within the set of relevant information. However, this depends also on the external circumstances like sensors, etc.
Appropriate amount of data	The *product state concept*s main objective is to identify a set of relevant information in order to reduce the amount of information and data to be handled.
Relevancy	The *product state concept*s main objective is to identify the set of relevant information. This contributes to ensure that the data and information captured is relevant for the chosen purpose.

Table 11: Quality issues in manufacturing

Top level issue	Issue	Example	Proposed solution
Quality level	High level of unremunerative / unplanned cost (Instone & Dale, 1989)	Operating and manufacturing costs (Instone et al. 1989)	Reduce costs by getting it "right first time" (Instone et al. 1989)
			Provide better product-related services (Instone et al. 1989)
			'Quality cost' management (Instone et al. 1989)
	Poor level of organization (Instone et al. 1989)	Shortage of parts (Instone et al. 1989)	
		Rescheduling of jobs (Instone et al. 1989)	
		Lack of information needed by various functions (Instone et al. 1989)	Quality control panels (Instone et al. 1989)
			Quality audits (Instone et al. 1989)
			Documentation of procedures (Instone et al. 1989)
		Goods leaving the assembly area not fully completed (Instone et al. 1989)	
Quality management	Leadership support (Instone et al. 1989; Sohal & Terziovski, 2000)	"[…] agreed that senior management need to take the lead if any qualityinitiative is going to be credible and effective." (Instone et al. 1989)	Reports, interviews and meetings (Instone et al. 1989)
		Short term vision (Sohal et al. 2000)	Companies that had invested in leadership training are more likely to succeed […]" (Sohal et al. 2000)
			Positive attitude towards quality, leadership education and training (Sohal et al. 2000)
	Staff involvement (Instone et al. 1989; Sohal et al. 2000)	People thinking quality is not their responsibility (Instone et al. 1989, p.25).	Emphasizing quality and introduction of quality plans to all employees in meetings and briefings (Instone et al. 1989)
			Developing appropriate performance indicators and rewards (Sohal et al. 2000)
	Customer / Supplier involvement (Sohal et al. 2000)		Manufacturing and services quality department (Badri et al. 1995)

			Integrating the voice of the customer and the supplier (Sohal et al. 2000)
	Poor level of organization (Badri, Davis & Davis, 1995)	Co-ordination between the quality department and other departments (Badri et al. 1995)	Encouraging automation in the process (Badri et al. 1995)
	Product design (Badri et al. 1995)		Manufacturing and services quality department (Badri et al. 1995)
Defective products	Operator error (Dhafr, Ahmad, Burgess & Canagassababady, 2006)	Inadequate skills (Dhafr et al. 2006)	Control of the process variables that affect the quality of the product, application of a lean manufacturing tool - Jidoka (Dhafr et al. 2006)
		Ignorance of operation instruction (Dhafr et al. 2006)	
	Machine fault (Dhafr et al. 2006)	Faulty machine design (Dhafr et al. 2006)	
		Temporary fault (Dhafr et al. 2006)	
	Inadequate surrounding / environments (Dhafr et al. 2006)	Inadequate cleaning (Dhafr et al. 2006)	
		Contamination not noticed (Dhafr et al. 2006)	
	Fault in process (Dhafr et al. 2006)	Inadequate procedure (Dhafr et al. 2006)	
		Procedure temporary inadequate (Dhafr et al. 2006)	
	Faulty raw material (Dhafr et al. 2006)	Supplier error (Dhafr et al. 2006)	
	Rejected products, due to: (Dhafr et al. 2006)	Inclusion (71.43%) (Dhafr et al. 2006)	Appropriate cleaning of parts; use of rust-free jigs; reducing particles in the air; control measures for quality of the paint; appropriate material handling procedures to minimize possibilities of scratch (Dhafr et al. 2006)
		Scratches (12.91%) (Dhafr et al. 2006)	
		Fibers (11.48%) (Dhafr et al. 2006)	
		Under spray, paint run, warped, second pass, pin holes (Dhafr et al. 2006)	
Quality loss	Chance variations and systematic errors (Robles & Roy, 2004)	Resulting from tool wear (Robles et al. 2004)	
		Process sequence errors: Disk filing, shaping, polishing, milling, grinding (Robles et al. 2004)	
	Tolerance optimization (Robles et al. 2004)	Limitations of process capabilities (Robles et al. 2004)	Statistical tolerance analysis (Robles et al. 2004)
		Limitations of machining capacities (Robles et al. 2004)	
		Measurement errors (Robles et al. 2004)	
	Internal failure (Shetwan, Vitanov & Tjahjono, 2011)	Reworking (Shetwan et al. 2011)	
		Scrapping (Shetwan et al. 2011)	
		Replacement (Shetwan et al. 2011)	
	External failure	Replacement (Shetwan et al. 2011)	

	(Shetwan et al. 2011, p.475)	Repairing (Shetwan et al. 2011)	
		Quality loss (Shetwan et al. 2011)	
	Errors from machine tools (Jiang, Jia, Wang & Zheng, 2012)	Wear of guideway (Jiang et al. 2012)	Maintaining the equipment (Jiang et al. 2012)
		Fault of electric motors (Jiang et al. 2012)	
	Errors from cutting tools (Jiang et al. 2012)		
	Fixtures (Jiang et al. 2012)		
	Errors of the machining process (Jiang et al. 2012)	Perpendicularity of the final feature (Jiang et al. 2012)	
		Cylindricity of the final feature (Jiang et al. 2012)	
Quality monitoring	Inspection capability (Shetwan et al. 2011)	Inspection time (Shetwan et al. 2011)	
		Type-I error: producer risk – rejecting good item (Shetwan et al. 2011)	
		Type-II error: consumer risk – accepting non-conforming product (Shetwan et al. 2011)	
	Defective parts (Shetwan et al. 2011)	Non-conforming product (Shetwan et al. 2011)	

state characteristic	state (t=0)	state (t=1)	transformation yes/no ($Xn_{(t=0)} \neq Xn_{(t=1)}$) n: 1....i	relevant state characteristics (target area 1)
X1	X1	X1	+	X1
X2	X2	X2	-	-
X3	X3	X3	+	X3
X4	X4	X4	+	X4
X5	X5	X5	-	-
:	:	:	:	:
Xi	Xi	Xi	+	Xi

Figure 120: Relevant state characteristics (target area 1)

state characteristic	target state characteristics					relevant state characteristics (target area 2)
	X1	X2	X3	X4	X5	
X1	+	+	+	+	-	X1
X2	+	-	+	-	-	X2
X3	-	-	-	-	-	-
X4	+	-	-	-	-	X4
X5	-	-	-	-	-	-
:	:	:	:	:	:	:
Xi	-	+	+	-	-	Xi

Figure 121: Relevant state characteristics (target area 2)

state characteristic	process parameters					relevant state characteristics (target area 3)
	Y1	Y2	Y3	Y4	Y5	
X1	-	-	-	-	-	-
X2	+	-	-	+	-	X2
X3	-	+	-	-	-	X3
X4	+	+	-	+	-	X4
X5	-	-	-	-	-	-
:	:	:	:	:	:	:
Xi	-	-	-	-	-	-

Figure 122: Relevant state characteristics (target area 3)

state characteristic	relevant state characteristics of			relevant state characteristics (combined)
	target area 1	target area 2	target area 3	
X1	X1	-	-	X1
X2	X2	X2	-	X2
X3	-	-	-	-
X4	-	-	-	-
X5	X5	X5	X5	X5
X6			X6	X6
X7	X7	X7	-	X7
X8	X8	X8	X8	X8
X9	-	-	-	-
X10	-	-	X10	X10
X11	-	-	-	-
:	:	:	:	:
Xi	Xi	Xi	-	Xi

Figure 123: Combining relevant state characteristics of the three target areas in stage 2

Table 12: Specific advantages of ML algorithms in manufacturing appl. derived from literature

Advantage	ML algorithm	Application area	References
more complex knowledge bases can be built in shorter time with less engagement of experts	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
successful supporting sensor integration, signal processing, uncertainty handling, real-time and adaptive functioning	ANN [Artificial Neural Networks]	Intelligent manufacturing	(Monostori, 2003)
efficient classification of new problem instances with unknown classes and represent domain concepts in a compact and transparent way suitable for human inspection presenting new insights	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
machine learning analysis has proved beneficial in detecting inconsistencies and suggesting corrections of existing prescriptions	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
machine learning methods [...] support transfer of relevant information to the technology planning level	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
savings in both the operating time and the investments into wheel stock	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
fast (compared to other techniques), simple and their generated models are easy to understand	Inductive learning	Manufacturing	(Pham & Afify, 2005)
successfully employed to detect multi-modal distributions as well as non-linear distributions	novelty detection methods	faulty wafer detection (semiconductor manufacturing)	(Kim, Kang, Cho, Lee & Doh, 2012)
overcame the major problems of statistical process control: linearity and unimodality of data	FDC (Fault Detection and Classification)	faulty wafer detection (semiconductor manufacturing)	(Kim et al., 2012)
No need for assumptions of distribution, and nonlinear problems can be addressed	FDC (Fault Detection and Classification)	faulty wafer detection (semiconductor manufacturing)	(Kim et al., 2012)
formulation of NN or k-NN algorithms is very simple	Nearest neighbour / k-NN	dynamic scheduling in flexible manufacturing systems	(Priore, de la Fuente, Puente & Parreño, 2006)

most well-known and widely used as pattern classifiers and function approximators	backpropagation neural networks	dynamic scheduling in flexible manuf. systems	(Priore et al., 2006)
provide lowest test error	nearest neighbour (k-NN)	dynamic scheduling in flexible manuf. systems	(Priore et al., 2006)
reduce the effort involved in determining the knowledge required to make decisions	ML (inductive learning or neural networks)	dynamic scheduling (semiconductor manuf.)	(Priore, De La Fuente, Gomez & Puente, 2001)
useful when input buffer size is limited and small, and there is a great variation in processing times for parts in the bottleneck machines	inductive learning, C4.5	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
provides the most adequate dispatching rule	back-propagation neural network	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
superior performance	competitive neural networks	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
deals with noise in data more efficiently	inductive learning, C4.5	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
advantage of generating rules that are intelligible to humans (compared with neural networks)	inductive learning, fuzzy logic	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
very effcient at classification, despite their simplicity	case based reasoning	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
Inductive learning classifiers obtain similar and sometimes better accuracies compared with other classification techniques	Inductive learning	manufacturing	(Pham & Afify, 2005)
Rule sets extracted were more accurate and compact than those obtained using its immediate predecessor	RULES-5	manufacturing	(Pham & Afify, 2005)
ability to model complex target concepts and the fact that information present in the training instances is never lost	Instance based methods	manufacturing	(Pham & Afify, 2005)
well-suited to problems in which the training data correspond to noisy, complex sensor data, such as inputs from cameras and microphones	Neural networks	manufacturing	(Pham & Afify, 2005)
wide applicability	Neural networks	manufacturing	(Pham & Afify,

			2005)
potentially greater ability to avoid local minima than is possible with the simple greedy search employed by most learning techniques	Genetic algorithm (GA)	manufacturing	(Pham & Afify, 2005)
natural platform for combining domain knowledge and empirical learning	Bayesian networks	manufacturing	(Pham & Afify, 2005)
improved to handle large data sets efficiently	Decision-tree algorithms	manufacturing	(Pham & Afify, 2005)
flexible and can be applied to a number of other design and manufacturing processes to reduce costs and improve productivity	C4.5 algorithm	manufacturing	(Pham & Afify, 2005)
Adding hidden layers to a feed-forward network enlarges the space of hypotheses that can be represented by the network	neural networks	scheduling	(Mönch, Zimmermann & Otto, 2006)
simple and powerful form of learning algorithms	Inductive decision trees	scheduling	(Mönch et al., 2006)
functional dependencies between input and output variables can be described by rules	inductive decision trees	scheduling	(Mönch et al., 2006)
successful training phase is pssible within seconds	inductive decision trees	scheduling	(Mönch et al., 2006)
computational effort is much smaller by following the machine learning approach	ML (inductive decision trees, neural networks)	scheduling	(Mönch et al., 2006)
CBR with the 'Activity' weighting method had a better prediction rate, outperforming the CBR-alone and all other weighting methods	Hybrid (neural networks and case-based reasoning)	Yield management in semiconductor manufacturing companies	(Lee & Ha, 2009)

Table 13: Challenges & limitations of ML algorithms in manufacturing application

Challenge / limitation	ML algorithm	Application area	References
ML algorithms performance is strongly influenced by inconsistent decisions of operators & different preferences concerning the learning system	learning from examples / decision trees / rules	decision making in machining processes	(Filipic & Junkar, 2000)
Very large optimization problems present major challenge for application of powerful global optimization techniques like GA and as centralized ap-	genetic algorithm	uncertainty, complexity and change in manufacturing	(Monostori, 2003)

proaches, are not totally devoid of all known drawbacks of centralized/hierarchical control systems			
beyond a given problem size even multi-agent approaches may becomes unrealistic, first of all due to the rapidly increasing communication burden	Multi-agent approaches	uncertainty, complexity and change in manufacturing	(Monostori, 2003)
larger batch size leads to a larger error	inductive decision trees, neural networks	scheduling	(Mönch et al., 2006)
a larger number of families leads to poorer results	inductive decision trees, neural networks	scheduling	(Mönch et al., 2006)
time and experience needed to perform optimization of Fuzzy membership functions that correspond to attribute intervals of decision tree	inductive decision trees, neural networks	scheduling	(Mönch et al., 2006)
in practice available attributes often do not contain all the information necessary to unambiguously determine the classes of an example.	Supervised classification	manufacturing	(Pham & Afify, 2005)
inability to handle noisy data	Rule induction (RULE-5 algorithm)	manufacturing	(Pham & Afify, 2005)
cost of classifying new instances can be high	Instance-based learning	manufacturing	(Pham & Afify, 2005)
high computational cost	Genetic algorithm	manufacturing	(Pham & Afify, 2005)
typically consider all attr. of instances when attempting to retrieve similar training instances from memory	Instance-based approaches (nearest neighbor)	manufacturing	(Pham & Afify, 2005)
Difficulty to understand produced models	Neural Networks	manufacturing	(Pham & Afify, 2005)
time-consuming training	Neural Networks	manufacturing	(Pham & Afify, 2005)
inference can have a high time complexity and as tools for classification learning Bayesian networks are not yet as mature or well tested as other approaches	Bayesian networks	manufacturing	(Pham & Afify, 2005)
suffers from feature weighting; when it measures the distance between cases, some features should be weighted differently	case-based reasoning (CBR)	Yield management in semiconductor manufacturing companies	(Lee & Ha, 2009)
assumption that the input data are	Gaussian density	faulty wafer de-	(Kim et al.,

generated from a single Gaussian distribution	estimation	tection (semiconductor manufacturing)	2012)
True Positive Rate (TPR) / False Positive Rate (FPR) ratio influenced by set misclassification cost	novelty detection models (Gaussian density estimation, G. mixture model, Parzen window)	faulty wafer detection (semiconductor manufacturing)	(Kim et al. 2012)
outliers in the data can degrade model performance	novelty detection models (Gaussian density estimation, G. mixture model, Parzen window)	faulty wafer detection (semiconductor manufacturing)	(Kim et al. 2012)
Conventional binary classification models tend to place too much emphasis on majority class	Binary classification algorithms	faulty wafer detection (semiconductor manuf.)	(Kim et al. 2012)
training examples and learning algorithm must be fitting	Knowledge-based systems	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
training set is a subset of the universe of all possible cases	Knowledge-based systems	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
system's performance depends on number and range of control attributes taken into account in the design of training examples	Knowledge-based systems	dynamic scheduling (semiconductor manufacturing)	(Priore et al., 2001); (Priore et al., 2006)
system can be prone to inadequate generalizations in extremely imprecise situations	Knowledge-based systems	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001); (Priore et al., 2006)
no identification of important attributes if not considered initially -> Process must be repeated if performance measurements change	Neural Networks	dynamic scheduling (semiconductor manuf.)	(Priore et al., 2001)
lack of a method to systematically search for an optimum no. of output nodes to the neural network, and that it is compared to a random system rather than the best possible combination of the proposed dispatching rules	competitive neural network	dynamic scheduling (semiconductor manufacturing)	(Priore et al., 2001)
Difficulty of determination of - optimum no. of training examples - adequate monitoring period - mechanism or filter to smooth transitory states	Knowledge-based systems	dynamic scheduling (semiconductor manufacturing)	(Priore et al., 2001)

Figure 124: Screenshot of Cluster3.0 software to create the cluster Dendogram

Table 14: Eliminated examples incl. no. of missing values and label

No. of example	'positive' / 'negative'	No. of missing values
1	negative (Yield 38)	16
172	positive (Yield 39.66)	11
173	positive (Yield 39.68)	11
174	positive (Yield 42.23)	11
175	negative (Yield 38.48)	11
176	positive (Yield 39.49)	11

```
Performance:
PerformanceVector (
------accuracy: 78.24% +/- 1.91% (mikro: 78.24%)
ConfusionMatrix:
True:   Defect140       PASS
Defect140:      1840    655
PASS:   258     1442
------precision: 84.91% +/- 2.52% (mikro: 84.82%) (positive class: PASS)
ConfusionMatrix:
True:   Defect140       PASS
Defect140:      1840    655
PASS:   258     1442
------recall: 68.77% +/- 3.78% (mikro: 68.76%) (positive class: PASS)
ConfusionMatrix:
True:   Defect140       PASS
Defect140:      1840    655
PASS:   258     1442
------AUC (optimistic): 0.834954318049283 +/- 0.015349009539686 (mikro: 0.834954318049283) (positive class: PASS)
------AUC: 0.834954318049283 +/- 0.015349009539686 (mikro: 0.834954318049283) (positive class: PASS)
------AUC (pessimistic): 0.834954318049283 +/- 0.015349009539686 (mikro: 0.834954318049283) (positive class: PASS)
)
SVM.C   = 1.5
SVM.convergence_epsilon= 0.001
```

Figure 125: Optimization results x-val with linear kernel TOM(RR)

Table 15: Feature ranking of TOM (RR) incl. Weight values by RapidMiner (v5.3)

Rank	feature	weight	Rank	feature	weight
1	para.20	1.0	44	para.37	0.19073308720125864
2	para.46	0.9431972103438553	45	para.22	0.17965204384777628
3	para.23	0.8464995383086241	46	para.35	0.17520430381272134
4	para.50	0.8352833142593754	47	para.2	0.1657479320668089
5	para.33	0.8229103820488326	48	para.8	0.16169403401941393
6	para.56	0.7756820916332756	49	para.13	0.15176382850956638
7	para.32	0.7675416276804052	50	para.55	0.14283514298535313
8	para.60	0.749193123828978	51	para.59	0.14283514298535313
9	para.3	0.7406241674830152	52	para.17	0.13840327319873905
10	para.16	0.7341441214201473	53	para.44	0.13125343691007132
11	para.9	0.7336618695289076	54	para.42	0.12454214027707596
12	para.14	0.7122208751888185	55	para.72	0.11473269596225709
13	para.69	0.7087935762760935	56	para.30	0.10440208602478708
14	para.61	0.6942744565358098	57	para.62	0.10373404877225662
15	para.48	0.5920141474501461	58	para.80	0.09744052824308289
16	para.40	0.5327406977305587	59	para.66	0.09463194343840874
17	para.10	0.5322276525211775	60	para.79	0.09101524842481891
18	para.58	0.5232503308553361	61	para.76	0.09053871706160208
19	para.7	0.5088399235689387	62	para.87	0.08240904343515952
20	para.6	0.4860916337002901	63	para.18	0.08033421039918315
21	para.11	0.48149353056291866	64	para.75	0.07496266780365791
22	para.24	0.4799253897020237	65	para.41	0.0693218455939779
23	para.54	0.4711020118823952	66	para.4	0.06924983973843427
24	para.36	0.46361412755010806	67	para.31	0.06496569971664111
25	para.38	0.4485329360291879	68	para.49	0.06112259066582686
26	para.81	0.42309516223387544	69	para.57	0.058928097142938694
27	para.47	0.4010679323978458	70	para.70	0.05429731076153086
28	para.28	0.39988480752979627	71	para.63	0.051576110138954995
29	para.65	0.3882502167171662	72	para.53	0.04808178170191992
30	para.73	0.37907180557897474	73	para.86	0.047763036285393926
31	para.25	0.3760098002890117	74	para.83	0.04511856973150561
32	para.64	0.3542035417456957	75	para.27	0.04473973585514782
33	para.39	0.3343960071837557	76	para.34	0.04362704328448821
34	para.77	0.3332916203012297	77	para.45	0.04346133182092924
35	para.19	0.32441843773640944	78	para.15	0.03721077646461305
36	para.51	0.2793760367487238	79	para.78	0.021039814030233365
37	para.29	0.24360210981477157	80	para.74	0.018627141845566384
38	para.52	0.23672616359340248	81	para.43	0.010050237256845713
39	para.26	0.23538699948122366	82	para.67	0.008891523120404967
40	para.5	0.2308081505950978	83	para.82	0.007684004320234522
41	para.68	0.22717347818444095	84	para.21	0.007231314510371265
42	para.12	0.19655264733132205	85	para.84	0.0
43	para.71	0.19401523378644578			

Table 16: Eliminated examples/vectors with more than 6% of missing values (SECOM data set)

Example No.	Quality ass.	Time-stamp	Missing values	Example No.	Quality ass.	Time-stamp	Missing values	Example No.	Quality ass.	Time-stamp	Missing values
example 1	-1	19.07.08 11:55	44	example 141	-1	06.08.08 19:24	40	example 270	-1	18.08.08 15:30	44
example 2	-1	19.07.08 12:32	36	example 143	-1	06.08.08 20:17	36	example 273	-1	18.08.08 16:19	44
example 5	-1	19.07.08 15:22	40	example 146	-1	06.08.08 23:40	36	example 283	1	19.08.08 03:59	44
example 7	-1	19.07.08 19:44	40	example 152	-1	07.08.08 08:55	52	example 284	-1	19.08.08 04:58	36
example 18	-1	22.07.08 08:41	40	example 166	-1	09.08.08 02:37	36	example 285	-1	19.08.08 05:09	48
example 21	-1	22.07.08 15:30	48	example 174	-1	09.08.08 21:07	40	example 293	-1	19.08.08 06:24	36
example 27	-1	27.07.08 11:10	44	example 176	-1	09.08.08 22:37	40	example 297	-1	19.08.08 08:07	36
example 51	1	29.07.08 18:08	44	example 178	-1	09.08.08 23:34	36	example 298	-1	19.08.08 08:33	36
example 54	-1	29.07.08 23:19	47	example 184	-1	10.08.08 11:16	43	example 300	-1	19.08.08 09:05	80
example 60	-1	31.07.08 13:57	36	example 186	-1	10.08.08 12:22	36	example 301	-1	19.08.08 10:46	36
example 75	-1	03.08.08 14:25	44	example 189	1	10.08.08 20:07	36	example 304	-1	19.08.08 11:13	36
example 90	-1	04.08.08 16:15	87	example 196	-1	12.08.08 04:23	60	example 308	-1	19.08.08 11:57	40
example 94	-1	04.08.08 18:24	96	example 199	-1	12.08.08 11:29	44	example 309	-1	19.08.08 11:58	44
example 95	-1	04.08.08 19:58	68	example 202	-1	15.08.08 03:26	48	example 310	-1	19.08.08 12:30	36
example 96	-1	04.08.08 20:32	96	example 205	-1	15.08.08 09:38	40	example 317	-1	19.08.08 20:53	64
example 98	-1	04.08.08 21:43	36	example 215	-1	16.08.08 06:52	36	example 318	-1	20.08.08 00:05	36
example 99	-1	04.08.08 22:51	48	example 236	1	17.08.08 21:26	44	example 322	1	20.08.08 02:27	40
example 101	-1	05.08.08 00:04	36	example 239	1	17.08.08 23:04	36	example 323	-1	20.08.08 03:00	40
example 102	-1	05.08.08 01:12	40	example 246	-1	18.08.08 04:12	36	example 326	-1	20.08.08 07:12	36
example 104	-1	05.08.08 02:36	36	example 247	-1	18.08.08 04:35	36	example 327	1	20.08.08 08:40	40
example 115	-1	05.08.08 07:12	36	example 249	-1	18.08.08 05:25	36	example 328	1	20.08.08 09:17	40
example 119	-1	05.08.08 09:48	40	example 252	-1	18.08.08 06:26	36	example 330	-1	20.08.08 16:08	48
example 134	-1	06.08.08 09:57	84	example 253	-1	18.08.08 07:11	44	example 331	-1	20.08.08 16:16	36
example 135	-1	06.08.08 12:33	40	example 258	-1	18.08.08 10:13	36	example 336	-1	20.08.08 18:35	44
example 136	-1	06.08.08 13:35	40	example 263	-1	18.08.08 12:49	44	example 337	1	20.08.08 19:20	64
example 138	-1	06.08.08 18:00	40	example 266	-1	18.08.08 14:04	40	example 342	-1	20.08.08 22:13	36

Example No.	Quality ass.	Time-stamp	Missing values	Example No.	Quality ass.	Time-stamp	Missing values	Example No.	Quality ass.	Time-stamp	Missing values
example 345	1	20.08.08 23:43	44	example 428	-1	22.08.08 23:49	40	example 495	-1	28.08.08 20:20	44
example 347	-1	21.08.08 02:46	48	example 433	-1	23.08.08 04:52	44	example 498	-1	28.08.08 21:52	44
example 359	-1	21.08.08 13:11	40	example 436	-1	23.08.08 05:58	36	example 500	-1	28.08.08 23:40	36
example 363	-1	21.08.08 14:47	36	example 442	1	23.08.08 15:29	44	example 501	-1	28.08.08 23:42	48
example 370	-1	21.08.08 16:36	40	example 450	-1	25.08.08 09:29	36	example 506	-1	29.08.08 03:27	48
example 373	-1	21.08.08 18:04	40	example 451	-1	27.08.08 00:46	40	example 512	-1	29.08.08 05:54	100
example 374	1	21.08.08 18:05	40	example 455	-1	27.08.08 14:18	44	example 513	-1	29.08.08 05:54	84
example 375	-1	21.08.08 18:39	36	example 458	-1	27.08.08 22:58	36	example 514	-1	29.08.08 05:57	36
example 376	-1	21.08.08 18:53	44	example 459	-1	28.08.08 03:03	36	example 516	-1	29.08.08 06:35	44
example 383	-1	21.08.08 21:32	36	example 460	-1	28.08.08 03:04	56	example 522	-1	29.08.08 07:22	36
example 384	-1	21.08.08 22:25	40	example 461	-1	28.08.08 03:29	36	example 523	-1	29.08.08 07:24	44
example 385	-1	21.08.08 22:41	44	example 464	-1	28.08.08 04:10	40	example 524	-1	29.08.08 07:33	36
example 387	-1	21.08.08 23:27	36	example 465	-1	28.08.08 04:20	36	example 527	-1	29.08.08 08:20	44
example 389	-1	21.08.08 23:57	40	example 466	-1	28.08.08 04:54	44	example 529	-1	29.08.08 08:45	48
example 392	-1	22.08.08 00:47	36	example 467	-1	28.08.08 05:00	36	example 531	-1	29.08.08 12:22	36
example 398	-1	22.08.08 02:25	48	example 469	-1	28.08.08 06:39	44	example 533	-1	29.08.08 13:04	36
example 400	-1	22.08.08 03:18	36	example 470	-1	28.08.08 06:56	36	example 534	-1	29.08.08 13:14	48
example 404	-1	22.08.08 05:21	36	example 477	-1	28.08.08 10:28	39	example 535	-1	29.08.08 13:27	48
example 406	-1	22.08.08 05:32	40	example 478	-1	28.08.08 10:32	36	example 538	-1	29.08.08 14:18	36
example 407	1	22.08.08 06:00	36	example 479	-1	28.08.08 11:19	39	example 539	-1	29.08.08 14:30	40
example 408	-1	22.08.08 07:04	36	example 484	-1	28.08.08 16:43	48	example 540	-1	29.08.08 15:01	36
example 414	-1	22.08.08 10:18	36	example 487	-1	28.08.08 17:27	44	example 541	-1	29.08.08 15:43	44
example 418	-1	22.08.08 12:37	36	example 488	-1	28.08.08 17:32	36	example 542	-1	29.08.08 16:26	44
example 421	-1	22.08.08 14:45	36	example 492	-1	28.08.08 18:25	40	example 543	-1	29.08.08 20:21	44
example 422	-1	22.08.08 15:25	40	example 493	-1	28.08.08 18:47	36	example 546	-1	29.08.08 21:46	40
example 427	-1	22.08.08 23:47	40	example 494	-1	28.08.08 18:57	40	example 550	-1	29.08.08 22:49	40

Example No.	Quality ass.	Time-stamp	Missing values
example 552	-1	30.08.08 00:01	36
example 553	-1	30.08.08 00:05	40
example 554	-1	30.08.08 00:57	36
example 555	-1	30.08.08 01:29	36
example 556	-1	30.08.08 02:08	36
example 557	-1	30.08.08 02:22	48
example 558	-1	30.08.08 02:32	36
example 559	-1	30.08.08 02:39	36
example 564	-1	30.08.08 04:55	44
example 565	-1	30.08.08 05:05	48
example 567	-1	30.08.08 05:38	44
example 571	-1	30.08.08 08:10	48
example 575	-1	30.08.08 09:13	36
example 579	-1	30.08.08 10:16	48
example 582	-1	30.08.08 11:18	40
example 583	-1	30.08.08 11:57	36
example 585	-1	30.08.08 14:23	36
example 586	-1	30.08.08 14:37	44
example 587	-1	30.08.08 15:06	36
example 590	-1	30.08.08 15:50	51
example 591	-1	30.08.08 15:57	40
example 593	-1	30.08.08 17:05	40
example 595	-1	30.08.08 19:21	44
example 598	-1	30.08.08 20:53	48
example 600	-1	30.08.08 23:51	36
example 601	-1	31.08.08 02:57	40
example 602	1	31.08.08 04:46	48
example 605	-1	31.08.08 10:36	36
example 606	1	31.08.08 10:59	36
example 608	-1	31.08.08 11:03	40
example 610	-1	31.08.08 15:13	36
example 611	-1	31.08.08 16:32	44
example 612	-1	31.08.08 20:24	36
example 614	-1	31.08.08 21:46	40
example 615	-1	31.08.08 21:58	36
example 617	-1	31.08.08 22:48	40
example 620	-1	01.09.08 00:39	40
example 621	-1	01.09.08 00:45	36
example 624	-1	01.09.08 05:32	39
example 627	-1	01.09.08 06:52	36
example 628	-1	01.09.08 08:18	36
example 635	1	01.09.08 19:54	40
example 636	-1	01.09.08 20:51	40
example 637	-1	01.09.08 22:05	40
example 638	-1	01.09.08 23:05	40
example 641	-1	01.09.08 23:28	40
example 643	-1	01.09.08 23:45	40
example 647	-1	02.09.08 00:46	44
example 648	-1	02.09.08 01:09	44
example 650	-1	02.09.08 01:10	40
example 652	-1	02.09.08 01:50	40
example 656	-1	02.09.08 02:52	39
example 657	-1	02.09.08 03:33	36
example 659	-1	02.09.08 03:35	48
example 660	-1	02.09.08 04:19	48
example 662	-1	02.09.08 05:46	48
example 663	-1	02.09.08 06:03	40
example 664	-1	02.09.08 06:19	44
example 666	-1	02.09.08 06:48	43
example 667	-1	02.09.08 07:22	36
example 671	-1	02.09.08 09:36	44
example 676	-1	02.09.08 10:33	40
example 677	-1	02.09.08 11:08	52
example 679	-1	02.09.08 11:47	40
example 680	-1	02.09.08 11:49	48
example 681	-1	02.09.08 12:01	52
example 683	-1	02.09.08 12:26	36
example 684	-1	02.09.08 12:30	52
example 685	-1	02.09.08 12:44	36
example 686	-1	02.09.08 13:05	36
example 687	-1	02.09.08 13:08	44
example 688	-1	02.09.08 13:27	40
example 690	-1	02.09.08 14:09	44
example 697	-1	02.09.08 16:58	44
example 698	-1	02.09.08 17:10	48
example 701	-1	02.09.08 18:52	84
example 702	-1	02.09.08 19:18	44
example 703	-1	02.09.08 20:18	36

Example No.	Quality ass.	Time-stamp	Missing values
example 707	-1	03.09.08 00:18	40
example 708	-1	03.09.08 00:45	39
example 710	1	03.09.08 01:15	48
example 719	-1	03.09.08 18:16	36
example 721	-1	03.09.08 20:10	40
example 722	-1	04.09.08 08:01	40
example 729	-1	04.09.08 14:11	36
example 731	-1	04.09.08 16:30	36
example 733	-1	04.09.08 16:56	48
example 734	-1	04.09.08 17:30	40
example 736	-1	05.09.08 21:48	88
example 738	-1	07.09.08 00:12	40
example 744	-1	07.09.08 18:02	36
example 751	-1	07.09.08 23:33	40
example 752	-1	08.09.08 00:32	40
example 753	-1	08.09.08 00:40	72
example 765	-1	08.09.08 19:33	40
example 767	-1	08.09.08 21:36	40
example 769	1	08.09.08 22:17	40
example 780	-1	11.09.08 00:17	44
example 786	-1	11.09.08 09:24	40
example 787	-1	11.09.08 12:47	40
example 792	-1	11.09.08 15:05	36
example 794	-1	11.09.08 16:28	36
example 795	-1	11.09.08 21:13	36
example 802	-1	12.09.08 06:41	48
example 808	-1	12.09.08 13:08	64
example 810	-1	12.09.08 15:40	36
example 811	-1	12.09.08 16:23	99
example 815	-1	12.09.08 20:28	96
example 828	-1	13.09.08 12:48	36
example 830	-1	13.09.08 16:04	40
example 832	1	13.09.08 20:06	48
example 841	-1	14.09.08 19:13	36
example 842	-1	14.09.08 20:00	36
example 847	-1	15.09.08 02:04	84
example 849	-1	15.09.08 11:40	36
example 855	-1	15.09.08 22:13	40
example 857	-1	15.09.08 22:54	36
example 862	-1	16.09.08 08:52	44
example 866	-1	16.09.08 18:23	40
example 867	-1	16.09.08 19:34	44
example 874	-1	17.09.08 07:28	36
example 878	-1	18.09.08 01:43	36
example 885	1	18.09.08 10:48	36
example 896	-1	19.09.08 01:57	36
example 921	-1	19.09.08 20:05	44
example 924	-1	19.09.08 20:17	36
example 929	-1	19.09.08 21:14	68
example 949	-1	20.09.08 12:36	44
example 951	-1	20.09.08 14:00	36
example 996	-1	21.09.08 18:12	92
example 1008	-1	22.09.08 10:43	40
example 1012	-1	22.09.08 13:03	36
example 1016	-1	22.09.08 16:25	36
example 1055	-1	24.09.08 07:23	92
example 1076	-1	25.09.08 09:31	40
example 1140	-1	27.09.08 20:46	36
example 1141	-1	27.09.08 21:45	52
example 1153	-1	28.09.08 17:05	100
example 1207	-1	30.09.08 23:34	76
example 1235	-1	02.10.08 07:58	68
example 1356	-1	06.10.08 08:57	36
example 1372	-1	06.10.08 17:26	36
example 1449	-1	10.10.08 12:30	36
example 1451	-1	10.10.08 15:50	40
example 1460	-1	11.10.08 07:42	44
example 1462	-1	11.10.08 14:43	36
example 1468	-1	12.10.08 01:50	36
example 1470	-1	12.10.08 03:21	40
example 1543	1	15.10.08 22:54	52
example 1544	-1	15.10.08 23:00	52
example 1545	-1	15.10.08 23:45	48
example 1546	-1	16.10.08 02:16	48
example 1547	-1	16.10.08 02:16	48
example 1548	-1	16.10.08 02:17	60
example 1549	-1	16.10.08 02:22	52
example 1550	-1	16.10.08 02:55	56

Example No.	Quality ass.	Time-stamp	Missing values
example 1552	-1	16.10.08 04:02	60
example 1553	-1	16.10.08 04:02	60
example 1554	-1	16.10.08 04:04	52
example 1555	-1	16.10.08 04:47	60
example 1556	-1	16.10.08 04:50	60
example 1557	-1	16.10.08 04:54	56
example 1558	-1	16.10.08 05:08	48
example 1559	-1	16.10.08 05:13	52
example 1560	-1	16.10.08 05:44	48
example 1561	-1	16.10.08 05:58	60
example 1562	-1	16.10.08 15:02	140
example 1563	-1	16.10.08 15:13	52
example 1564	-1	16.10.08 20:49	52
example 1565	-1	17.10.08 05:26	148
example 1566	-1	17.10.08 06:01	60
example 1567	-1	17.10.08 06:07	152

Table 17: Eliminated features with missing values on reduced SECOM data set 1239 examples

Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'
feature 1	5	feature 143	1	feature 358	7	feature 544	1
feature 2	4	feature 144	2	feature 359	1027	feature 545	1
feature 3	1	feature 156	9	feature 363	42	feature 546	1
feature 4	1	feature 158	1119	feature 364	42	feature 547	47
feature 5	1	feature 159	1119	feature 383	729	feature 548	47
feature 7	1	feature 220	7	feature 384	729	feature 549	47
feature 8	2	feature 221	1027	feature 385	729	feature 550	47
feature 20	9	feature 225	42	feature 386	592	feature 551	47
feature 41	17	feature 226	42	feature 409	1	feature 552	47
feature 42	17	feature 245	729	feature 410	5	feature 553	47
feature 73	612	feature 246	729	feature 411	4	feature 554	47
feature 74	612	feature 247	729	feature 412	1	feature 555	47
feature 85	7	feature 248	592	feature 413	1	feature 556	47
feature 86	1027	feature 271	1	feature 414	1	feature 557	47
feature 90	42	feature 272	5	feature 416	1	feature 558	47
feature 91	42	feature 273	4	feature 417	2	feature 563	129
feature 110	729	feature 274	1	feature 429	9	feature 564	129
feature 111	729	feature 275	1	feature 492	7	feature 565	129
feature 112	729	feature 276	1	feature 493	1027	feature 566	129
feature 113	592	feature 278	1	feature 497	42	feature 567	129
feature 136	1	feature 279	2	feature 498	42	feature 568	129
feature 137	5	feature 291	9	feature 517	729	feature 569	129
feature 138	4	feature 293	1119	feature 518	729	feature 570	129
feature 139	1	feature 294	1119	feature 519	729	feature 579	703
feature 140	1	feature 346	612	feature 520	592	feature 580	703
feature 141	1	feature 347	612	feature 543	1	feature 581	703
						feature 582	703

Table 18: Eliminated features containing more than 5 (>5) missing values on reduced SECOM data set (1239 examples)

Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'	Feature No.	Σ 'NaN'
feature 20	9	feature 225	42	feature 386	592	feature 554	47
feature 41	17	feature 226	42	feature 429	9	feature 555	47
feature 42	17	feature 245	729	feature 492	7	feature 556	47
feature 73	612	feature 246	729	feature 493	1027	feature 557	47
feature 74	612	feature 247	729	feature 497	42	feature 558	47
feature 85	7	feature 248	592	feature 498	42	feature 563	129
feature 86	1027	feature 291	9	feature 517	729	feature 564	129
feature 90	42	feature 293	1119	feature 518	729	feature 565	129
feature 91	42	feature 294	1119	feature 519	729	feature 566	129
feature 110	729	feature 346	612	feature 520	592	feature 567	129
feature 111	729	feature 347	612	feature 547	47	feature 568	129
feature 112	729	feature 358	7	feature 548	47	feature 569	129
feature 113	592	feature 359	1027	feature 549	47	feature 570	129
feature 156	9	feature 363	42	feature 550	47	feature 579	703
feature 158	1119	feature 364	42	feature 551	47	feature 580	703
feature 159	1119	feature 383	729	feature 552	47	feature 581	703
feature 220	7	feature 384	729	feature 553	47	feature 582	703
feature 221	1027	feature 385	729				

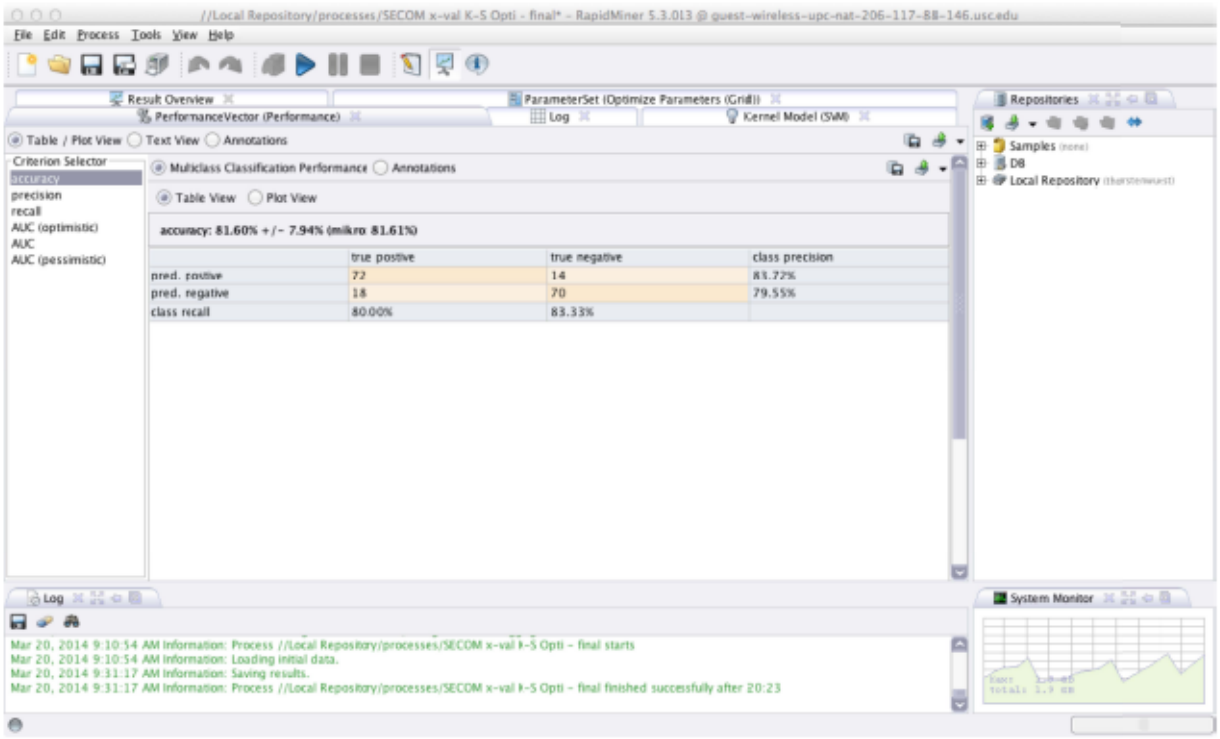

Figure 126: SECOM data set x-val result (accuracy)

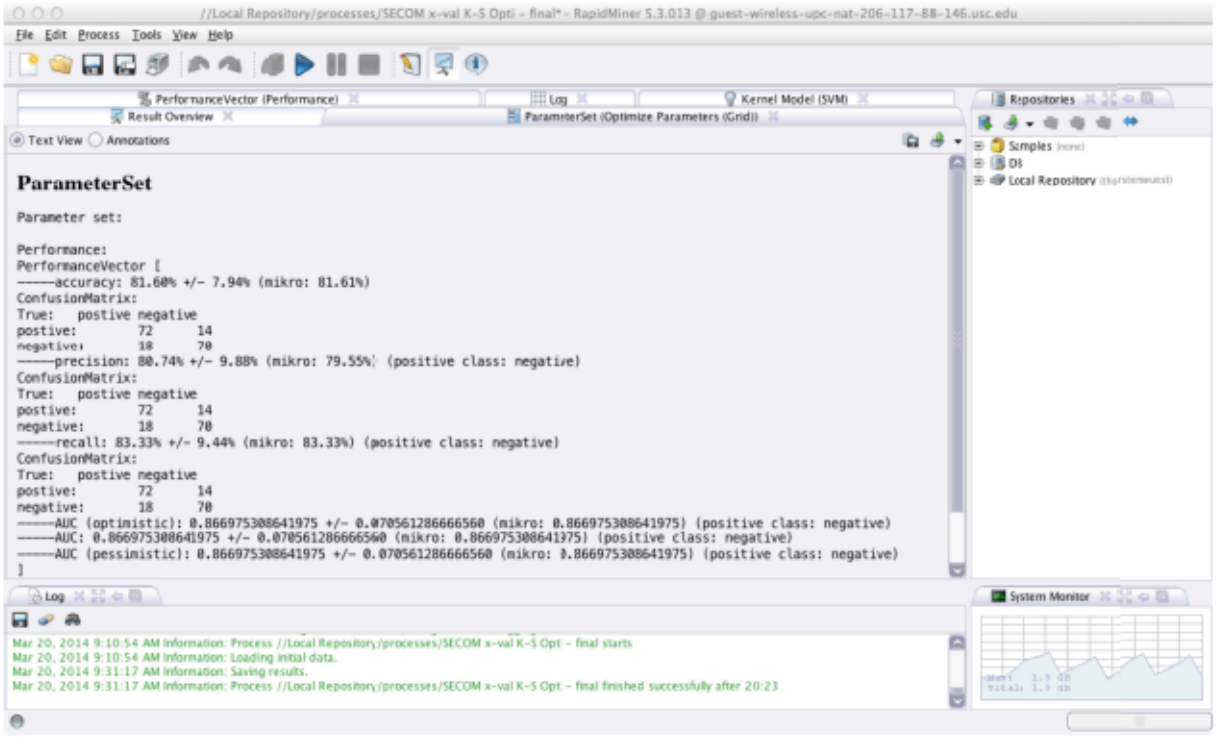

Figure 127: SECOM x-val results (parameters – last optimization cycle)

Table 19: Feature Ranking RR Manufacturing Programme (complete)

Rank	TOM(RR)	Rank	TOM(RR)	Rank	DICK(RR)	Rank	DICK(RR)	Rank	HARRY(RR)	Rank	HARRY(RR)
1	para.51	51	para.37	1	para.DICK.29	51	para.DICK.14	1	para.HARRY.43	51	para.HARRY.22
2	para.21	52	para.17	2	para.DICK.7	52	para.DICK.38	2	para.HARRY.33	52	para.HARRY.14
3	para.50	53	para.26	3	para.DICK.41	53	para.DICK.4	3	para.HARRY.24	53	para.HARRY.34
4	para.33	54	para.52	4	para.DICK.12	54	para.DICK.16	4	para.HARRY.32	54	para.HARRY.1
5	para.6	55	para.39	5	para.DICK.27	55	para.DICK.3	5	para.HARRY.47	55	para.HARRY.48
6	para.36	56	para.87	6	para.DICK.21	56	para.DICK.10	6	para.HARRY.42	56	para.HARRY.18
7	para.14	57	para.8	7	para.DICK.48			7	para.HARRY.58	57	para.HARRY.10
8	para.9	58	para.12	8	para.DICK.31			8	para.HARRY.30	58	para.HARRY.5
9	para.47	59	para.10	9	para.DICK.9			9	para.HARRY.67	59	para.HARRY.6
10	para.59	60	para.27	10	para.DICK.56			10	para.HARRY.11	60	para.HARRY.38
11	para.29	61	para.15	11	para.DICK.17			11	para.HARRY.2	61	para.HARRY.53
12	para.55	62	para.18	12	para.DICK.23			12	para.HARRY.52	62	para.HARRY.46
13	para.61	63	para.28	13	para.DICK.37			13	para.HARRY.36	63	para.HARRY.4
14	para.60	64	para.63	14	para.DICK.51			14	para.HARRY.64	64	para.HARRY.26
15	para.44	65	para.67	15	para.DICK.22			15	para.HARRY.65	65	para.HARRY.44
16	para.45	66	para.66	16	para.DICK.28			16	para.HARRY.39	66	para.HARRY.16
17	para.34	67	para.56	17	para.DICK.34			17	para.HARRY.45	67	para.HARRY.3
18	para.32	68	para.30	18	para.DICK.13			18	para.HARRY.59	68	para.HARRY.35
19	para.64	69	para.40	19	para.DICK.5			19	para.HARRY.55		
20	para.2	70	para.77	20	para.DICK.6			20	para.HARRY.60		
21	para.5	71	para.79	21	para.DICK.55			21	para.HARRY.13		
22	para.35	72	para.65	22	para.DICK.45			22	para.HARRY.19		
23	para.46	73	para.69	23	para.DICK.1			23	para.HARRY.28		
24	para.48	74	para.71	24	para.DICK.36			24	para.HARRY.50		
25	para.31	75	para.4	25	para.DICK.52			25	para.HARRY.37		
26	para.38	76	para.7	26	para.DICK.2			26	para.HARRY.23		
27	para.11	77	para.72	27	para.DICK.11			27	para.HARRY.17		
28	para.42	78	para.74	28	para.DICK.42			28	para.HARRY.56		
29	para.20	79	para.70	29	para.DICK.47			29	para.HARRY.9		
30	para.16	80	para.80	30	para.DICK.32			30	para.HARRY.15		
31	para.13	81	para.82	31	para.DICK.43			31	para.HARRY.31		
32	para.84	82	para.86	32	para.DICK.33			32	para.HARRY.49		
33	para.53	83	para.68	33	para.DICK.24			33	para.HARRY.40		
34	para.3	84	para.78	34	para.DICK.53			34	para.HARRY.66		
35	para.43	85	para.76	35	para.DICK.46			35	para.HARRY.7		
36	para.22			36	para.DICK.30			36	para.HARRY.29		
37	para.62			37	para.DICK.26			37	para.HARRY.62		
38	para.19			38	para.DICK.39			38	para.HARRY.57		
39	para.54			39	para.DICK.44			39	para.HARRY.41		
40	para.57			40	para.DICK.19			40	para.HARRY.12		
41	para.81			41	para.DICK.50			41	para.HARRY.27		
42	para.75			42	para.DICK.15			42	para.HARRY.68		
43	para.49			43	para.DICK.49			43	para.HARRY.21		
44	para.58			44	para.DICK.40			44	para.HARRY.25		
45	para.23			45	para.DICK.18			45	para.HARRY.8		
46	para.24			46	para.DICK.35			46	para.HARRY.54		
47	para.25			47	para.DICK.25			47	para.HARRY.20		
48	para.83			48	para.DICK.8			48	para.HARRY.61		
49	para.73			49	para.DICK.54			49	para.HARRY.63		
50	para.41			50	para.DICK.20			50	para.HARRY.51		

Rank	TD(RR)	Rank	TD(RR)	Rank	TD(RR)
1	para.DICK.43	51	para.82	101	para.40
2	para.DICK.32	52	para.DICK.34	102	para.46
3	para.DICK.47	53	para.66	103	para.51
4	para.DICK.42	54	para.53	104	para.54
5	para.DICK.33	55	para.DICK.18	105	para.19
6	para.DICK.24	56	para.64	106	para.15
7	para.DICK.11	57	para.DICK.5	107	para.11
8	para.DICK.2	58	para.DICK.6	108	para.3
9	para.DICK.52	59	para.17	109	para.61
10	para.DICK.36	60	para.5	110	para.41
11	para.DICK.1	61	para.84	111	para.52
12	para.26	62	para.DICK.25	112	para.21
13	para.DICK.45	63	para.DICK.10	113	para.9
14	para.DICK.55	64	para.DICK.8	114	para.73
15	para.DICK.30	65	para.DICK.54	115	para.58
16	para.DICK.13	66	para.DICK.20	116	para.36
17	para.DICK.28	67	para.31	117	para.47
18	para.DICK.37	68	para.DICK.14	118	para.50
19	para.62	69	para.DICK.53	119	para.71
20	para.6	70	para.DICK.46	120	para.33
21	para.DICK.23	71	para.8	121	para.38
22	para.DICK.17	72	para.DICK.26	122	para.35
23	para.70	73	para.DICK.44	123	para.86
24	para.DICK.56	74	para.DICK.38	124	para.69
25	para.DICK.9	75	para.18	125	para.2
26	para.DICK.39	76	para.DICK.4	126	para.23
27	para.DICK.31	77	para.25	127	para.37
28	para.45	78	para.79	128	para.34
29	para.DICK.19	79	para.63	129	para.32
30	para.DICK.50	80	para.81	130	para.59
31	para.DICK.7	81	para.43	131	para.55
32	para.DICK.29	82	para.DICK.16	132	para.56
33	para.DICK.15	83	para.DICK.35	133	para.30
34	para.DICK.41	84	para.DICK.3	134	para.27
35	para.DICK.12	85	para.77	135	para.28
36	para.DICK.49	86	para.10	136	para.13
37	para.68	87	para.87	137	para.14
38	para.7	88	para.83	138	para.24
39	para.DICK.27	89	para.39	139	para.67
40	para.DICK.21	90	para.48	140	para.57
41	para.DICK.48	91	para.16	141	para.20
42	para.DICK.40	92	para.60		
43	para.74	93	para.42		
44	para.78	94	para.49		
45	para.72	95	para.65		
46	para.DICK.51	96	para.4		
47	para.76	97	para.22		
48	para.80	98	para.75		
49	para.DICK.22	99	para.12		
50	para.44	100	para.29		

Rank	TDH(RR)	Rank	TDH(RR)	Rank	TDH(RR)	Rank	TDH(RR)	Rank	TDH(RR)
1	para.HARRY.67	51	para.HARRY.7	101	para.DICK.54	151	para.36	201	para.5
2	para.DICK.32	52	para.HARRY.29	102	para.HARRY.18	152	para.69	202	para.37
3	para.HARRY.11	53	para.DICK.55	103	para.HARRY.5	153	para.73	203	para.30
4	para.DICK.43	54	para.HARRY.62	104	para.DICK.9	154	para.12	204	para.59
5	para.HARRY.2	55	para.HARRY.64	105	para.HARRY.6	155	para.17	205	para.55
6	para.HARRY.52	56	para.DICK.13	106	para.DICK.51	156	para.2	206	para.47
7	para.DICK.33	57	para.60	107	para.HARRY.68	157	para.9	207	para.56
8	para.HARRY.36	58	para.19	108	para.DICK.20	158	para.51	208	para.27
9	para.DICK.24	59	para.HARRY.57	109	para.DICK.48	159	para.34	209	para.44
10	para.HARRY.1	60	para.DICK.28	110	para.HARRY.53	160	para.53		
11	para.77	61	para.HARRY.41	111	para.DICK.12	161	para.46		
12	para.DICK.26	62	para.DICK.52	112	para.HARRY.46	162	para.86		
13	para.HARRY.45	63	para.HARRY.65	113	para.HARRY.25	163	para.62		
14	para.HARRY.55	64	para.81	114	para.DICK.14	164	para.7		
15	para.DICK.30	65	para.HARRY.12	115	para.HARRY.8	165	para.21		
16	para.HARRY.60	66	para.DICK.53	116	para.HARRY.54	166	para.11		
17	para.DICK.39	67	para.DICK.40	117	para.23	167	para.49		
18	para.HARRY.13	68	para.HARRY.39	118	para.DICK.38	168	para.28		
19	para.DICK.44	69	para.HARRY.27	119	para.HARRY.20	169	para.58		
20	para.HARRY.42	70	para.DICK.37	120	para.6	170	para.3		
21	para.DICK.47	71	para.HARRY.59	121	para.HARRY.14	171	para.64		
22	para.HARRY.47	72	para.HARRY.19	122	para.43	172	para.52		
23	para.HARRY.32	73	para.DICK.18	123	para.DICK.27	173	para.87		
24	para.DICK.19	74	para.HARRY.50	124	para.HARRY.26	174	para.26		
25	para.HARRY.28	75	para.13	125	para.DICK.21	175	para.15		
26	para.DICK.50	76	para.DICK.6	126	para.29	176	para.16		
27	para.HARRY.37	77	para.DICK.25	127	para.DICK.41	177	para.66		
28	para.DICK.46	78	para.HARRY.15	128	para.DICK.4	178	para.32		
29	para.24	79	para.8	129	para.20	179	para.41		
30	para.HARRY.43	80	para.HARRY.21	130	para.63	180	para.45		
31	para.DICK.42	81	para.65	131	para.HARRY.38	181	para.39		
32	para.HARRY.23	82	para.DICK.5	132	para.DICK.31	182	para.68		
33	para.HARRY.33	83	para.HARRY.48	133	para.HARRY.44	183	para.40		
34	para.DICK.11	84	para.HARRY.61	134	para.14	184	para.70		
35	para.HARRY.24	85	para.DICK.10	135	para.83	185	para.25		
36	para.DICK.2	86	para.38	136	para.DICK.29	186	para.4		
37	para.HARRY.17	87	para.HARRY.49	137	para.DICK.7	187	para.72		
38	para.HARRY.58	88	para.DICK.8	138	para.33	188	para.57		
39	para.DICK.15	89	para.DICK.34	139	para.31	189	para.35		
40	para.75	90	para.HARRY.63	140	para.67	190	para.48		
41	para.HARRY.56	91	para.HARRY.51	141	para.HARRY.4	191	para.18		
42	para.DICK.36	92	para.DICK.23	142	para.DICK.16	192	para.82		
43	para.HARRY.9	93	para.HARRY.22	143	para.50	193	para.80		
44	para.HARRY.31	94	para.DICK.17	144	para.HARRY.16	194	para.74		
45	para.DICK.35	95	para.HARRY.34	145	para.54	195	para.84		
46	para.DICK.49	96	para.79	146	para.DICK.3	196	para.76		
47	para.HARRY.30	97	para.DICK.56	147	para.HARRY.3	197	para.78		
48	para.DICK.1	98	para.HARRY.40	148	para.HARRY.35	198	para.10		
49	para.HARRY.66	99	para.HARRY.10	149	para.71	199	para.61		
50	para.DICK.45	100	para.DICK.22	150	para.22	200	para.42		

Table 20: Feature Ranking CHEM Manufacturing Programme (complete)

Rank	TOM(CHEM)	Rank	TOM(CHEM)	Rank	DICK(CHEM)	Rank	HARRY(CHEM	Rank	TD(CHEM)	Rank	TD(CHEM)	Rank	TD(CHEM)	Rank	TDH(CHEM)	Rank	TDH(CHEM)	Rank	TDH(CHEM)
1	BiologicalMaterial07	50	ManufacturingProcess35	1	Parameter 29	1	Parameter29	1	ManufacturingProcess16	50	ParameterI 20	99	ParameterI 40	1	ParameterII29	50	ManufacturingProcess11	99	ParameterI05
2	ManufacturingProcess32	51	ManufacturingProcess16	2	Parameter 38	2	Parameter11	2	ManufacturingProcess05	51	ParameterI 13	100	ParameterI 25	2	ParameterII11	51	ManufacturingProcess44	100	ParameterI33
3	ManufacturingProcess09	52	ManufacturingProcess27	3	Parameter 11	3	Parameter13	3	ManufacturingProcess06	52	ManufacturingProcess34	101	ParameterI 07	3	ParameterII13	52	ManufacturingProcess03	101	BiologicalMaterial04
4	ManufacturingProcess34	53	ManufacturingProcess18	4	Parameter 13	4	Parameter32	4	ManufacturingProcess33	53	ParameterI 21	102	ParameterI 12	4	ParameterII32	53	ManufacturingProcess12	102	ParameterI16
5	ManufacturingProcess13	54	ManufacturingProcess20	5	Parameter 32	5	Parameter02	5	BiologicalMaterial06	54	ParameterI 32	103	ParameterI 26	5	ParameterII02	54	ManufacturingProcess29	103	ParameterI38
6	ManufacturingProcess19	55	ManufacturingProcess26	6	Parameter 02	6	Parameter09	6	ManufacturingProcess42	55	ParameterI 30	104	ParameterI 37	6	ParameterII09	55	ParameterI27	104	ParameterI19
7	ManufacturingProcess30	56	BiologicalMaterial04	7	Parameter 09	7	Parameter28	7	ManufacturingProcess13	56	ParameterI 02	105	ManufacturingProcess17	7	ParameterII28	56	BiologicalMaterial03	105	ParameterI23
8	ManufacturingProcess39	57	ManufacturingProcess17	8	Parameter 45	8	Parameter17	8	BiologicalMaterial12	57	ParameterI 41			8	ManufacturingProcess42	57	BiologicalMaterial09	106	ParameterI13
9	ManufacturingProcess21			9	Parameter 43	9	Parameter31	9	ManufacturingProcess14	58	ManufacturingProcess09			9	ParameterII17	58	ParameterII27	107	ParameterI06
10	ManufacturingProcess01			10	Parameter 28	10	Parameter25	10	ManufacturingProcess22	59	ParameterI 46			10	ParameterII31	59	ManufacturingProcess30	108	ParameterI32
11	BiologicalMaterial05			11	Parameter 34	11	Parameter07	11	ParameterI 38	60	ManufacturingProcess10			11	ParameterII25	60	ParameterI46	109	ParameterI48
12	BiologicalMaterial02			12	Parameter 39	12	Parameter26	12	BiologicalMaterial01	61	ParameterI 09			12	ParameterII07	61	ManufacturingProcess23	110	ParameterI02
13	ManufacturingProcess10			13	Parameter 47	13	Parameter12	13	ManufacturingProcess38	62	ParameterI 08			13	ParameterII26	62	ManufacturingProcess38	111	ParameterI11
14	BiologicalMaterial09			14	Parameter 17	14	Parameter10	14	BiologicalMaterial08	63	ParameterI 45			14	ParameterII12	63	ManufacturingProcess40	112	ParameterI01
15	ManufacturingProcess03			15	Parameter 31	15	Parameter03	15	ManufacturingProcess45	64	ParameterI 14			15	ParameterII10	64	ManufacturingProcess09	113	ParameterI09
16	ManufacturingProcess31			16	Parameter 42	16	Parameter04	16	ManufacturingProcess11	65	ParameterI 43			16	ManufacturingProcess22	65	ParameterI40	114	ParameterI35
17	ManufacturingProcess37			17	Parameter 25	17	Parameter22	17	BiologicalMaterial07	66	ParameterI 36			17	ParameterII03	66	ManufacturingProcess36	115	ParameterI45
18	ManufacturingProcess12			18	Parameter 07	18	Parameter15	18	ManufacturingProcess36	67	ManufacturingProcess41			18	ParameterII04	67	ManufacturingProcess28	116	ParameterI43
19	BiologicalMaterial10			19	Parameter 37	19	Parameter01	19	ManufacturingProcess43	68	ParameterI 28			19	ParameterII22	68	ManufacturingProcess35	117	ParameterI15
20	ManufacturingProcess23			20	Parameter 26	20	Parameter06	20	ManufacturingProcess15	69	ParameterI 44			20	ParameterII15	69	ManufacturingProcess37	118	ParameterI28
21	ManufacturingProcess02			21	Parameter 12	21	Parameter23	21	ManufacturingProcess08	70	ParameterI 05			21	ParameterII01	70	BiologicalMaterial02	119	ParameterI22
22	ManufacturingProcess15			22	Parameter 10	22	Parameter19	22	ManufacturingProcess12	71	ParameterI 34			22	BiologicalMaterial10	71	ManufacturingProcess45	120	ParameterI34
23	ManufacturingProcess38			23	Parameter 03	23	Parameter16	23	BiologicalMaterial10	72	ParameterI 33			23	BiologicalMaterial01	72	ManufacturingProcess32	121	ParameterI04
24	ManufacturingProcess36			24	Parameter 04	24	Parameter05	24	ManufacturingProcess04	73	ParameterI 39			24	ParameterII06	73	ParameterI41	122	ParameterI39
25	ManufacturingProcess14			25	Parameter 22	25	Parameter14	25	ManufacturingProcess02	74	ParameterI 16			25	ParameterII23	74	ParameterI30	123	ParameterI03
26	ManufacturingProcess40			26	Parameter 15	26	Parameter08	26	ManufacturingProcess07	75	ParameterI 19			26	BiologicalMaterial11	75	ManufacturingProcess13	124	ParameterI47
27	BiologicalMaterial12			27	Parameter 35	27	Parameter20	27	ManufacturingProcess20	76	ManufacturingProcess03			27	ManufacturingProcess04	76	ManufacturingProcess24	125	ParameterI10
28	ManufacturingProcess06			28	Parameter 01	28	Parameter24	28	ManufacturingProcess32	77	ManufacturingProcess27			28	ParameterII19	77	ManufacturingProcess05	126	ParameterI17
29	ManufacturingProcess08			29	Parameter 48	29	Parameter18	29	ManufacturingProcess28	78	ManufacturingProcess25			29	ManufacturingProcess34	78	ManufacturingProcess43	127	ManufacturingProcess31
30	ManufacturingProcess24			30	Parameter 06	30	Parameter21	30	BiologicalMaterial04	79	ParameterI 23			30	ManufacturingProcess07	79	ParameterI21	128	ParameterI31
31	ManufacturingProcess22			31	Parameter 23	31	Parameter30	31	ManufacturingProcess24	80	BiologicalMaterial02			31	BiologicalMaterial06	80	ManufacturingProcess06	129	ParameterI29
32	BiologicalMaterial11			32	Parameter 19	32	Parameter27	32	ManufacturingProcess40	81	ParameterI 06			32	BiologicalMaterial05	81	ManufacturingProcess33	130	ParameterI42
33	ManufacturingProcess05			33	Parameter 16			33	ManufacturingProcess30	82	ParameterI 47			33	ManufacturingProcess15	82	ParameterI18	131	ParameterI12
34	ManufacturingProcess44			34	Parameter 33			34	ManufacturingProcess18	83	ParameterI 48			34	ManufacturingProcess21	83	ParameterI24	132	ParameterI26
35	ManufacturingProcess41			35	Parameter 05			35	ManufacturingProcess29	84	ParameterI 27			35	ParameterII16	84	ManufacturingProcess01	133	ParameterI25
36	ManufacturingProcess42			36	Parameter 44			36	ManufacturingProcess19	85	ParameterI 01			36	ParameterII05	85	ParameterI20	134	ParameterI07
37	BiologicalMaterial08			37	Parameter 36			37	BiologicalMaterial09	86	BiologicalMaterial11			37	ParameterII14	86	ManufacturingProcess20	135	ParameterI37
38	ManufacturingProcess11			38	Parameter 14			38	ManufacturingProcess35	87	ParameterI 35			38	ParameterII08	87	ManufacturingProcess27	136	ManufacturingProcess17
39	ManufacturingProcess33			39	Parameter 08			39	BiologicalMaterial05	88	ParameterI 15			39	ManufacturingProcess19	88	ManufacturingProcess16	137	BiologicalMaterial07
40	ManufacturingProcess25			40	Parameter 20			40	ParameterI 29	89	ManufacturingProcess01			40	ParameterII20	89	ManufacturingProcess02		
41	ManufacturingProcess04			41	Parameter 24			41	BiologicalMaterial03	90	ParameterI 22			41	ManufacturingProcess14	90	ManufacturingProcess08		
42	BiologicalMaterial01			42	Parameter 18			42	ManufacturingProcess23	91	ParameterI 17			42	ManufacturingProcess41	91	ParameterI08		
43	ManufacturingProcess29			43	Parameter 21			43	ParameterI 24	92	ParameterI 04			43	ParameterII24	92	ParameterI14		
44	ManufacturingProcess28			44	Parameter 30			44	ManufacturingProcess21	93	ManufacturingProcess26			44	ParameterII18	93	ManufacturingProcess10		
45	BiologicalMaterial03			45	Parameter 41			45	ManufacturingProcess31	94	ParameterI 03			45	ParameterII21	94	ManufacturingProcess25		
46	ManufacturingProcess43			46	Parameter 46			46	ManufacturingProcess44	95	ParameterI 31			46	ManufacturingProcess39	95	ParameterI36		
47	ManufacturingProcess45			47	Parameter 27			47	ParameterI 18	96	ParameterI 42			47	BiologicalMaterial12	96	ManufacturingProcess18		
48	ManufacturingProcess07			48	Parameter 40			48	ParameterI 11	97	ManufacturingProcess37			48	BiologicalMaterial08	97	ManufacturingProcess26		
49	BiologicalMaterial03							49	ManufacturingProcess39	98	ParameterI 10			49	ParameterII30	98	ParameterI44		

Table 21: Feature Ranking SECOM 412 & 528 Part I

Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528
1	feature 171	1	feature 171	1	75	feature 522	75	feature 522	1	149	feature 85	149	feature 85	1	223	feature 272	223	feature 272	1
2	feature 165	2	feature 165	1	76	feature 274	76	feature 274	1	150	feature 4	150	feature 4	1	224	feature 144	224	feature 144	1
3	feature 301	3	feature 301	1	77	feature 176	77	feature 176	1	151	feature 454	151	feature 454	1	225	feature 273	225	feature 273	1
4	feature 250	4	feature 250	1	78	feature 131	78	feature 131	1	152	feature 2	152	feature 2	1	226	feature 167	226	feature 167	1
5	feature 521	5	feature 521	1	79	feature 147	79	feature 147	1	153	feature 541	153	feature 541	1	227	feature 477	227	feature 477	1
6	feature 388	6	feature 388	1	80	feature 559	80	feature 559	1	154	feature 170	154	feature 170	1	228	feature 206	228	feature 206	1
7	feature 574	7	feature 574	1	81	feature 313	81	feature 313	1	155	feature 108	155	feature 108	1	229	feature 271	229	feature 271	1
8	feature 336	8	feature 336	1	82	feature 66	82	feature 66	1	156	feature 303	156	feature 303	1	230	feature 501	230	feature 501	1
9	feature 526	9	feature 526	1	83	feature 412	83	feature 412	1	157	feature 251	157	feature 251	1	231	feature 106	231	feature 106	1
10	feature 425	10	feature 425	1	84	feature 295	84	feature 295	1	158	feature 9	158	feature 9	1	232	feature 528	232	feature 528	1
11	feature 339	11	feature 339	1	85	feature 342	85	feature 342	1	159	feature 213	159	feature 213	1	233	feature 488	233	feature 488	1
12	feature 338	12	feature 338	1	86	feature 204	86	feature 204	1	160	feature 163	160	feature 163	1	234	feature 289	234	feature 289	1
13	feature 60	13	feature 60	1	87	feature 296	87	feature 296	1	161	feature 133	161	feature 133	1	235	feature 411	235	feature 411	1
14	feature 429	14	feature 429	1	88	feature 48	88	feature 48	1	162	feature 185	162	feature 185	1	236	feature 172	236	feature 172	1
15	feature 100	15	feature 100	1	89	feature 426	89	feature 426	1	163	feature 228	163	feature 228	1	237	feature 300	237	feature 300	1
16	feature 349	16	feature 349	1	90	feature 578	90	feature 578	1	164	feature 318	164	feature 318	1	238	feature 575	238	feature 575	1
17	feature 215	17	feature 215	1	91	feature 77	91	feature 77	1	165	feature 89	165	feature 89	1	239	feature 29	239	feature 29	1
18	feature 76	18	feature 76	1	92	feature 589	92	feature 589	1	166	feature 487	166	feature 487	1	240	feature 479	240	feature 479	1
19	feature 472	19	feature 472	1	93	feature 122	93	feature 122	1	167	feature 335	167	feature 335	1	241	feature 270	241	feature 270	1
20	feature 104	20	feature 104	1	94	feature 356	94	feature 356	1	168	feature 353	168	feature 353	1	242	feature 201	242	feature 201	1
21	feature 495	21	feature 495	1	95	feature 305	95	feature 305	1	169	feature 430	169	feature 430	1	243	feature 18	243	feature 18	1
22	feature 203	22	feature 203	1	96	feature 441	96	feature 441	1	170	feature 287	170	feature 287	1	244	feature 136	244	feature 136	1
23	feature 445	23	feature 445	1	97	feature 280	97	feature 280	1	171	feature 312	171	feature 312	1	245	feature 54	245	feature 54	1
24	feature 282	24	feature 282	1	98	feature 109	98	feature 109	1	172	feature 278	172	feature 278	1	246	feature 545	246	feature 545	1
25	feature 34	25	feature 34	1	99	feature 211	99	feature 211	1	173	feature 473	173	feature 473	1	247	feature 35	247	feature 35	1
26	feature 103	26	feature 103	1	100	feature 471	100	feature 471	1	174	feature 197	174	feature 197	1	248	feature 64	248	feature 64	1
27	feature 302	27	feature 302	1	101	feature 135	101	feature 135	1	175	feature 470	175	feature 470	1	249	feature 123	249	feature 123	1
28	feature 461	28	feature 461	1	102	feature 151	102	feature 151	1	176	feature 36	176	feature 36	1	250	feature 304	250	feature 304	1
29	feature 576	29	feature 576	1	103	feature 44	103	feature 44	1	177	feature 392	177	feature 392	1	251	feature 297	251	feature 297	1
30	feature 22	30	feature 22	1	104	feature 13	104	feature 13	1	178	feature 21	178	feature 21	1	252	feature 494	252	feature 494	1
31	feature 174	31	feature 174	1	105	feature 82	105	feature 82	1	179	feature 542	179	feature 542	1	253	feature 254	253	feature 254	1
32	feature 161	32	feature 161	1	106	feature 129	106	feature 129	1	180	feature 153	180	feature 153	1	254	feature 87	254	feature 87	1
33	feature 523	33	feature 523	1	107	feature 202	107	feature 202	1	181	feature 483	181	feature 483	1	255	feature 585	255	feature 585	1
34	feature 407	34	feature 407	1	108	feature 181	108	feature 181	1	182	feature 25	182	feature 25	1	256	feature 321	256	feature 321	1
35	feature 65	35	feature 65	1	109	feature 120	109	feature 120	1	183	feature 139	183	feature 139	1	257	feature 546	257	feature 546	1
36	feature 212	36	feature 212	1	110	feature 322	110	feature 322	1	184	feature 351	184	feature 351	1	258	feature 255	258	feature 255	1
37	feature 458	37	feature 458	1	111	feature 460	111	feature 460	1	185	feature 281	185	feature 281	1	259	feature 348	259	feature 348	1
38	feature 115	38	feature 115	1	112	feature 46	112	feature 46	1	186	feature 47	186	feature 47	1	260	feature 78	260	feature 78	1
39	feature 422	39	feature 422	1	113	feature 438	113	feature 438	1	187	feature 431	187	feature 431	1	261	feature 478	261	feature 478	1
40	feature 428	40	feature 428	1	114	feature 20	114	feature 20	1	188	feature 332	188	feature 332	1	262	feature 62	262	feature 62	1
41	feature 486	41	feature 486	1	115	feature 540	115	feature 540	1	189	feature 169	189	feature 169	1	263	feature 469	263	feature 469	1
42	feature 105	42	feature 105	1	116	feature 269	116	feature 269	1	190	feature 427	190	feature 427	1	264	feature 377	264	feature 377	1
43	feature 588	43	feature 588	1	117	feature 39	117	feature 39	1	191	feature 118	191	feature 118	1	265	feature 434	265	feature 434	1
44	feature 320	44	feature 320	1	118	feature 406	118	feature 406	1	192	feature 58	192	feature 58	1	266	feature 52	266	feature 52	1
45	feature 24	45	feature 24	1	119	feature 1	119	feature 1	1	193	feature 292	193	feature 292	1	267	feature 51	267	feature 51	1
46	feature 361	46	feature 361	1	120	feature 146	120	feature 146	1	194	feature 160	194	feature 160	1	268	feature 19	268	feature 19	1
47	feature 239	47	feature 239	1	121	feature 455	121	feature 455	1	195	feature 325	195	feature 325	1	269	feature 84	269	feature 84	1
48	feature 101	48	feature 101	1	122	feature 369	122	feature 369	1	196	feature 577	196	feature 577	1	270	feature 216	270	feature 216	1
49	feature 317	49	feature 317	1	123	feature 175	123	feature 175	1	197	feature 168	197	feature 168	1	271	feature 268	271	feature 268	1
50	feature 130	50	feature 130	1	124	feature 394	124	feature 394	1	198	feature 453	198	feature 453	1	272	feature 209	272	feature 209	1
51	feature 311	51	feature 311	1	125	feature 189	125	feature 189	1	199	feature 413	199	feature 413	1	273	feature 421	273	feature 421	1
52	feature 306	52	feature 306	1	126	feature 63	126	feature 63	1	200	feature 10	200	feature 10	1	274	feature 362	274	feature 362	1
53	feature 23	53	feature 23	1	127	feature 291	127	feature 291	1	201	feature 198	201	feature 198	1	275	feature 367	275	feature 367	1
54	feature 218	54	feature 218	1	128	feature 511	128	feature 511	1	202	feature 96	202	feature 96	1	276	feature 240	276	feature 240	1
55	feature 544	55	feature 544	1	129	feature 572	129	feature 572	1	203	feature 584	203	feature 584	1	277	feature 343	277	feature 343	1
56	feature 366	56	feature 366	1	130	feature 173	130	feature 173	1	204	feature 256	204	feature 256	1	278	feature 196	278	feature 196	1
57	feature 345	57	feature 345	1	131	feature 368	131	feature 368	1	205	feature 88	205	feature 88	1	279	feature 67	279	feature 67	1
58	feature 72	58	feature 72	1	132	feature 288	132	feature 288	1	206	feature 69	206	feature 69	1	280	feature 210	280	feature 210	1
59	feature 156	59	feature 156	1	133	feature 333	133	feature 333	1	207	feature 435	207	feature 435	1	281	feature 476	281	feature 476	1
60	feature 279	60	feature 279	1	134	feature 33	134	feature 33	1	208	feature 561	208	feature 561	1	282	feature 358	282	feature 358	1
61	feature 492	61	feature 492	1	135	feature 93	135	feature 93	1	209	feature 573	209	feature 573	1	283	feature 207	283	feature 207	1
62	feature 186	62	feature 186	1	136	feature 155	136	feature 155	1	210	feature 92	210	feature 92	1	284	feature 75	284	feature 75	1
63	feature 57	63	feature 57	1	137	feature 222	137	feature 222	1	211	feature 337	211	feature 337	1	285	feature 586	285	feature 586	1
64	feature 457	64	feature 457	1	138	feature 140	138	feature 140	1	212	feature 15	212	feature 15	1	286	feature 11	286	feature 11	1
65	feature 350	65	feature 350	1	139	feature 45	139	feature 45	1	213	feature 83	213	feature 83	1	287	feature 480	287	feature 480	1
66	feature 217	66	feature 217	1	140	feature 188	140	feature 188	1	214	feature 59	214	feature 59	1	288	feature 132	288	feature 132	1
67	feature 357	67	feature 357	1	141	feature 468	141	feature 468	1	215	feature 95	215	feature 95	1	289	feature 355	289	feature 355	1
68	feature 307	68	feature 307	1	142	feature 424	142	feature 424	1	216	feature 137	216	feature 137	1	290	feature 164	290	feature 164	1
69	feature 208	69	feature 208	1	143	feature 49	143	feature 49	1	217	feature 119	217	feature 119	1	291	feature 154	291	feature 154	1
70	feature 219	70	feature 219	1	144	feature 340	144	feature 340	1	218	feature 543	218	feature 543	1	292	feature 410	292	feature 410	1
71	feature 40	71	feature 40	1	145	feature 99	145	feature 99	1	219	feature 490	219	feature 490	1	293	feature 71	293	feature 71	1
72	feature 138	72	feature 138	1	146	feature 152	146	feature 152	1	220	feature 102	220	feature 102	1	294	feature 418	294	feature 418	1
73	feature 489	73	feature 489	1	147	feature 512	147	feature 512	1	221	feature 432	221	feature 432	1	295	feature 442	295	feature 442	1
74	feature 223	74	feature 223	1	148	feature 114	148	feature 114	1	222	feature 143	222	feature 143	1	296	feature 587	296	feature 587	1

Table 22: Feature Ranking SECOM 412 & 528 Part II

Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528	Rank	SECOM_412	Rank	SECOM_528	412 = 528
297	feature 8	297	feature 8	1	372	feature 162	372	feature 162	1			447	feature 451				522	feature 98	
298	feature 440	298	feature 440	1	373	feature 485	373	feature 485	1			448	feature 450				523	feature 70	
299	feature 199	299	feature 199	1	374	feature 390	374	feature 390	1			449	feature 423				524	feature 53	
300	feature 56	300	feature 56	1	375	feature 439	375	feature 439	1			450	feature 415				525	feature 50	
301	feature 414	301	feature 414	1	376	feature 324	376	feature 324	1			451	feature 405				526	feature 43	
302	feature 30	302	feature 30	1	377	feature 408	377	feature 408	1			452	feature 404				527	feature 14	
303	feature 117	303	feature 117	1	378	feature 27	378	feature 27	1			453	feature 403				528	feature 6	
304	feature 420	304	feature 420	1	379	feature 178	379	feature 178	1			454	feature 402						
305	feature 378	305	feature 378	1	380	feature 393	380	feature 393	1			455	feature 401						
306	feature 38	306	feature 38	1	381	feature 484	381	feature 484	1			456	feature 400						
307	feature 283	307	feature 283	1	382	feature 148	382	feature 148	1			457	feature 399						
308	feature 496	308	feature 496	1	383	feature 205	383	feature 205	1			458	feature 398						
309	feature 116	309	feature 116	1	384	feature 341	384	feature 341	1			459	feature 397						
310	feature 16	310	feature 16	1	385	feature 177	385	feature 177	1			460	feature 396						
311	feature 149	311	feature 149	1	386	feature 409	386	feature 409	1			461	feature 395						
312	feature 79	312	feature 79	1	387	feature 387	387	feature 387	1			462	feature 382						
313	feature 389	313	feature 389	1	388	feature 525	388	feature 525	1			463	feature 381						
314	feature 68	314	feature 68	1	389	feature 107	389	feature 107	1			464	feature 380						
315	feature 448	315	feature 448	1	390	feature 94	390	feature 94	1			465	feature 379						
316	feature 183	316	feature 183	1	391	feature 7	391	feature 7	1			466	feature 376						
317	feature 229	317	feature 229	1	392	feature 500	392	feature 500	1			467	feature 375						
318	feature 61	318	feature 61	1	393	feature 214	393	feature 214	1			468	feature 374						
319	feature 475	319	feature 475	1	394	feature 55	394	feature 55	1			469	feature 373						
320	feature 299	320	feature 299	1	395	feature 391	395	feature 391	1			470	feature 372						
321	feature 298	321	feature 298	1	396	feature 417	396	feature 417	1			471	feature 371						
322	feature 443	322	feature 443	1	397	feature 126	397	feature 126	1			472	feature 370						
323	feature 419	323	feature 419	1	398	feature 571	398	feature 571	1			473	feature 365						
324	feature 121	324	feature 121	1	399	feature 182	399	feature 182	1			474	feature 331						
325	feature 200	325	feature 200	1	400	feature 249	400	feature 249	1			475	feature 330						
326	feature 433	326	feature 433	1	401	feature 253	401	feature 253	1			476	feature 329						
327	feature 583	327	feature 583	1	402	feature 344	402	feature 344	1			477	feature 328						
328	feature 134	328	feature 134	1	403	feature 5	403	feature 5	1			478	feature 327						
329	feature 37	329	feature 37	1	404	feature 128	404	feature 128	1			479	feature 326						
330	feature 127	330	feature 127	1	405	feature 560	405	feature 560	1			480	feature 323						
331	feature 524	331	feature 524	1	406	feature 141	406	feature 141	1			481	feature 316						
332	feature 12	332	feature 12	1	407	feature 449	407	feature 449	1			482	feature 315						
333	feature 334	333	feature 334	1	408	feature 446	408	feature 446	1			483	feature 314						
334	feature 354	334	feature 354	1	409	feature 97	409	feature 97	1			484	feature 285						
335	feature 491	335	feature 491	1	410	feature 157	410	feature 157	1			485	feature 277						
336	feature 81	336	feature 81	1	411	feature 444	411	feature 444	1			486	feature 267						
337	feature 360	337	feature 360	1	412	feature 276	412	feature 276	1			487	feature 266						
338	feature 447	338	feature 447	1			413	feature 539				488	feature 265						
339	feature 125	339	feature 125	1			414	feature 538				489	feature 264						
340	feature 145	340	feature 145	1			415	feature 537				490	feature 263						
341	feature 28	341	feature 28	1			416	feature 536				491	feature 262						
342	feature 474	342	feature 474	1			417	feature 535				492	feature 261						
343	feature 436	343	feature 436	1			418	feature 534				493	feature 260						
344	feature 527	344	feature 527	1			419	feature 533				494	feature 259						
345	feature 224	345	feature 224	1			420	feature 532				495	feature 258						
346	feature 309	346	feature 309	1			421	feature 531				496	feature 257						
347	feature 456	347	feature 456	1			422	feature 530				497	feature 244						
348	feature 124	348	feature 124	1			423	feature 529				498	feature 243						
349	feature 416	349	feature 416	1			424	feature 516				499	feature 242						
350	feature 562	350	feature 562	1			425	feature 515				500	feature 241						
351	feature 481	351	feature 481	1			426	feature 514				501	feature 238						
352	feature 590	352	feature 590	1			427	feature 513				502	feature 237						
353	feature 290	353	feature 290	1			428	feature 510				503	feature 236						
354	feature 17	354	feature 17	1			429	feature 509				504	feature 235						
355	feature 3	355	feature 3	1			430	feature 508				505	feature 234						
356	feature 308	356	feature 308	1			431	feature 507				506	feature 233						
357	feature 80	357	feature 80	1			432	feature 506				507	feature 232						
358	feature 220	358	feature 220	1			433	feature 505				508	feature 231						
359	feature 286	359	feature 286	1			434	feature 504				509	feature 230						
360	feature 32	360	feature 32	1			435	feature 503				510	feature 227						
361	feature 437	361	feature 437	1			436	feature 502				511	feature 195						
362	feature 166	362	feature 166	1			437	feature 499				512	feature 194						
363	feature 252	363	feature 252	1			438	feature 482				513	feature 193						
364	feature 310	364	feature 310	1			439	feature 467				514	feature 192						
365	feature 319	365	feature 319	1			440	feature 466				515	feature 191						
366	feature 26	366	feature 26	1			441	feature 465				516	feature 190						
367	feature 184	367	feature 184	1			442	feature 464				517	feature 187						
368	feature 284	368	feature 284	1			443	feature 463				518	feature 180						
369	feature 31	369	feature 31	1			444	feature 462				519	feature 179						
370	feature 352	370	feature 352	1			445	feature 459				520	feature 150						
371	feature 275	371	feature 275	1			446	feature 452				521	feature 142						