Forschungsbereich
Maschinenbauinformatik
und Virtuelle
Produktentwicklung
Univ.-Prof. Dr.-Ing. Detlef Gerhard

Machine Learning in Manufacturing

Master Thesis

Jakob Giner

Forschungsbereich
MaschinenbauInformatik
und Virtuelle
Produktentwicklung
Univ.-Prof. Dr.-Ing. Detlef Gerhard

Machine Learning in Manufacturing

An outline of machine learning fundamentals and concepts in manufacturing, validated by the implementation of a real-life use case in the production of a midsize company

Master Thesis

Author:
Jakob Giner

Supervisor:
Dipl. Ing. Martin Hennig

Supervisor Company:
Marius Stehling, MA

Institute of Engineering Design and Product Development

Vienna, August 2019

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtliche und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

…………………………………..
Jakob Giner

Abstract:

Machine learning has been applied successfully in recent years in applications in manufacturing such as monitoring of real-time data, detecting and interpreting patterns in data or as part of decision-making support systems. Industrial manufacturing is currently experiencing an unprecedented trend towards digitalization, commonly referred to as Industry 4.0 and machine learning, as an aspect of artificial intelligence, is believed to play a decisive role in it.

The work is divided into a theoretical and a practical part. In the theoretical part an outline of machine learning with a special focus on machine learning in manufacturing is given. Different learning types are explained and different classes of algorithms are described and summed up in an illustrative overview. Challenges and opportunities of machine learning in manufacturing are identified as well as concepts to facilitate the execution of machine learning projects. Finally, the current state-of-the-art of machine learning in manufacturing is investigated. In the practical part results and concepts found in the theoretical part are tested and proofed on their applicability in a real-life environment by implementing an unsupervised outlier detection for welding data. The machine learning project is carried out according to the cross-industrial standard process for data mining (CRISP-DM). In a first step welding data is analysed, pre-processed and fit to a standardized format. Subsequently, two different approaches for outlier detection models are implemented and compared, namely local outlier factor (LOF) and one-class SVM. To conclude, both parts of the work are summed up and main findings are pointed out. Furthermore, an outlook on possible further actions is given.

Keywords: Machine Learning, Artificial Intelligence, Industry 4.0, Smart Manufacturing, Smart Factory, Welding

Index

Delete Index in first row after refreshing Index

1 Introduction

1.1 Motivation and Problem Description/Research Issue

“Three industrial revolutions have so far led to paradigm changes in the domain of manufacturing – mechanization through water and steam power, mass production in assembly lines, and automation using information technology. However, over the past years, industries together with researchers and policy makers worldwide have increasingly advocated an upcoming fourth industrial revolution” [1]. In Europe, especially in Germany, the term Industry 4.0 was introduced in 2011 at the Hannover Messe. In other countries for instance the USA “smart manufacturing” or in South Korea “smart factory” is a different terminology for the same phenomenon [2]. The fourth industrial revolution is driven by the concept of the internet of things (IoT), the development of cyber-physical-systems (CPS) in manufacturing and the increasing availability of large amounts of data, collected in the manufacturing process [1]. One of the key aspects of this new technologies (next to many other such as connection within the system, data exchange, or cybersecurity) is data processing. Machine learning is a very promising method to process big amounts of data and can be used to implement a data-based manufacturing strategy and to benefit from information collected with the data.

As implementation of the industry 4.0 concept in manufacturing is still in its early stage and discovery phase, this work should give an experience value to the company in order to give an orientation for future projects when implementing machine learning in manufacturing.

In the course of this work an overview of the fundamentals of machine learning with a special focus on machine learning in an industrial environment shall be provided. Also, a solid overview of the latest developments and progress made in the field of machine learning in manufacturing shall be given. The work will be split into two parts. One part providing the theoretical background and a practical part where findings of the theoretical part can be tested and applied.

The research issue for the theoretical part has been defined as follows:

- Which challenges arise from the deployment of machine learning in an industrial environment?
- What are the opportunities that result from the deployment of machine learning in manufacturing?
- Which concepts and methods show particular suitability for application of machine learning in manufacturing?

For the practical part the research issue has been defined as follows:

- Validation of the theoretical concepts and specifications found in the theoretical part.

1.2 Objective and Approach

The objective is, as mentioned, to provide to the company an experience value for future projects in the field of machine learning. On the one hand, by providing a solid overview that describes fundamentals such as basic terminology and methods that are used in machine learning. On the other hand, by providing useful instruments and concepts that can facilitate the implementation of upcoming projects. The results of the theoretical part shall be used and tested in a practical part following the theoretical part.

Objectives for the theoretical part have been defined as follows:

- Identifying the current state-of-the-art of machine learning in manufacturing.
- Providing an application-oriented overview of the available algorithms and methods.
- Providing a roadmap or overview on how to proceed when implementing a machine learning project in manufacturing (problem description, data acquisition, data analysis and pre-processing, etc.)

Objectives for the practical part have been defined as follows:

- Selection of a real-life use case in the production of a midsize company
- Prototypical implementation with Python and the selection of Python based framework (e.g. Scikit-learn, etc.)
- Analysis and evaluation of the results.

In chapter 2 the basic principles of machine learning will be explained. A short introduction to the topic, definition and the history will be given. The terminology will be introduced, different kinds of learning methods will be illustrated, and the process of a machine learning project will be explained.

In chapter 3 the special application of machine learning in manufacturing will be investigated. Challenges and opportunities will be shown. Different algorithms will be grouped into methods and these methods will be presented. Data quality will be defined and a model on how to assess data will be given. Finally, a state-of-the-art research of machine learning in manufacturing will be presented.

In chapter 4 the application of machine learning in a practical use case will be carried out to verify the findings of the previous chapters and to gain practical experience on the field of machine learning. The practical use case will be executed according to the process organisation found in chapter 2 and every step will be explained.

In chapter 5 the findings will be summarized and the whole project will be reviewed and evaluated. Finally, an outlook on future possibilities for the continuation of the work will be given.

2 Introduction to Machine Learning

2.1 Definition

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." [3]

This definition was given by Tom M. Mitchell in 1997. It is very broad in order to include any computer program that improves its performance at some task through experience. While regular programs strictly follow static program instructions, machine learning algorithms operate by building a model from a given training set in order to make predictions or decisions and have the ability to learn without explicitly being programmed [4].

Machine learning is seen as a subfield of artificial intelligence and at the intersection of statistics, computer science and data science but most of the time when talking about machine learning people refer to the extraction of knowledge from data. The application of machine learning methods to large databases is called data mining [5]. Hence, to stick to this metaphor, if the knowledge is the gold and the data is the mountain, machine learning would be the tool that keeps our mine exploitable and makes the data usable.

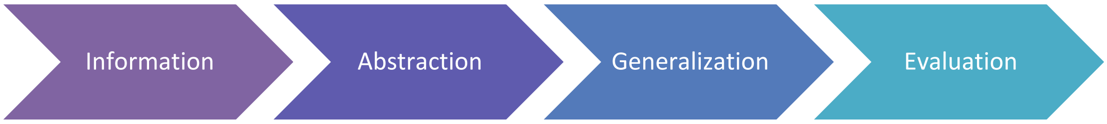
As of today, Machine Learning is already widely used in a variety of applications in everyday life. From recommendation which playlist to listen or which movie to watch, to more commercial applications like which product to order in the online shop or which advertising content may fit most to the target group. Especially when using more elaborate websites from some of the big players in online advertising market, it is very likely, that every part of the site contains multiple machine learning models [6].

Outside those everyday life applications, machine learning is also used in a variety of other fields, such as medicine, science, biology, jurisprudence etc. where experts try to develop forecast methods based on the historical data and observations. A very common example is the weather forecast, where scientists try to forecast the time variable weather conditions based on a system of equations with a big number of input parameters. If the system of equation gets too complex and the input parameters reach a level beyond the capabilities of common solving methods, machine learning approaches can provide powerful possibilities to solve those problems. Besides machine learning methods can also help to learn from the data and often reveal so far undiscovered relations within the data set, their task is not only to produce algorithms making accurate predictions but also to provide insights of the predictive structure of the data [7].

2.2 Basic principles of learning and terminology

The basic objective of machine learning is, as mentioned above, to improve the performance based on experience. Hence, in every learning process no matter if human learning or machine learning needs to go through four different stages [8]:

- Gathering and storage of information
- Abstraction, which involves the translation of information to broader representations
- Generalization, where abstracted information is used to derive knowledge
- Evaluation of the knowledge



In the following different types of machine learning will be described. For a better understanding a clarification of the basic terminology is necessary.

Attributes are quantifiable data points, characteristics or observations that describe the regarded object. If the object is the human body this would be attributes like height or weight.

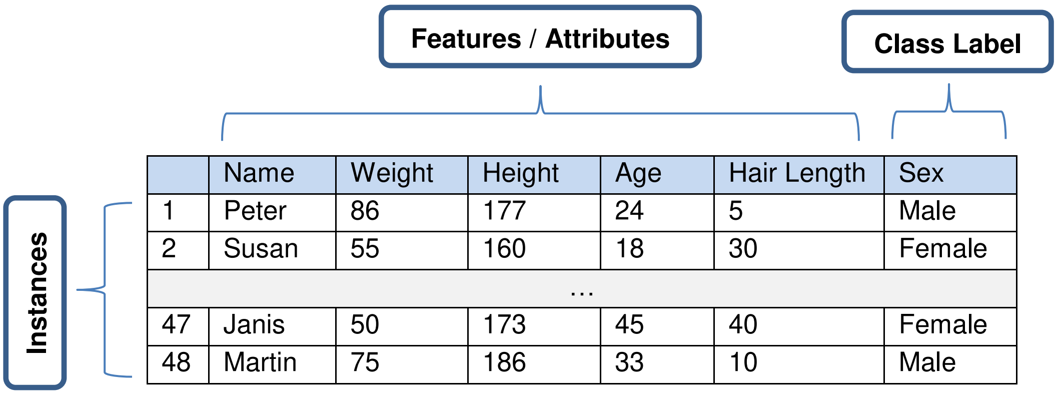
Features are representations of data to describe an object. Hence, features can be attributes but also a combination of attributes like for example body mass index. Though, in everyday language sometimes the differentiation between attributes and features is fuzzy.

Instances consist of a certain number of features and are also named feature vectors. One instance represents one record in the data set.

Label is the output class that is given to an instance.

Outliers or anomalies (also sometimes referred to as noise or exceptions) are instances or a subset of instances that deviate noticeable from other instances in the data set [9].

Figure 2-1: Basic terminology inspired by [78]



	Name	Weight	Height	Age	Hair Length	Sex
1	Peter	86	177	24	5	Male
2	Susan	55	160	18	30	Female
…						
47	Janis	50	173	45	40	Female
48	Martin	75	186	33	10	Male

The shown example in Figure 2-1 dataset consists of 5 features and 48 instances. This could be written in a 48x5 Matrix. Individual rows or columns are called vectors. For instance, each feature is a 48-dimensional column vector.

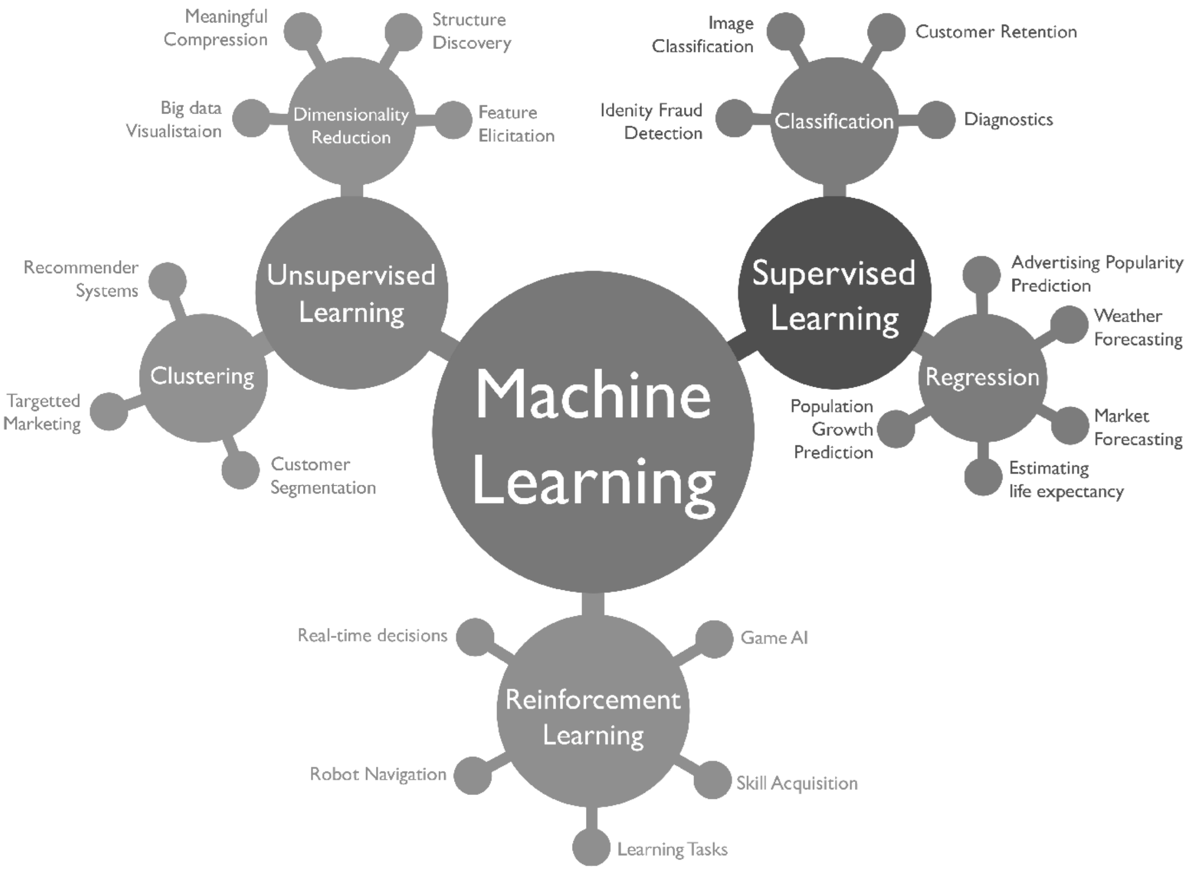
2.3 Types of learning algorithms

A huge range of different machine learning algorithms is available today. There are two approaches to categorize machine learning algorithms [10]:

- Grouping by learning type
- Grouping by similarity in form and function

A vast majority of scientific publications group the learning algorithms into three different learning types. Supervised learning, unsupervised learning and reinforcement learning.

Figure 2-2: Overview of different learning types in machine learning provided by [16]



2.3.1 Supervised Learning

The input, called training data, for supervised learning algorithms is always labelled. That means that input information comes with the right output labels. The supervised learning algorithm tries to identify the relation between features and class labels in order to model a general rule and predict the right answer, when applied to new information.

The main types of supervised learning algorithms are [8]:

Regression: These algorithms are used to predict a numeric output value based on some input features obtained from the data. The output values in this case are continuous and not discrete. An example for a regression task would be to estimate the price of a new car, based on its engine power.

Classification: These algorithms build predictive models with discrete classes as output value. Classification algorithms can predict the class labels of previously unseen data based on the similarity of its feature to a certain class. An example would be to determine if the output class is female or male based on the features weight and hair length, or an image recognition algorithm.

Figure 2-4: Examples for Regression (l) and Classification (r)

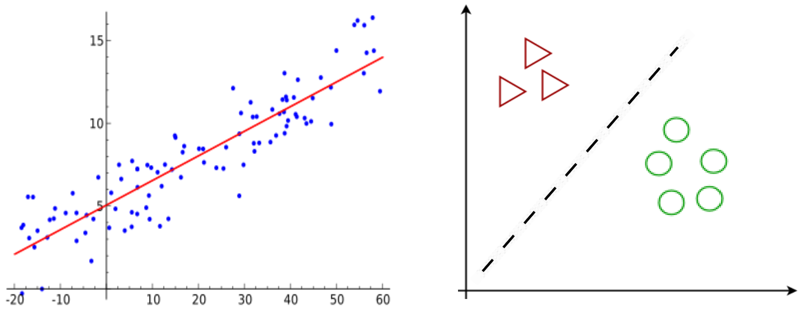
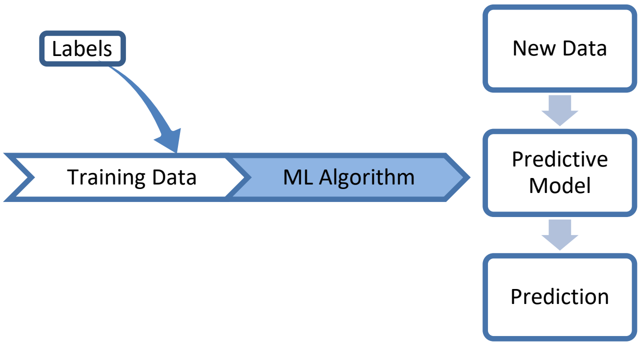


Figure 2-4: Supervised Learning Principle



2.3.2 Unsupervised Learning

The major difference to supervised learning, where the right answer is already given in the training data, is the absence of labels or given structure in the data. Hence, there is no external input that provides a correct label or class. Unsupervised learning algorithms look for commonalities, density and relations within the data. They are used to detect previously unknown correlations in the data, to detect new features or to estimate the importance of the features.

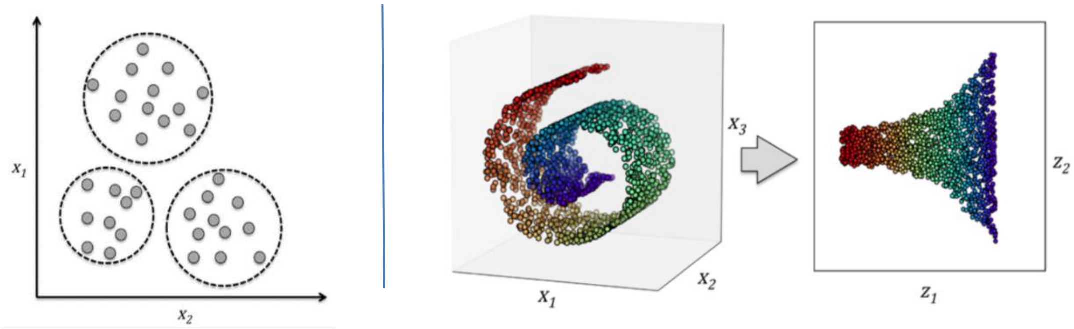
The main types of unsupervised learning algorithms are:

Clustering algorithms: The main objective of these algorithms is to cluster or group input data points into different classes or categories using just the features derived from the input data alone and no other external information. Unlike classification, the output labels are not known beforehand in clustering. [8] These algorithms are used e.g. to discover new customer groups based on their purchasing behaviour. (k-means, k-medoids, and hierarchical clustering)

Association rule learning algorithms: These algorithms are used to discover rules and patterns in data sets to explain correlations between different variables and attributes. This can give useful new insights into the data set and generate new knowledge. (Apriori and Frequent Pattern Growth)

Dimensionality reduction algorithms: Dimensionality reduction algorithms reduce the features under consideration by summarizing the essential characteristics with fewer features. It is a commonly used approach in feature pre-processing to remove noise from data, which can also degrade the predictive performance of certain algorithms and compress the data onto a smaller dimensional subspace while retaining most of the relevant information. [6]

Figure 2-5: Clustering based on two features x_1 und x_2 (left) and the dimensionality reduction from 3D to 2D (right)



2.3.3 Reinforcement Learning

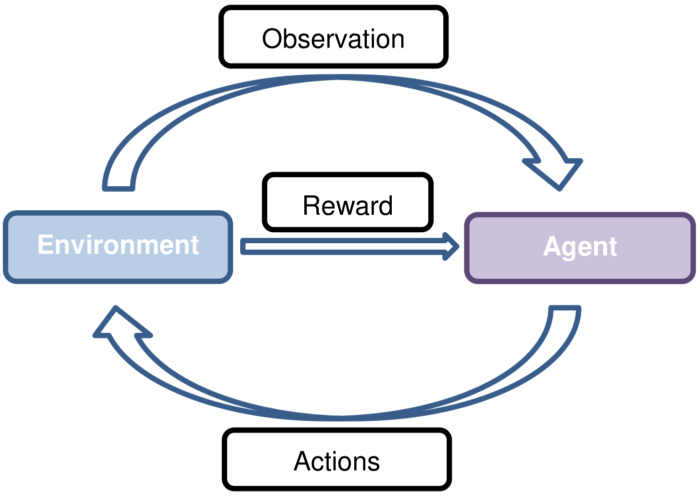
The third learning type of machine learning is reinforcement learning, which is probably the most similar one to human learning. In this type a learning algorithm, called agent, is implemented that improves its performance based on the interaction with the environment. Hence, the agent must be able to sense and be able to affect the state of its environment [11]. The feedback on how well the agent performed in the environment is provided by a numerical reinforcement signal (reward function). By trial and error, the agent uncovers which actions generate the best results. [2]

It is different from supervised learning because it works in an environment without labelled data and instant feedback. It is also different from unsupervised learning, because reinforcement learning does not look for structure or relations within the data. The goal of the reinforcement learning algorithm is to maximize its reward function.

For a better understanding, the process of reinforcement learning can be compared to the learning process of a child learning how to ride a bike. Showing pictures of trees to a child in order to teach it how to distinguish trees from other objects would correspond to supervised learning. Keeping the balance and riding a bike is something a child learns on its own by trial and error and with the feedback of the reward signal (in this case pleasure if it works and pain if falling from the bike).

Reinforcement Learning is used for skill acquisition in games such as chess, Go, or computer games, to improve robotic navigation in a dynamic environment and to improve self-driving cars.

Figure 2-6: Reinforcement Learning Principle



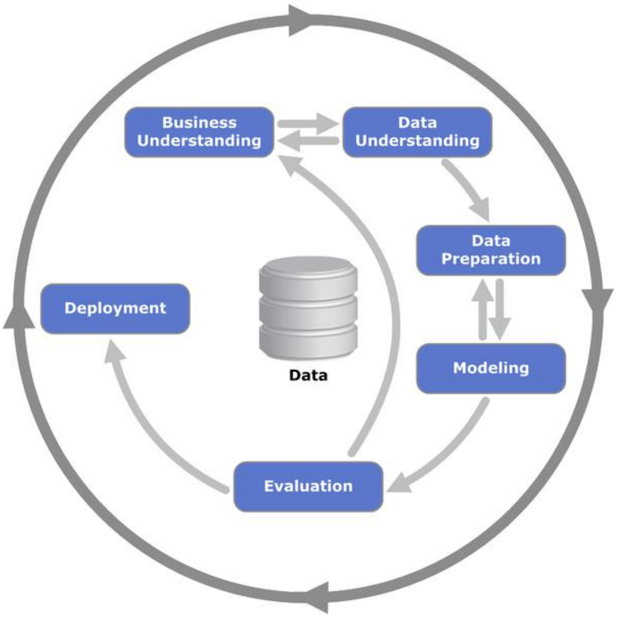
2.4 Implementation of a Machine Learning System

Different approaches and roadmaps exist for the implementation of machine learning applications. The “Cross-Industry Standard Process for Data Mining” (short: CRISP-DM) was designed in 1996 and is a very proven and reliable process model since then. Other prominent approaches are SEMMA (Sample, Explore, Modify, Model and Assess) and the KDD Process. However, in this work the CRISP-DM Model will be used as this process is seen to be the most established one with well-defined single process steps [12].

The CRISP-DM Model breaks down the machine learning project into 6 phases (see Figure 2-7) and these phases into several steps (see Figure 2-8) [13]:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

Figure 2-7: CRISP-DM Model Process Diagram [14]



The IBM CRISP-DM Manual [13] released by the IBM Corporation gives very detailed instructions on how the implementation process of a machine learning and data mining application works. A summary of the manual will be given on the next pages. A much more detailed summary of the CRISP-DM process can be found in [15].

Business Understanding:

To start with the project, it is important to get an understanding of the business perspective of the project and the business goals. This ensures that everybody involved in the project knows about the real objective of the project. The business goals can then be translated to machine learning goals. The output of this first phase is a preliminary project plan that describes the intended realisation of the machine learning objectives, as well as success criteria. Success criteria are necessary to measure, whether the business objectives as well as the machine learning objectives have been reached.

Data Understanding:

This phase starts with an initial data collection, where out of a variety of sources all the available data is gathered together. Next step is to describe the data in terms of quantity and quality. Important is the amount and the condition of the data, such as sample size, formats, dimensions, etc. but also the question if the data can meet the requirements to reach the objectives. Data Exploring is the next step, where the data is processed or visualized to get a deeper insight and to discover errors, patterns or correlations within the data. Verifying the data quality is the last step, where the data is checked for missing data, data errors, measurement errors or other inconsistencies.

There is an interaction between data understanding and business understanding because the formulation of the data mining problem and the project plan requires at least a basic idea of the available data [16]. If the data basis is found not to be sufficient to reach demands of the intended machine learning goals or business goals, the business understanding part has to be adjusted.

Data Preparation:

This phase covers all the activities that are necessary to provide the final data set which will be used for the machine learning application and is the most time consuming one. It can take 50-70% of the projects time [17].

First step is to select the data that will be used. If it is no relevant or not useable due to technical constraints it can be excluded. The reason for exclusion or inclusion should be documented. After selecting, the data must be cleaned. Thus, at this stage, the data analyst must either select clean subsets of data or incorporate more ambitious techniques such as estimating missing data through modelling analyses [15]. The clean and selected data can then be further processed also referred to as data construction. In this step new attributes or features can be derived from already existing ones. For instance, if the features torque and rotational speed are available in a data set, the feature power can be derived. It is also possible to generate completely new records. As it is not uncommon to have several data sources like tables or records, the next step, called data integration, is to join the data together in one table. If necessary, the final step is to convert the data into a new format or design that is suitable for further processing.

Modelling:

In this phase the modelling techniques that will be applied are selected, applied and then calibrated. Sometimes it is necessary to go back to the Data Preparation phase in order to reformat the data to a format that fits. This is the phase where the diligent planning and preparation in the previous steps pays off, so that if everything up until this phase was carried out carefully this should not consume more than 10% of the total project time. [17]

There is no specific model for every specific problem, hence several algorithms might be able to come to a solution. Normally one or more models are selected and compared subsequently. After selecting the model, the data set is split into a training set and a test set. The training set is used to build the model in a learning process, the test set is used to measure the quality of the model. In many cases cross-validation is applied. In this case another data set, the validation set, is used to monitor the learning process in order to avoid overfitting [6]. Overfitting happens if the model corresponds too closely to the training data and misses out to generalize the training data (see Figure 3-5).

Now the creation of the chosen models starts, where the training data is used to train the models in order to be able to predict an output. Last step is the model assessment, where the models will be evaluated according to the success criteria set in the business understanding phase. If not, another model has to be chosen or the project can be taken back one step to refine data preparation.

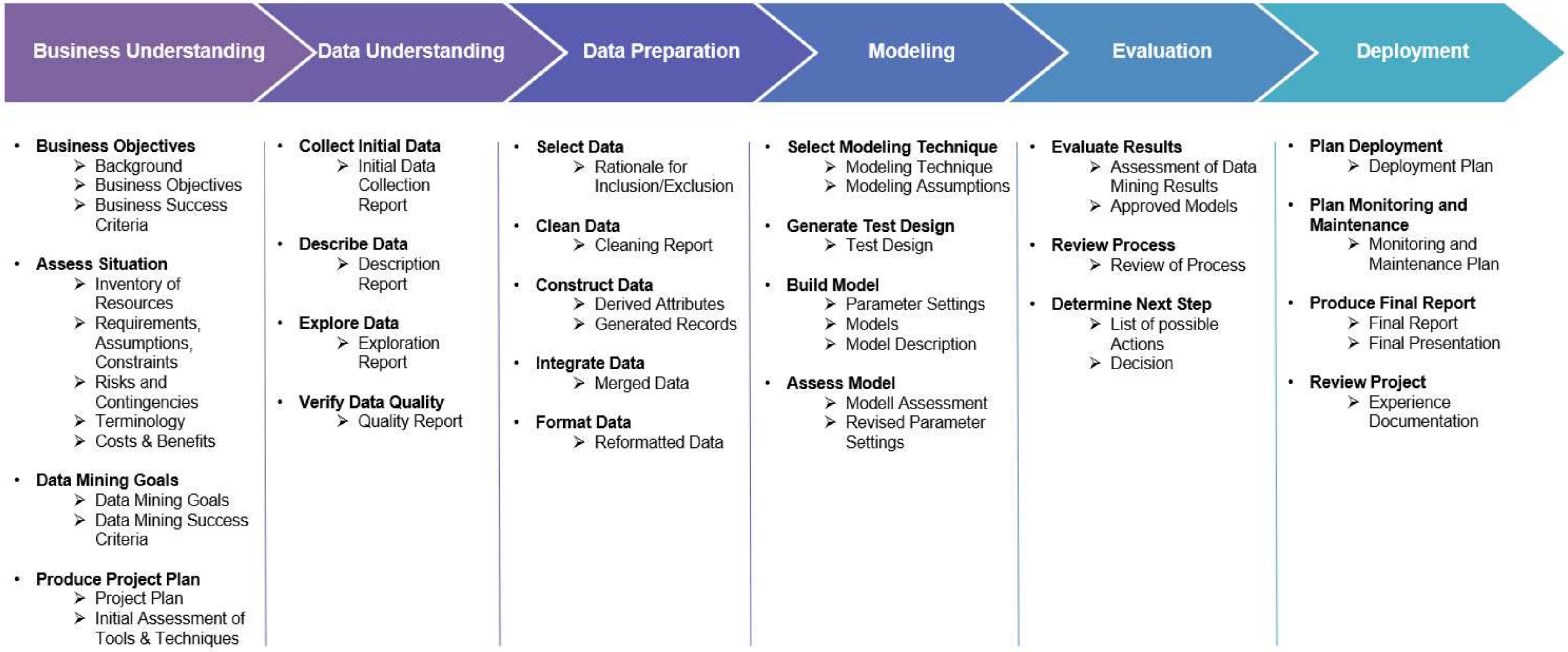
Evaluation:

At this phase the machine learning project is almost finished. The results are evaluated in terms of compliance with the initial business objectives. Both, the business success criteria as well as the machine learning success criteria should be fulfilled. Next step is to reflect the process of the project to gain valuable experience for future projects. If both steps have a positive outcome the project can pass into the last phase, otherwise the project has to go back to a previous phase.

Deployment:

The deployment phase can either be as easy as presenting the knowledge found during the project in a report, or as difficult as implementing the machine learning model into a manufacturing process of a company. However, the deployment of the results should be planned first and in case that the machine learning project is going to be continued, a strategy for monitoring and maintenance of the results must be made. Last step of the project is the preparation of a final report where the project is reviewed and reflected.

Figure 2-8: CRISP-DM Process, Phases and Steps as described in [13]



Business Understanding

- **Business Objectives**
  - Background
  - Business Objectives
  - Business Success Criteria
- **Assess Situation**
  - Inventory of Resources
  - Requirements, Assumptions, Constraints
  - Risks and Contingencies
  - Terminology
  - Costs & Benefits
- **Data Mining Goals**
  - Data Mining Goals
  - Data Mining Success Criteria
- **Produce Project Plan**
  - Project Plan
  - Initial Assessment of Tools & Techniques

Data Understanding

- **Collect Initial Data**
  - Initial Data Collection Report
- **Describe Data**
  - Description Report
- **Explore Data**
  - Exploration Report
- **Verify Data Quality**
  - Quality Report

Data Preparation

- **Select Data**
  - Rationale for Inclusion/Exclusion
- **Clean Data**
  - Cleaning Report
- **Construct Data**
  - Derived Attributes
  - Generated Records
- **Integrate Data**
  - Merged Data
- **Format Data**
  - Reformatted Data

Modeling

- **Select Modeling Technique**
  - Modeling Technique
  - Modeling Assumptions
- **Generate Test Design**
  - Test Design
- **Build Model**
  - Parameter Settings
  - Models
  - Model Description
- **Assess Model**
  - Modell Assessment
  - Revised Parameter Settings

Evaluation

- **Evaluate Results**
  - Assessment of Data Mining Results
  - Approved Models
- **Review Process**
  - Review of Process
- **Determine Next Step**
  - List of possible Actions
  - Decision

Deployment

- **Plan Deployment**
  - Deployment Plan
- **Plan Monitoring and Maintenance**
  - Monitoring and Maintenance Plan
- **Produce Final Report**
  - Final Report
  - Final Presentation
- **Review Project**
  - Experience Documentation

3 Machine Learning in Manufacturing

After the basic introduction to machine learning, this chapter is dedicated to the application of machine learning in manufacturing. After a short historical review, the possibilities and challenges will be discussed, as well as the technical requirements that go hand in hand with the implementation of machine learning. Finally, the current state of the art will be investigated and presented.

3.1 Historical Review and Status Quo of Digitalization

Machine learning as an academic discipline exists since around the 1960s, but it was in the years of 1990 that it gained momentum with the rise of the internet, the availability of huge data storage capabilities and increasing computing power. In the years of 2000 big dot-com companies, financial institutes and insurance companies realized the value lying hidden in their data database servers. They were the first to exploit the knowledge covered in implicit or explicit form. This was when the terms data mining and big data became familiar to a general public. Since around 2005 cloud storage facilities gained popularity and enabled the “birth” of the Internet of Things (IoT) somewhere between 2008 and 2009 [18]. This was the time, when manufacturing companies also started to get interested on a larger scale in machine learning, but they are still underrepresented in this field compared to other businesses [19].

Since 2010 the integration of IoT in a large number of businesses continued and as of today the status quo could be described as a transition phase from computer supported manufacturing towards a full digitalization or Industry 4.0. [18]

Digitalization in manufacturing industry is seen as a key challenge in the upcoming years [20], but manufacturing industry is also facing other big challenges that will affect the design of future manufacturing facilities and even accelerate shift towards digitalization. According to [2] and [21] the key challenges for the next decade are the following:

- Agile and flexible production capacity and supply chains
- Mass customization making manufacturing more complex and dynamic
- Sustainable manufacturing (processes) and products
- Application of information management, and AI systems
- Collaborative manufacturing and extended enterprises
- Industrial application of rapid manufacturing techniques
- Traceability and monitoring of production
- High value manufacturing in terms of society, environment and return on investment

- Tendency towards product-service systems (Servitization of manufacturing)

Machine learning can help to address these key challenges and will most likely play an increasingly important role in manufacturing in the upcoming years [2].

3.2 Challenges and Opportunities of Machine Learning applications in manufacturing

This chapter examines general challenges manufacturing companies face, when implementing machine learning followed by challenges machine learning applications have to cope with when used in a manufacturing environment. Additionally, an overview of the possibilities will be given.

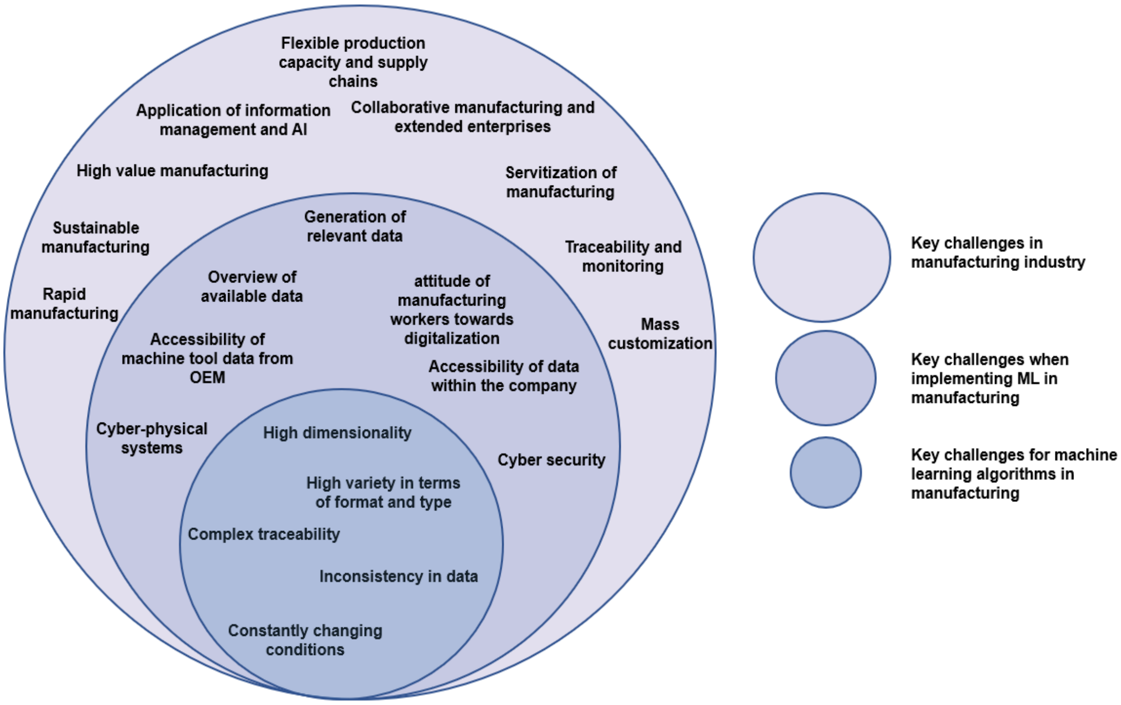
Looking at the key challenges for manufacturing industry as described in the previous chapter, a trend towards complexity and flexibility is visible. Not only in the manufacturing process but also in the product itself. To answer those challenges some data driven solution approaches have been proposed, but to implement these approaches an unprecedented amount of data is necessary.

- The first challenge manufacturing companies face, is to generate the data that is needed as well as to access the generated data in a way it is useable. Very frequently original equipment manufacturers (OEM) do not give access to their machine tool data although the data would be available. When investing in new production resources, accessibility of data should be considered. Once data is generated, another challenge is to get an overview of the data in the process in order to make it easy to access for everyone. This can be managed by creating a data map.
- Accessibility to data is one of the big paradigms of Industry 4.0, where the machine shall be embedded into a network called cyber-physical-system (CPS), which connects all physical devices to the internet [20]. This is challenging especially for established manufacturing industries that include usually machine tools from multiple suppliers with various PLCs, OS versions, sensor systems, protocols and age [22]. Accessibility or to integrate all these machines, with different sensors, data types and formats in the network is a second challenge.
- With accessibility of information another problem arises. Cybersecurity can be seen as the third challenge. Interlinked machines, sensors and PLC in a network are attractive targets for cyberattacks. This issue is already believed today to cost every year billions of dollars in revenue worldwide and it will continue growing throughout the next years [23]. However, machine learning can help detecting attacks and improving cyber defence.
- A fourth challenge is to work on the attitude of manufacturing workers towards digitalization, especially on the shop floor level, as they often oppose the idea of implementing a more cyber based manufacturing system. For many years the focus has been mainly on physical manifestation of manufacturing technology [22]. This means companies have to insure a continuing professional development towards a more cyber based education in order to enable workers to cope with changing demands [20].

These major challenges must be considered when moving towards digitalization and application of machine learning tools. In the following, some challenges for machine learning algorithms that are encountered in a manufacturing environment will be described.

Data collected in manufacturing is usually characterized by high variety in terms of format and type (number, string, status information) and high dimensionality (sample size < number of features) [2]. Besides data often comes in complex formats with a lack of traceability of the captured data. The deployment of equipment (machines, sensors, etc.) from a variety of OEM makes traceability hard throughout the product lifecycle and the individual process stages, if there is no identification of the component associated with sensor data or reliable time stamps [24]. Faulty sensors can lead to inconsistency or missing data points in samples and must be treated separately. To do so a profound knowledge of the process and data is necessary. Finally, the manufacturing environment is not stable but underlies constantly changing conditions, such as wear, material changes, product developments, batch production, discontinuous production processes, etc.

Figure 3-1: Summary of key challenges in manufacturing and machine learning


Flexible production capacity and supply chains
Application of information management and AI
Collaborative manufacturing and extended enterprises
High value manufacturing
Servitization of manufacturing
Sustainable manufacturing
Generation of relevant data
Traceability and monitoring
Rapid manufacturing
Overview of available data
attitude of manufacturing workers towards digitalization
Mass customization
Accessibility of machine tool data from OEM
Accessibility of data within the company
Cyber-physical systems
High dimensionality
Cyber security
High variety in terms of format and type
Complex traceability
Inconsistency in data
Constantly changing conditions
Key challenges in manufacturing industry
Key challenges when implementing ML in manufacturing
Key challenges for machine learning algorithms in manufacturing

It might be demanding to overcome challenges mentioned above, but machine learning also offers a very extensive amount of opportunities and algorithms, that can meet in particular the requirements of a manufacturing environment. [25] gives an overview of the opportunities.

- High dimensionality of data often comes paired with redundancy and correlations within the data that might even be trivial in a high degree. Machine learning algorithms have the ability to cope with high dimensionality, but also can be used to reduce the dimensionality of data sets and thereby also reduce the complexity of data without losing information. Hence, initially complex data can be presented in a much more tangible form. [2] If the dimensionality and complexity of data is not reduced, this can lead to a loss of performance of the machine learning model and disturb the learning process [19]. Machine learning can help to overcome the increasing complexibility of modern manufacturing industry.

- Different machine learning models are able to handle different formats and types of data. This is attractive especially in manufacturing, as data is usually not available in a homogeneous form.
- Some machine learning algorithms also have shown to be robust when it comes to changing conditions, as found in manufacturing environment. This means when confronted to new slightly changed data or data with noise, the machine learning model is still able to work at an acceptable performance.
- Machine learning can find patterns and relations in information that is not traceable or unlabelled and contribute to the discovery of so far unknown information. [25] As they can learn from new incoming data as well as from new results, they adapt to changing environment conditions.
- Finally, a great opportunity with particular interest for manufacturing industry is the increasing availability of user friendly and high performing open source software solutions to implement machine learning in their processes. Most of the tools are Python-based as this programming language. Some software solutions can be found in Table 1.

Table 1 Table of popular open-source libraries for machine learning

Tool	Description
TensorFlow	This machine learning System, developed by Google Brain Team and published in 2017, can operate in very heterogeneous environments such as servers, edge devices or browsers. It supports multiple machine learning methods and has a focus on neural networks. CPU as well as GPU units can be used for computing [26]
Scikit-learn	This well-known machine learning library has a big user community and provides most of the state-of-the-art machine learning algorithms since 2010. As the community is very active and many tutorials can be found, Scikit-learn is popular for beginners. [6]
Theano	This library was released in 2007 and therefor one of the oldest ones. Theano supports CPU and GPU. [27]
Keras	Keras is a library released in 2015 with a focus on neural networks and deep learning. It can run on the top of other libraries such as TensorFlow and Theano. [28]

3.3 Machine Learning methods

First, there is no standardized wording in machine learning. In this work, methods are referred to as similarities of machine learning algorithms in the way they operate and the function they have. The following pages will give a very brief overview and description of the most common machine learning methods.

In [29] a comprehensive overview of machine learning methods is given with deep insights into the methods. [10] provides a shorter, more compiled representation of machine learning methods. Both sources were used orientate and to identify the following methods:

- Bayesian method
- Decision Tree
- Dimensionality Reduction
- Instance based
- Clustering
- Neural Networks
- Ensemble
- Regularization
- Association rule learning
- Regression
- Support Vector Machine

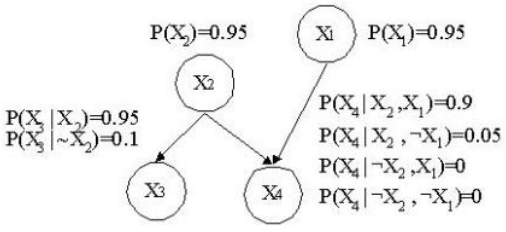
Bayesian methods:

These algorithms are based on the Bayes' Theorem. The Bayes' Theorem provides information about the conditional probability of an event. That means the probability of an event A if event B occurs. [30] The Bayes' Theorem is given as the following equation:

$$P_{(A|\,B)} = \frac{P_{(B|\,A)} * P_A}{P_B} \tag{1}$$

A Bayesian network shows probability relationship in a set of variables. The information is gathered in a network of nodes that are connected to each other in a way that there are no

Figure 3-2 Structure of a Bayes Network [31]



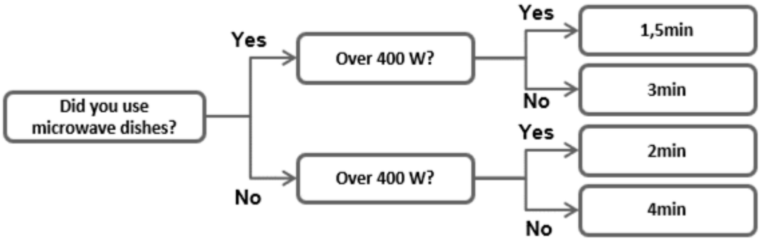
cycles [29]. In the nodes, dependencies between the nodes are represented as parameters or probabilities (see Figure 3-2). The learning process consists of 2 tasks. Learning the dependencies between the nodes and learning the parameters [31].

Bayesian methods have the ability to learn fast with respect to the number of instances. They show good tolerance to missing values and to noise. In comparison to other methods, Bayesian methods work poorly on data sets with many features and show in general less accuracy.

Decision Tree:

A decision tree is a widely used classifier and visualization of a hierarchical decision-making process (see Figure 3-3). Starting point is the root node, which is connected with at least two branches to a decision node, representing a possible decision scenario with a discrete outcome [5]. At the end of the tree is a terminal leaf, a decision node without outcoming branch. The leaf node represents a decision or outcome in form of a label or numerical value [32].

Figure 3-3: Decision Tree example, how long needs a microwave to warm up food?



The learning process starts by finding the best-dividing feature for all attributes. This best-dividing feature is then the starting point referred to as root node. This process is then reapplied to divide the part into ever smaller parts. This process has to be stopped before the algorithms fits the training data in order to avoid overfitting [31].

A big advantage of decision trees is its transparency in decision making.

Dimensional Reduction:

See Chapter 2.3.2

Instance based methods:

These learning methods uses labelled instances data as a training data set. In contrast to other learning methods, instance-based methods do not generalize the training data in order to derive a model to predict a class. [33] Instance based methods compare a new instance without label to already existing ones to classify them based on the similarity to labelled data.

Similarity is defined through the distance between several points in the defined area. [34] Unlike other methods, instance-based methods evolve and adapt with new incoming instances by integrating them into the data sets, which makes them flexible to changes of the environment.

As new instances are stored, this can lead to high memory requirements. Besides, especially if a high number of data points is compared to the new incoming instance, computational time to classify can increase rapidly [31].

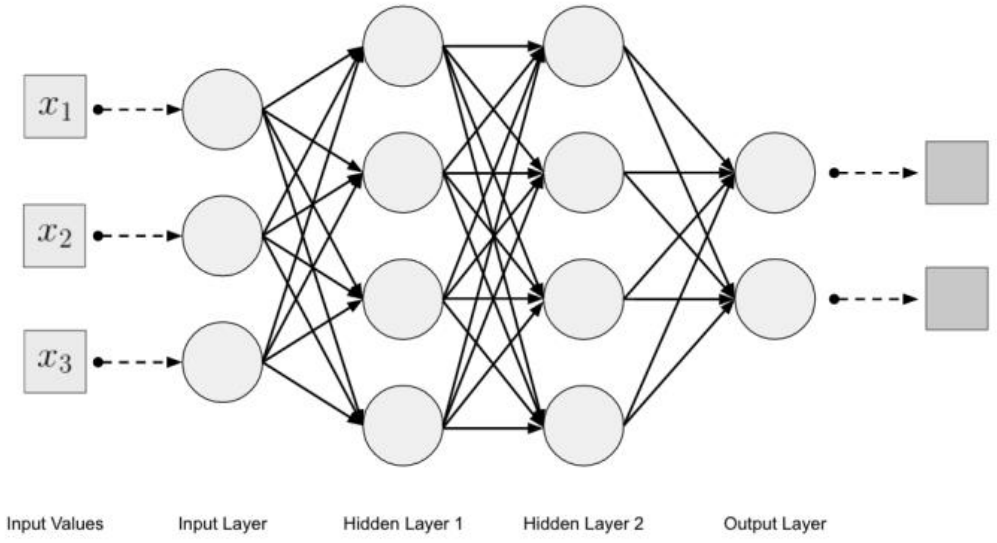
Clustering methods:

See Chapter 2.3.2

Neural Networks:

The architecture of Artificial Neural Networks (ANN) was inspired by nature more precisely by the animal and human brain. It is a network of artificial neurons that are connected together in a certain way but without a centralized control unit. [35] The most fundamental building block of a neural network is the single neuron, also called perceptron, which is a linear and binary classifier. The single neurons are connected in a hierarchical order. Each level in the hierarchical order represents one layer. The connections between the neurons have weights that can be increased or decreased in order to minimize the error on the output. [35] This is what happens in the training mode. The layers in the model represent different levels of abstraction in the data. For instance, to give an idea the first layer might recognize edges in picture, the second one recognizes ears, eyes and noses with the edges and the output layer recognizes a face. ANN's can be used for supervised, unsupervised and reinforcement learning, which makes them very versatilely applicable.

Figure 3-4: Multilayer Neural Network [35]



A special form of ANN's are deep learning networks. These networks became famous to a larger audience in the last years due to the remarkable success of deep learning in text recognition, image recognition, etc.

Deep neural networks (DNN) are artificial neural networks with a special architecture. More precisely they have more layers and the connections between the layers can be much more complex. As mentioned above, every neuron is a linear binary classifier. By introducing hidden layers and hidden neurons, nonlinearity is introduced to the model [35]. In comparison to single layer neural networks, deep neural networks have therefore the ability to build increasingly more complex and nonlinear models with every layer. This leads to a much more elaborated and automated extraction of features.

A more complex architecture of DNN also leads to higher computing effort and increased computing time.

Ensemble:

"Ensemble" is the French word for "together". In machine learning an ensemble is a method that itself is composed of an individually trained set of methods, like decision trees, logistic regression, ANN's, etc. It uses different algorithms for the same task and combines the output of each algorithm to a single output that is more accurate [36].

Finding the right combinations of algorithms for each use case needs some experience but can result in lower error, less overfitting, less bias and less variance in the results. One disadvantage all ensemble methods share is a poor transparency. As in some cases dozens of algorithms work simultaneously it can be hard to trace back the factors that lead to an improved outcome [29].

Regularization:

A special form of regression (Chapter 2.3.1) is regularization. This method is an extension to other methods that reduces overfitting by introducing a penalty term to the objective function and thus constrains the complexity and variance of the method [37]. Regularization is usually used to improve the quality of regression methods. Popular algorithms are Ridge Regression and Lasso.

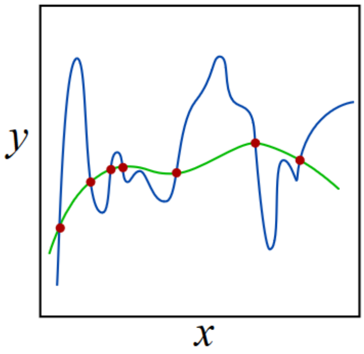
A typical relation for linear regression is given as follows:

$$y(x) = w_0 + w_1 * x_1 + w_2 * x_2 + \cdots + w_n * x_n \tag{2}$$

$y(x)$ represents the prediction the model makes, w_0 to w_n are the parameters of the model and x_1 to x_n denote the features.

Ridge regression limits the parameters at a very low level with the effect of limiting the effect of each feature as well [6]. In contrast to Ridge Regression, Lasso restricts the parameters as well but is even putting some of them to zero and hence ignoring some features completely [6]. As shown in Figure 3-5, both functions include the given data points, but the green function generalizes better and shows less overfitting.

Figure 3-5: Regression model without (blue) and after regularization (green) [38]



Regression:

See Chapter 2.3.1

Association rule learning methods:

See Chapter 2.3.2

Support Vector Machine:

The SVM is a relatively recent method and was developed back in the 90's. It can be used for supervised learning and tasks such as classification, regression or outlier detection [39]. The idea of SVM is to “best split the data into groups”, whereby best means that the hyperplane that divides the data points into groups has the biggest margin possible (Figure 3-4). SVM is a linear classifier but by applying a “kernel function” SVM can also be used as a non-linear classifier [25].

Kernel and parameter settings of the SVM are heavily influencing the performance, hence sufficient effort and time should be devoted to that. When well adjusted, SVM can well handle high dimensional data and work well on smaller data sets. In practical application, SVM have shown high performance and to be versatilely applicable [25].

It has been shown in various industrial applications such as condition monitoring, image recognition and time series forecasting, that SVM are show robust and high accurate performance in classification tasks [25].

Figure 3-6: Linear SVM

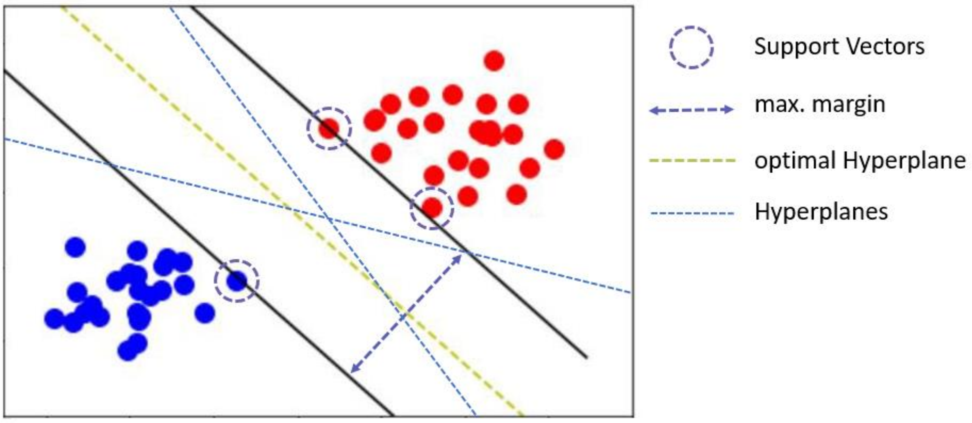
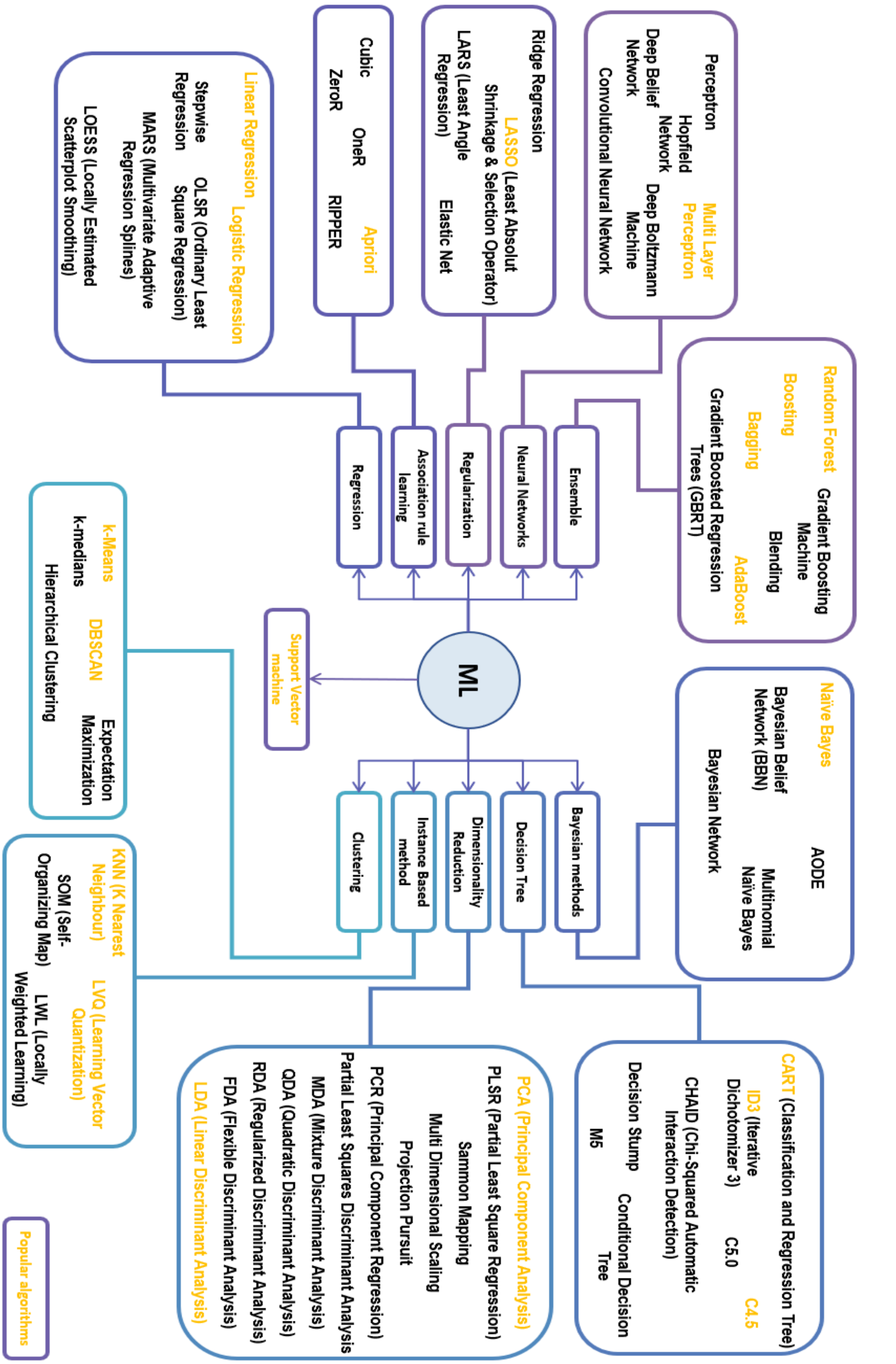


Figure 3-7 lists some popular algorithms categorized by working principle and function they have. Most of the algorithms shown below are used for supervised learning, some of them for unsupervised learning and reinforcement learning as well. However, some of them can be applied for regression as well as for classification (SVM) and some of them would fit into several categories (LVQ). Figure 3-5 should give an understanding where to put which algorithm but in practice the location of algorithms and methods is not completely strict.

Figure 3-7: Overview of machine learning methods and algorithms based on [10] and [29]


ML
Ensemble
Neural Networks
Regularization
Association rule learning
Regression
Support Vector machine
Bayesian methods
Decision Tree
Dimensionality Reduction
Instance Based method
Clustering
Random Forest
Gradient Boosting Machine
Boosting
Blending
Bagging
AdaBoost
Gradient Boosted Regression Trees (GBRT)
Perceptron
Multi Layer Perceptron
Hopfield Network
Deep Belief Network
Deep Boltzmann Machine
Convolutional Neural Network
Ridge Regression
LASSO (Least Absolut Shrinkage & Selection Operator)
LARS (Least Angle Regression)
Elastic Net
Cubic
Apriori
ZeroR
OneR
RIPPER
Linear Regression
Logistic Regression
Stepwise Regression
OLSR (Ordinary Least Square Regression)
MARS (Multivariate Adaptive Regression Splines)
LOESS (Locally Estimated Scatterplot Smoothing)
k-Means
k-medians
DBSCAN
Expectation Maximization
Hierarchical Clustering
Naïve Bayes
AODE
Bayesian Belief Network (BBN)
Multinomial Naïve Bayes
Bayesian Network
CART (Classification and Regression Tree)
ID3 (Iterative Dichotomizer 3)
C5.0
C4.5
CHAID (Chi-Squared Automatic Interaction Detection)
Decision Stump
Conditional Decision Tree
M5
PCA (Principal Component Analysis)
PLSR (Partial Least Square Regression)
Sammon Mapping
Multi Dimensional Scaling
Projection Pursuit
PCR (Principal Component Regression)
Partial Least Squares Discriminant Analysis
MDA (Mixture Discriminant Analysis)
QDA (Quadratic Discriminant Analysis)
RDA (Regularized Discriminant Analysis)
FDA (Flexible Discriminant Analysis)
LDA (Linear Discriminant Analysis)
KNN (K Nearest Neighbour)
LVQ (Learning Vector Quantization)
SOM (Self-Organizing Map)
LWL (Locally Weighted Learning)
Popular algorithms

3.4 Applicability of Machine Learning Methods in manufacturing

This chapter is dedicated to the identification of machine learning methods that are particularly suitable for application in manufacturing. Also, in this chapter a matrix is created a systematic comparison of different machine learning methods to their applicability in typical manufacturing processes.

There is no predetermined path how to select the right algorithm for a specific task. In fact, in most cases multiple algorithms can lead to a satisfactory outcome for a given problem but it is important to find an algorithm that can comply with the given challenges. However, when choosing an algorithm, several key factors must be considered. The algorithm must be chosen regarding suitability to circumstances of the working area like data situation, the respective objectives and the individual application. For instance, neural networks work very well in image recognition when a big data set is available to train the ANN but if only few data instances are available and the goal of the application is to gain new insights into the process with the algorithm, ANN's are maybe not the right choice.

3.4.1 Assessing Data Quality

As seen in the CRISP-DM model, the most important part of every machine learning project is data understanding, accordingly, sufficient effort has to be put its assessment. Several works tried to examine how good data quality can be defined and how data quality can be evaluated. As the data quality has a high impact on the choice of the machine learning algorithm, it is easier to decide once data quality is rated.

In [40] the authors developed 15 characteristics of high-quality information based on a survey from the year 1996, to define data quality:

- Accessibility (easy and direct retrievable)
- Appropriate amount of data (available information must comply with demand)
- Believability (reliable sources provide information)
- Completeness (all the needed information is available at the defined time and throughout the whole process)
- Concise representation (format that is suitable and easy to process)
- Consistent representation (standardized format and representation)
- Ease of manipulation (information can be easily edited)

- Free of error (information corresponds to the reality)
- Interpretability (information is clear and unambiguous)
- Relevancy (provided information is the required information)
- Objectivity (information is unbiased)
- Reputation (information originates from a reliable source)
- Timeliness (recent and up-to-date information)
- Understandability (information is transparent and comprehensible)
- Value-added (information can be used in a beneficial way)

This line-up of aspects of information quality should give its observer an input when assessing the quality of the available data. Not all those aspects have to be met without exception. If aspects are more essential than others, depends highly on the use case. For instance, in machine learning, understandability and relevance might not be essential, whereas concise representation and free of error can greatly facilitate the task. However, accessibility and an appropriate amount of data is crucial.

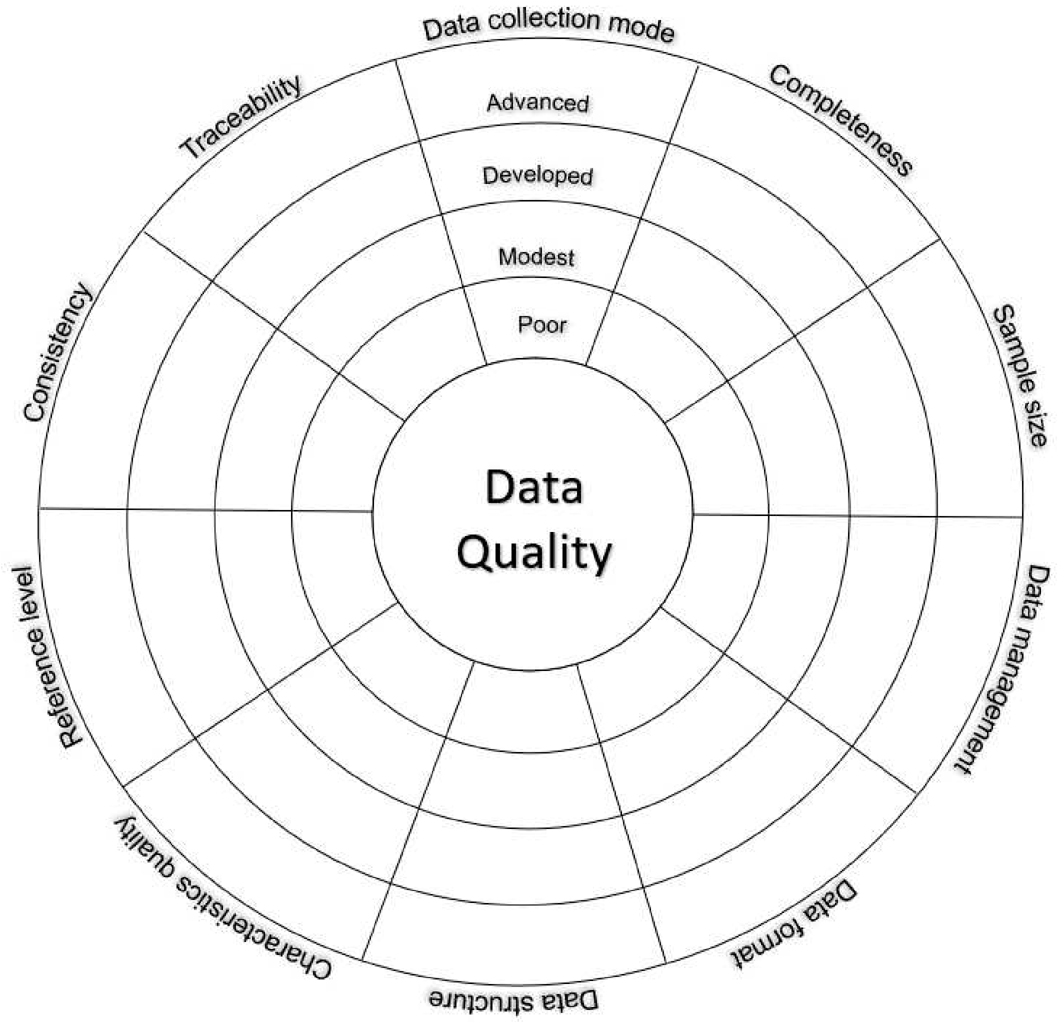
Especially in the manufacturing industry IT infrastructure and data landscape can be very heterogeneous, which can constitute a high barrier for the use of machine learning. Taking into account the special requirements in an industrial manufacturing environment, in [41] the authors present a model to assess the quality of the available data for an industrial environment. This model in form of a table (see Table 2Table 2: Model to assess data maturity in the context of machine learning) identifies 10 different criteria and divides into 4 different levels of data maturity in the context of machine learning. The model can be used in step 2 of the CRSIP-DM model “data understanding” to estimate the required time and effort that has to be put into data preparation.

Figure 3-8 was developed on the basis of the findings in [41] and can serve as a template to assess data in an intuitive and illustrative manner. In Figure 4-2 the template has been used to assess the data basis of different use cases.

Table 2: Model to assess data maturity in the context of machine learning [41]

	Data Maturity Level			
	Poor	Modest	Developed	Advanced
Data collection mode	Manually	Manually initiated	Largely automated	Online real time access
Completeness	Incomplete capture of characteristics	Capture of major characteristics	Capture of all relevant characteristics	Capture of all characteristics
Sample size	No historical data	Small sample size	Small and big sample sizes unequally distributed classes	Big sample sizes in every class
Data management	Paper records	Decentralized data storage	Centralized data storage in a data management system	Universal data warehouse
Data format	Hardly convertible formats	Convertible with moderate effort	Different but easily convertible formats	Standardized format
Data structure	Unstructured text or images	Semi-structured data	Structured but unscaled data	Structured, scaled data, standardized code
Characteristics quality	Sole set values	Highly aggregated actual values	Aggregated actual values from raw data with low scanning frequency	Raw data in real time
Reference level	Values of highest reference level	Values of high reference levels	Values of one reference level above	Values of element level
Consistency	No consistency in data	Multiple logical contradictions in data	Few logical contradictions in data	Continuous consistency
Traceability	No ID or time stamp	Multiple ID's or time stamps	Main ID or time stamps	Main ID or time stamps for one reference level

Figure 3-8: Template to assess data quality based on Table 2



3.4.2 Assessing Learning Algorithms

In manufacturing environment data usually is structured, traceable and in most cases can be labelled with the help of expert knowledge. For this reason, supervised learning seems to be the best fitting learning method for manufacturing by now [2]. Typical applications use prediction models and classifiers [2]. Unsupervised learning is frequently used to get a first impression of the data and assess is usability for application by looking at patterns, correlation in the data but also finds use in rule learning, visualization, dimensional reduction, etc. [19] Reinforcement Learning is by now barely used in manufacturing industry. However, some promising steps have been made towards the application of RL in manufacturing especially in robotics and automation.

[31] examines several supervised learning algorithms and gives an overview (see Table 1) of advantages and disadvantages of popular algorithms and methods in form of a table. This provides useful information about pros and cons of these algorithms and shows one of the most important characteristics of machine learning algorithms: Its capability of handling imperfect data.

*Table 3: Comparing learning algorithms (**** stars represent the best and * star the worst performance) [31]*

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the 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	***	*	****	***	*	***

3.5 State of the Art – An Overview of machine learning in manufacturing

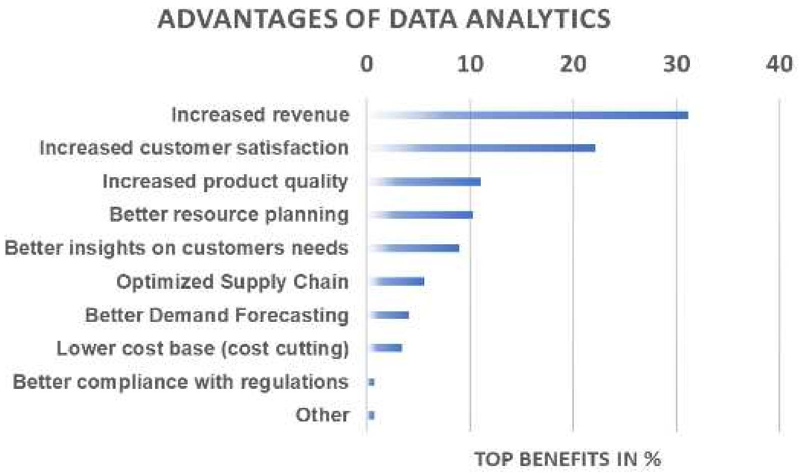
In Chapter 3.2 main challenges in the application of machine learning in manufacturing have already been discussed. In this chapter, an overview of today's application of machine learning in manufacturing will be given.

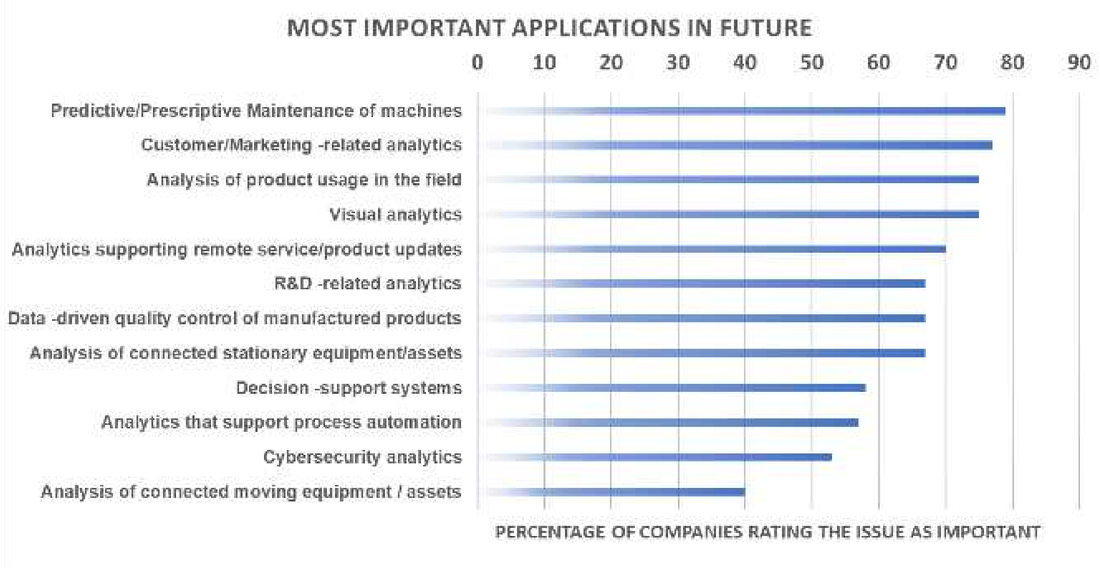
Although machine learning is a well-established field in academic research since some decades, application in industrial environment remained low. This is changing now. Modern production systems will evolve to intelligent and almost independent working process service providers [17]. Machine learning is an essential part in this development and has potential to improve the quality of a product as well as the product itself. Either with the help of expert systems for the machine operator or by providing value-adding services to the customer [17]. machine learning and its associated techniques are evaluated to have high impact on future manufacturing and even to be disruptive [42].

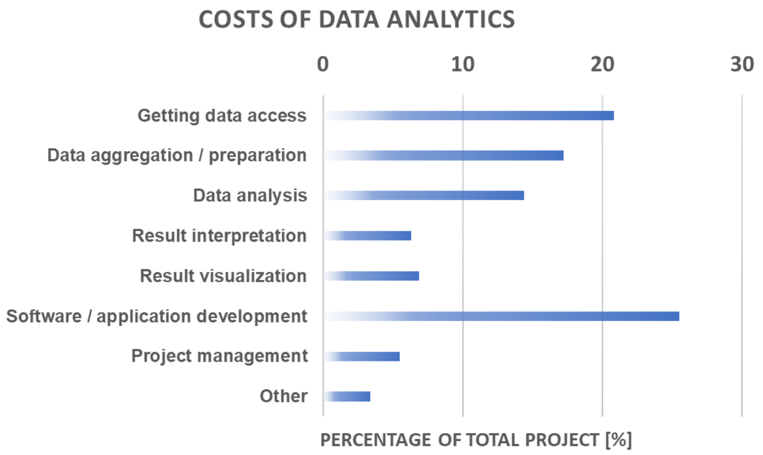
In [43], a survey on usage of data analytics in industrial environment, 69% of the participants estimate the role of data analytics to be crucial within five years. Although only 30% of the companies in the survey have completed projects, most companies have ongoing projects in this field, whereby a majority of almost two third relies on open-source tools. When the survey was carried out in 2016, only 32% of the participants believed to get good insights from the data they gather but estimated the importance of advanced analytic tools to increase sharply.

However, most companies in industrial environment seem to use data analytics and machine learning on shop floor level and the monitoring level (level 0 to 2 of ISA-95) but beneath business planning and manufacturing operations management [44].

Figure 3-9: Advantages, costs and most important data analytics applications according to [43]







The direction of development seems to go towards a real-time analysis of data. By integrating more sophisticated machine learning algorithms, the development will then go towards predictive and prescriptive real time analysis [43].

The following fields of application of machine learning methods can be found in literature [44], [17], [2], [45], [46], [47], [23], :

- Decision support when multiple interplaying variables influence a process
    - Production cost estimation [44]
    - Automated replies for requests of proposals of complex machine configuration [17]
    - Obtain similar solutions based on optimization, clustering or bagging to present it to a decision maker [44]
    - Optimal machine time scheduling [44]
    - Maintenance scheduling, Predictive maintenance [44]
    - Output estimation based on process features [45]
- Digital knowledge management of operations and plant
    - Identifying interplay between linked units [44]
    - Identifying most significant features [44]
    - Monitoring of tool and machine condition, as well as quality monitoring [2]
    - Quality Control [45]
    - Fault detection and analysis (e.g. classification) [44]
    - Implementing diagnosis and prognosis at all levels of the system [44]
    - Information consolidation and isolation [44]
    - Data driven innovation [17], [48]
    - Data integration of the whole product lifecycle [44]
    - Automated technical documentation [47]
- Intelligent autonomous diagnostic systems
    - Improve safety and flexibility [44]
    - Reduce wear and extent service life (e.g. for robots) [44]
    - Human-like machine vision (image recognition) [17]
    - Adaptive control for process optimization [17]
- Supporting Expert Systems
    - Improve and facilitate machine operation [17]
    - Interactive diagnostic maintenance tools for trouble shooting [46]
    - Cyber-attack detection systems [23]

Decision support systems based on machine learning find wide range application not just in manufacturing industry but also in medicine and finance, e.g. which medical treatment has the highest probability of success regarding the circumstances or which investment has good prospects of success considering economic data. In general, decision support systems are used in complex situations with multiple interplaying variables. In manufacturing the prediction of costs for a new product introduction with variables such as costs for material, labour, tools, development and research, etc. is such a scenario [44]. In [49] an approach with SVM to predict costs for airframe structures based on data of past projects is presented, found to provide accurate estimating performance.

In [17] the possibility of an automated processing of requests of proposals is introduced. Based on predictive models and historic data machine learning algorithms can provide an estimate of cost and can also suggest a suitable configuration of a machine.

Maintenance scheduling can save money by reducing repairing costs and provide a valuable service to the customer. One application is to schedule maintenance of a machine most efficiently just before failure and exploit almost full lifetime of the components. Another applications is the maximization of availability. Usually predictive maintenance is used to do so. In [50] a multiple classifier approach for predictive maintenance is described. Here the idea is not to give a remaining lifetime but to classify the machine into a certain “heath status” based on machine data. Depending on the health status, decision can be made like maintenance scheduling or how the machine is operated to extend lifetime.

Digital knowledge management is one key competence of modern industries. As the volume of data produced and stored is rapidly increasing, machine learning algorithms can be used on the one hand to compress data and on the other hand to extract most relevant data. Dimensional reduction algorithms (see Figure 3-5) can be used for that. In [51] the authors describe a system of automated data mining for quality control. Data is monitored by time, processed to extract non-trivial characteristics (such as variance for example) and anomalies are detected. With this method faults can be detected rapidly and relationship between product faults and machine data can be detected.

Another example is described in [52] where a feedforward neural network (FNN) is used to monitor geometrical parameters of the molten mass in a laser welding process. An SVM is then used to classify the welding seam for a defect diagnosis.

Data driven innovation is also part of digital knowledge management and can be a described as a side product of data analytics. Rule finding algorithms show relationship between features in the data set and give deeper insights into the data. The acquired knowledge can be used to find possibilities to improve the overall process efficiency with new product innovation [17]

Intelligent autonomous diagnostic systems have seen a boom in recent years with the rise of cloud computing and increasing computing power. Machine vision for drones and self-driving cars are prominent examples for this category. In manufacturing machine vision can be used for object identification, position detection, completeness check, shape and dimensional inspection, surface inspection, etc. [53]

One application of high potential is in robot supported production lines. Today’s industrial robots must be programmed for every work step they take. If something in the production changes, e.g. the geometry of a part or the production line design, everything has to be edited. This is time consuming, error-prone and does barely meet the requirement of industry 4.0. The company Autodesk is working on a project named “Brickbot”, where they use convolutional neural networks connected in series to enable industrial robots to work more autonomously. The robot is then able to detect objects in his surrounding, plan and decide what to do with the object and finally to manipulate these objects. [54]

Adaptive control is used to improve and optimize efficiency in processes. In [17] the process control of an offset printing press is described where machine learning is used to accelerate the quality control process. After refilling colours in the press, the colour must be adapted manually. By introducing an adaptive control this process can be done in shorter time and rejection can be reduced. Siemens uses a similar approach to optimize control parameters and to reduce emissions in a gas turbine. Two feed forward neural networks are put in series, whereby the second network models a quality function [55]. With this configuration it was possible to reduce NOx emissions around 15%-20%.

Supporting expert systems support the machine operator in his decision making. Rule learning algorithms can provide domain knowledge to the operator and reduce learning time. Another application is an automated spare part detection provided by Siemens [56]. Here a mobile tool, for example a mobile phone, can be used to identify spare parts easily and order them directly to support field service. Expert systems can also be used to support the IT in detecting cyber-attacks by monitoring the data flow. In paper [23] different approaches for machine learning in cyber security can be found. The use of machine learning accelerates the detection of potential cyber threats and improves cyber security.

Deep Learning in manufacturing:

Modern deep NN can use up to 150 layers in the model. These hidden layers enable the model to abstract features on its own by connecting the layers in series (also called cascades of multiple layers) [57] and reaching an ever increasing abstraction level per hidden layer. A remarkable achievement because no human handcrafted feature abstraction is needed. [57]

Deep learning networks evolved what is called the third boom period of ANN after 2000s. Since then many new deep learning models were developed e.g. deep believe networks, deep Boltzmann machine, deep convolutional networks, etc. [57] A description of different models of DNN can be found in [57]. DNN’s have been applied with great success for image recognition, video description, speech recognition, translation, social signal classification, game AI and many more [35]. In Industry DNN has been used for clustering, decision making, predictive maintenance, diagnostics, descriptive analytics, etc.

In quality control DNN have shown good performance. Machine vision can be used to detect faults such as surface imperfections after milling process or in steel processing [58] but can also be used to detect faults in welding seams [52] or incoming goods. After fault detection, a classification of faults can also be carried out and building on this decision finding machine learning methods can give advice of how to handle the fault. The paper [58] gives very good insight of how they used deep learning for surface inspection.

DNN can handle large amounts of data also in real time and show high performance in their applications. As they can extract features on its own, less time has to be spent on feature processing by human workforce. DNN show high potential for applications in manufacturing industry especially in an environment where data availability is going to increase. Hence, there are some obstacles that have to be taken in account when considering their use. Vast datasets are needed to train a DNN to a point it can work robustly. It might be time consuming and expensive to access this data. Another problem is called the “black-box problem”. As ANN’s are complex, consist of multiple hidden layers and their features are extracted in a hierarchical order of abstraction, it is not always possible to understand and therefore validate theoretically how the model came to its decision. Currently a big research area is the retraceability of ANN. [59] The black-box problem lessens the learning potential companies could take out of the machine learning methods and enhances distrust in the model.

4 Machine Learning – Application of machine learning in a practical case study

4.1 Introduction

The practical use case shall give valuable insights and experience values in the implementation of machine learning applications in the production. The objective of the project is to learn from obstacles and challenges that were encountered and to gain practical experience in the field of machine learning. Obstacles in project shall be described as well as the measures that led to a solution to provide experience for following projects. A solid documentation of the project can give useful experience values to subsequent projects in this field. The CRISP-DM process, as described in 2.4, will be applied to carry out the project. [13] and [60], provide both valuable information about the CRISP-DM process and were used as guidelines throughout the process.

The use case was carried out for the production facility of a midsize company in Upper Austria.

4.2 Process description

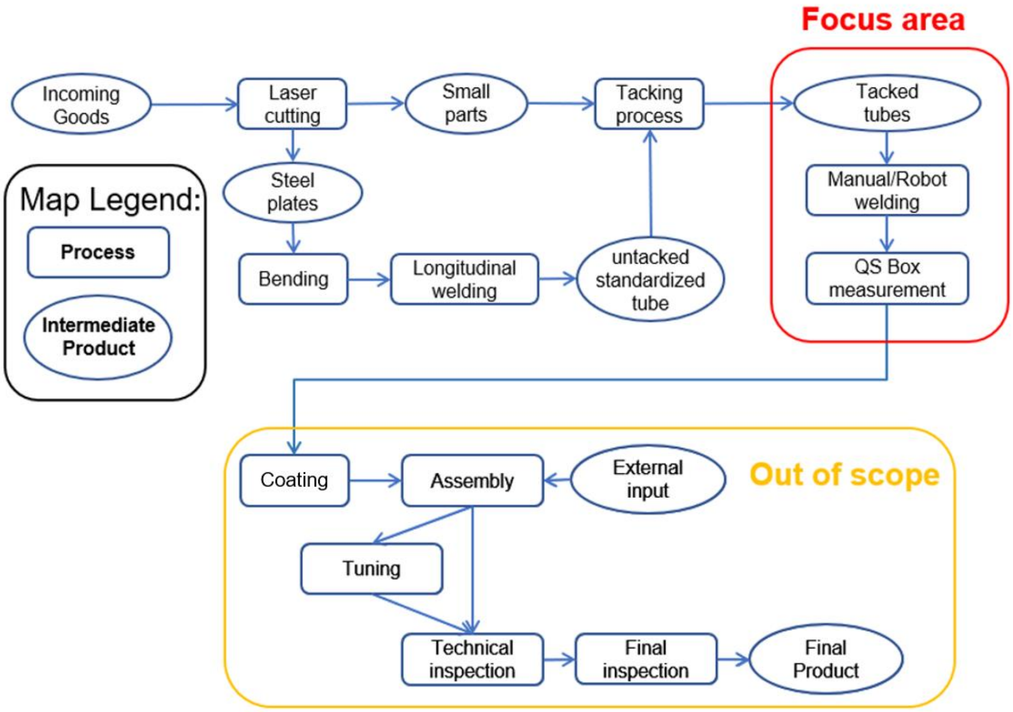
As shown in Figure 4-1, the production process starts when raw material in form of steel plates passes through the incoming goods inspection. After approval the steel plates are split into two parallel processes:

- Small-parts manufacturing
  In a laser cutting process a laser cutting machine cuts small parts out of the steel plates which are then further processed manually by milling, threading and bending.
- Production of standardized tubes
  Another laser cutting machine cuts the steel plates to the required form. Fully automated the plates come to the bending press where the plates are bended either to hexagonal tubes or P-shape tubes. Laser measurement technology is used to guarantee profile accuracy. After this step a longitudinal weld finishes the standardized tubes.

In the tacking process the previously manufactured small parts and standardized tubes are joined together by manual tack welding. After the tacking process robot welding as well as manual welding, depending on occupancy and components, is used to finalize the welding process. The resulting tube is then measured in the QS Box to guarantee dimensional accuracy and uniformity over time.

After the measurement process the tubes are prepared for the coating process and assembling process to finalize the product. As can be seen in Figure 4-1 these processes are not in the scope of this work and only mentioned for the sake of completeness.

Figure 4-1: Simplified overview of the production process



4.3 Use case selection

Several use cases were considered at the beginning. To explore the possibilities of a machine learning application a kick-off meeting was held, where the project team members explored possible use cases, the data basis of the process and the objective that should be reached. Three different use cases were discussed:

- Digital Controlled Bending
- Welding Data Examination
- Laser Cutting Data Examination

4.3.1 Scientific issues of use cases

Digital Controlled Bending is part of the bending process. After laser cutting, the steel plates arrive at the bending press where the steel plates are folded either to hexagonal tubes or Pshaped tubes. The bending press bends the plates with up to 1300 tons of press force with a positioning accuracy of around 0.01 mm. Although the process should be completely automated and deliver similar results over time, depending on different batches of material, wear condition of the machine and other influencing factors, the results vary. Corrective values have to be applied by the machine operator in order to comply with the geometrical limits. After the first component ran through the process, the machine operator measures the component and based on the measured geometry, then applies a corrective value. Usually a single throughput is sufficient to reach the geometrical limits. The know-how of the machine operator is essential when setting these corrective values. The corrective measures in the bending process lead to unexpected loss of time and production as well as to increased reject rates.

The idea of the use case is to examine the influencing factors that lead to a deviation of geometrical data. Machine learning algorithms can be used to discover pattern in the data in order to learn more about the influencing factors. In a second stage, a predictive model could be built, to support the machine operator when adjusting the corrective values by predicting values to be set.

Laser Cutting Data Examination is a use case in the laser cutting process. After incoming inspection of the steel plates, the steel plates pass through two different laser cutting machines both manufactured by Trumpf. One cutting machine to cut out small parts for the small part production and one to produce standardized tubes. Starting position is the fact, that despite seemingly similar input parameters, the cutting quality varies. Some components show weld spatters after cutting whereas other components show now cutting imperfections. A data basis has to be found and examined in order to discover factors that influence the cutting quality. This data should then be used to improve the laser cutting process.

Welding Data Examination is the idea of extracting previously unknown knowledge out of welding data. In the QS Box the standardized tubes are measured to guarantee dimensional accuracy. The measured data can then be used to trace back influencing factors that lead to greater deviation of dimensions. In the production process of the standardized tubes, the tubes pass through multiple processes that have the potential to influence the dimension. If pattern can be found in the data, this information can be used to improve future welding processes.

4.3.2 Use case discussion

The use case in the bending process is a very promising project which, if successfully carried out, would provide considerable added value to the company. Unfortunately, the data basis of the project lacks relevant information to carry out a machine learning project. The corrective values applied by the machine operator as well as the results of the measuring on which the corrective values are based on are not recorded yet. Hence, as no label in the dataset is available, supervised learning cannot be applied. Though, the corrective values could be recorded easily by introducing a fitting IT solution. The results of the measuring could be recorded as well if the tools used are able to record the results. It has been tried to collect this data but so far, no stable solution was found, the problem in the last try was the instability of the IT solution that records the corrective values. Another issue arises with the age of the bending press, as very few sensor data available. Very little data is yet recorded and the press has first to be equipped with fundamental sensor systems before starting with a machine learning project.

Nevertheless, this use case has high potential to improve considerably the throughput rate of the production of the company and reduce the rejection rate. Some steps have already been taken into this direction but have so far not shown success. Therefore, this use case should be pursued and borne in mind for future machine learning projects.

Laser cutting data examination shows a similar issue. The data basis lacks important information such as power consumption, laser focus and other technical parameters although the laser cutting machine is much more recent than the bending press. Although it can be believed that more data is generated, so far, Trumpf the OEM, does not give full access to the machine data. Probably this will change in future due to a general demand for more data, which would make this use case more feasible. Also, currently no information about the cutting quality is gathered, which makes the application of supervised learning impossible. Unsupervised learning has also limited applicability as the data basis is weak but could provide valuable new insights if the data basis can be improved.

The use case welding data examination is based on a solid data basis provided on the one hand by the PDM Link on the other hand by the QS Box. The process under observation is defined as the combination of two sub-processes. Robot welding of tacked units and measurement of the welded tubes in the QS box (see focus area in Figure 4-1). PDM Link is a production data monitoring system provided by Cloos, an OEM for welding equipment that can record a wide variety of welding parameters in the welding process. The QS Box, or quality control, measures components according to a standardized protocol and provides very detailed dimensional data for the components. A solid data basis, crucial precondition for every machine learning project whether it is supervised or unsupervised, is thus satisfied. To implement supervised learning, both data sources must be connected in order to label the welding data with a quality label coming from the QS Box. So far, the data sources are not connected and part of the use case is to establish this connection to be able to use supervised learning. Hence, in case it is not possible to connect both data sources, unsupervised learning must be applied to examine the welding data. However, in this case influencing factors that affect the dimensions of the components may probably not be found but the welding data will be analysed in a more general way like searching for striking patterns, outliers and other noticeable findings.

Other use cases for machine learning have not been considered for the closer selection but could be examined in another project. This was mainly because there was too little knowledge available about data availability or accessibility. The coating process for example was not considered for a use case. At present there is no scientific issue and no knowledge about the data basis. Still, as the machines in the process are recent and generate a fair amount of data, probably a use case could be found. Another potential use case could be found after the bending process and the subsequent longitudinal welding of the tube. After welding, the tube is measured and straightened. The results of the measuring and data from the straightening process could give conclusions on the bending process and welding process.

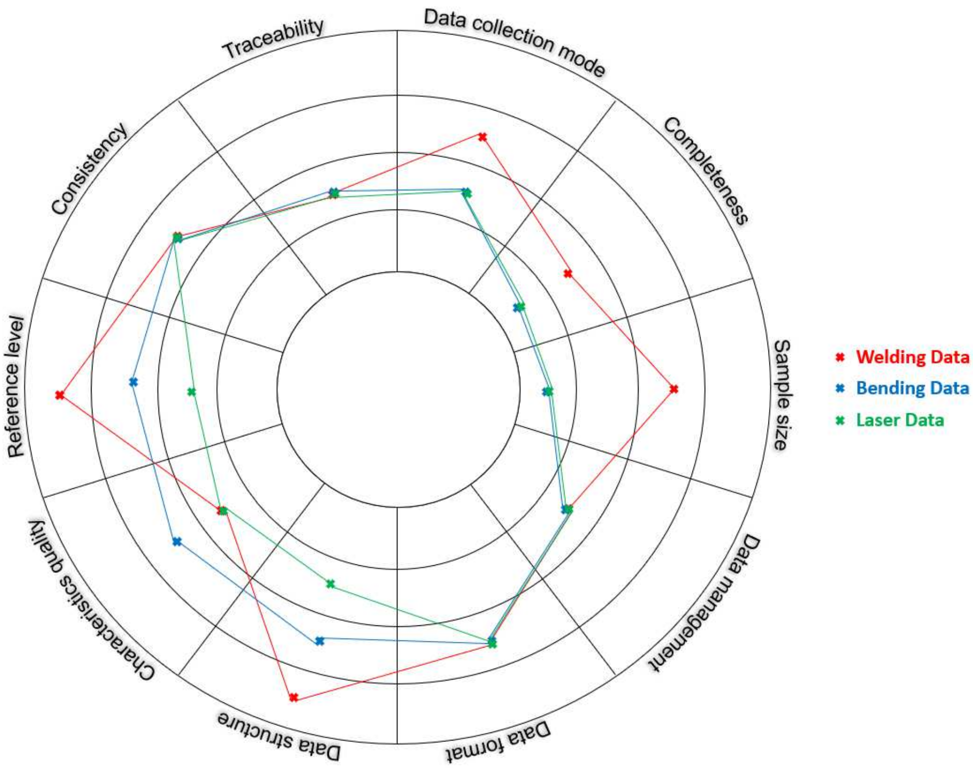
Figure 4-2 is an attempt to illustrate the data quality of all three use cases in a circular chart based on the findings in Table 2. In Table 4 an overview of the three use cases mentioned above can be found as it was found in the kick-off meeting of the project.

Table 4: Overview of possible use cases as found in the kick-off meeting (March 2019)

	Use cases		
	Digital controlled bending	Welding data examination	Laser cutting data examination
Data basis	- Correction value - Machine status (position) - Maintenance cycle - Program - Job Nr. - Material data - Job (SAP) - Drawing (PVA)	- Part / Type. Nr./Job Nr - Weld Nr. - Welding date - Voltage - Current - Current power consumption - Speed (welding speed) - Energy Consumption - Limit violation - Job Nr. (SAP) - Error code - Nr of incorrect welding seams - Parameter settings: - Method (e.g. MIGPuls) - Operating mode (e.g. S4) - Welding wire - Feed rate (Welding wire) - Welding gas - Gas flow - Material - Current set value - Voltage set value	- Power consumption - Start/Stop time - On/Off/pause status - Job Nr. - Error code - Remaining time - Current status [%]
Scientific problem	Although the process should be completely automated and deliver similar results over time, depending on different influencing factors, the results vary The objective is to identify these influencing factors in order to enable an automation of the process	PDM Cloos Link provides a variety of sensor data in high resolution. In addition, dimensional data is provided by the QS Box. The available data can be used to trace back influencing factors that lead to deviation of dimensions.	Despite seemingly similar input parameters, the cutting quality varies. The objective is to identify the influencing factors

Conclusion	Scientific problem has high potential for the production. Data basis has to be improved for execution.	It is unclear if new insights can be found. Data basis is robust.	Scientific problem exists. Data basis has to be improved for execution.

Figure 4-2: Data quality estimation according to Table 2 in form of a circular chart



4.3.3 Conclusion

Due to an insufficient data basis in the use cases of digital bending and laser cutting, a machine learning project cannot be executed before improving the data basis. At present, the data if rated as described in chapter 3.4.1, must be considered as poor, neither available in a sufficient amount nor accessible in a useable way. Though, these obstacles can be overcome with enough time and dedication. In future machine learning projects these use cases can be considered.

PDM Link is a powerful tool to monitor welding data, provided by Cloos, the OEM of welding machines in the production of the company. In the manual a broad range of features that can be recorded by the PDM Link are given. The most important ones can be seen in Table 4 and the sensor resolution is at around 100-300ms [61].

Considering the importance of a robust data basis for machine learning the welding data examination use case is the only use case that can be executed in the scheduled timeline of 6 months. The scientific issue was defined as extracting previously unknown knowledge out of welding data and to identify influencing factors of dimensional deviations. In the defined sub-process excellent welding data is available due to the PDM Link. Extensive measurements in the QS Box provide valuable dimensional data that can be used as label for supervised learning.

Weak point of this use case is at present the lacking connection between welding data and quality data and that it cannot be guaranteed that influencing factors can be found. It is unclear if influencing factors do even exist in the monitored process as dimensional deviations can also originate from preceding processes before the robot welding (e.g. the tacking process). The whole tube production would have to be integrated in a digital thread to capture all factors that influence the dimensions. In case supervised learning cannot be applied due to a lack of connection between the data sources, unsupervised learning will be applied to analyse the welding data.

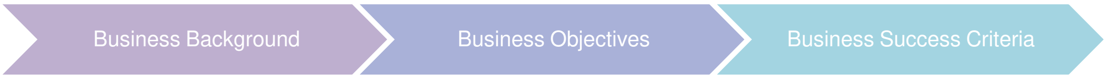
However, main objective of the project is to carry out a machine learning project in order to learn about obstacles and challenges that must be overcome, as well as to gain practical experience in this field. With these objectives in mind it was decided to focus on the welding data examination use case although that the scientific problem of the digital bending control use case would probably have more impact on the current production.

4.4 Business Understanding

As mentioned in 2.4 the objective of this section is to get an understanding of the business perspective and the business goals. The importance of this task is crucial and by taking as many people as possible on board who are involved in the project, it can be ensured that the right answer can be found on the right question [13]. Business understanding is divided in the following subtasks:

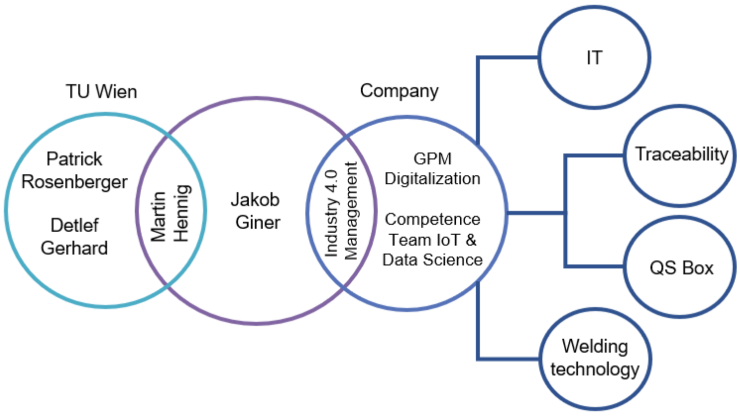
- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan

4.4.1 Determine business objectives



In the business background section, the market and personal background of the project shall be examined. The latest and future developments in machine learning as described in 3.1, 3.2 and 3.5 are the market background for the project. As the topic has prominent status and is said to be disruptive for many industries in the coming years, experience in this field is important to not lose ground. Primarily responsible for the project is Jakob Giner, student at TU Wien and author of this thesis. On the part of the company Alexander B., Christoph G. and Marius S. are the main contact persons, whereby Marius S. is the supervisor of the project. For specific questions persons from the respective departments were addressed, as can be seen in Figure 4-3. As part of the TU Wien, Martin Hennig holds the supervisor position and is the contact person.

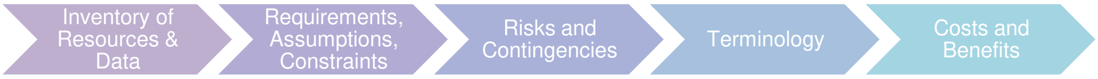
Figure 4-3: Persons involved in the project



Business objective is therefore to improve field experience in this field by implementing a machine learning project in the shop floor of the company and to generate knowledge by reviewing the project. If documented carefully and in a transparent manner, the documentation can serve as a valuable source of experience and orientation for future projects. In the kick-off event of the project is was agreed that fundamental functioning of the machine learning model and the process leading there should be focused on.

To measure the performance of the project, business success criteria must be set. As the project can be considered as a research project, business success criteria cannot be expressed in economic figures. The documentation can serve as a criterion. It should provide traceable insights and give a blueprint for future projects on how to apply machine learning in manufacturing. Also important is a critical look on problem areas and line out where future mistakes can be avoided. If the documentation meets these requirements, business success can be assumed.

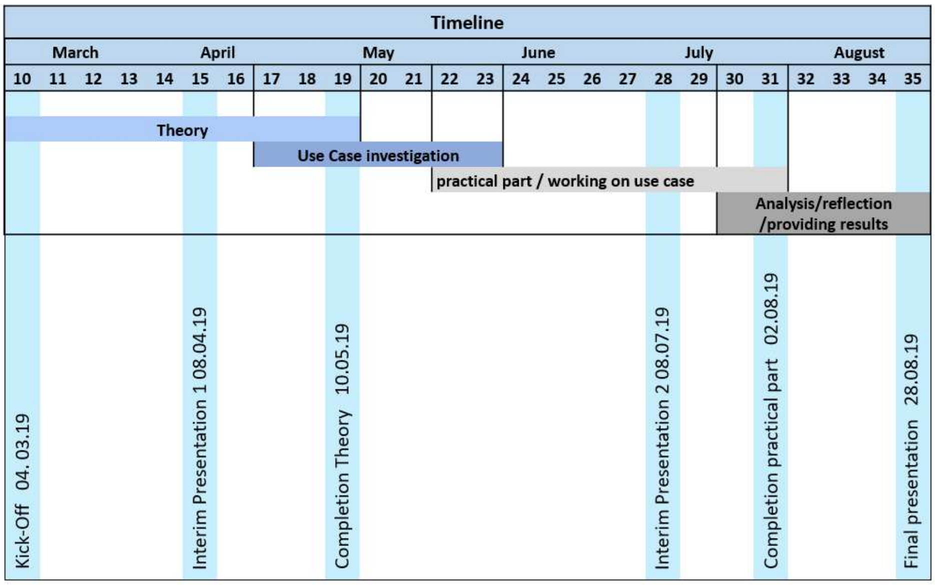
4.4.2 Assess Situation


Inventory of Resources & Data
Requirements, Assumptions, Constraints
Risks and Contingencies
Terminology
Costs and Benefits

The inventory of resources and data shall assess which resources will be needed and which data resources are available. The usage of economic resources for the project will be limited. No office space and technical equipment is needed to carry out the project. Periodic meetings will be held via Skype as well as on site in the company. Data resources of the of the welding data examination project consist of the PDM Cloos Link, a process data monitoring system that provides detailed welding data in real time, as well as dimensional data provided by the QS Box. Both data resources are available as CSV format. The data is yet not linked together, meaning that a specific component generates welding data as well as dimensional data, but the connection between this data is missing for now. As part of the project this issue should be fixed. The code for the machine learning project will be written in Python. Scikit Learn, an open source library providing a broad range of machine learning algorithms for Python [62] will be used to develop the machine learning model. Pandas, an open source library for data manipulation, will be used to pre-process the data.

Requirements, assumptions and constraints is a section to clarify expectations and identify future problems. The PDM data can be accessed only through Walter R. who has access to the data of the PDM Link. It cannot be accessed from outside the factory. The dimensional data can be accessed through Christian S. The data that is yet not linked together, but it is assumed that this connection can be provided within the expected timeline. Further it is assumed, that in the provided welding data patterns can be found. These patterns, combined with dimensional data coming from the QS Box, can then be used to discover knowledge on how patterns in welding data correspond to dimensional quality.

Figure 4-4: Timeline in calendar weeks as agreed in interim presentation 1



The risks and contingencies section evaluates possible risks throughout the project and comes up with contingency plan in case of an arising problem. Possible risks of the welding data examination project are listed in the table below.

Table 5: List of risks and contingencies

Risk	*Contingency plan*
No connection between welding and dimensional data can be made in time.	In this case, the data basis only qualifies for unsupervised learning as no label in the form of quality data is available. Patterns found with unsupervised learning can still be used to discover new knowledge about the welding process. An outlier (or anomaly) detection model will be established to use it for monitoring welding data as this can be done even with unlabelled data.
One of the main contact persons unexpectedly drops out of the team.	In this case a substitute must be found quickest possible. Probably it is possible to not lose completely the access to this person in order to be still able to execute the project.
Results do not show the expected outcome.	In this case the results have to be discussed in order to detect the problems that led to this outcome. Particular care has to be taken on the documentation of the project.
Time schedule cannot be met.	In this case a new time schedule must be agreed on.
No sensible relation between welding data and dimensional (or quality) data from the QS Box can be found.	Before coming to the Roy55 welding robot, where the standardized tube is welded, already preceding processes like tacking or bending can have considerable influence on dimensional results. These preceding processes cannot be included into the model within this project. As in case of unexpected outcome, special care has to be dedicated to the documentation in order to facilitate the reuse of the results in future projects.

In the terminology section lists technical words used in the project to avoid misunderstandings. In Chapter 2.2 the terminology for the project has already been worked out.

The costs and benefits section analyses the economic costs of a project and in the same time assesses the potential benefits that can be obtained. In the case of this project, costs can be reduced to the costs for the contract employee who works on the project. As no costs for office space or technical equipment arise, no further costs are involved. Beneficial for the company is improved field experience in the field of machine learning and the documentation of the implementation of this machine learning project that can serve as orientation for future projects.

4.4.3 Determine Data Mining Goals

Data mining goals are to firstly to access the data via PDM and through the CSV files provided by the QS Box. Secondly the data has to be linked in order to give every datapoint of the welding data a label in the form of quality data from the QS Box. Thirdly the welding data shall be searched for patterns in order to see correlations and relations within the data. Fourthly, if patterns can be found and welding data can be labelled, a relation between welding data and quality data shall be found. Fifthly, if supervised learning cannot be applied, unsupervised learning should be used to build an outlier detection model that detects reliably outliers in the welding data.

Five data mining success criteria can be defined for the five data mining goals. Firstly, the data of PDM and QS Box can be provided in format of a CSV and can be edited with Python. Secondly, it is possible to link every welding data point to a quality data point. Thirdly, a clustering algorithm can find noticeable pattern in the data. Fourthly, considerable relations between welding data and quality data can be seen. Fifthly, the outlier detection model detects unusual welding data reliably. Detected outliers will then be checked if they show unusual behaviour.

4.4.4 Produce Project Plan

The project plan provides information about stages, durations, resources, etc. of the project. In Figure 4-4 the timeline for the project can be found. As mentioned in 4.4.2 no particular resources need to be provided on the part of the company.

An initial assessment of tools and techniques has been carried out in chapter 3.

4.5 Data Understanding

A description of this task can be found in 2.4. Data Understanding is divided in the following subtasks:

- Collect initial data
- Describe data
- Explore data
- Verify data quality

4.5.1 Collect initial data

Initial Data Collection Report

Data for the project is coming mainly from two different sources. One source is the Process Data Monitoring (PDM) link provided by the welding technology company Cloos. This software was designed to record, archive and analyse operating and welding data coming from the welding process. Main objective of the software is to provide the user with background information and statistical evaluations in order to optimise the welding process. According to the manual of the PDM link [61] process data as listed in Table 6 can be accessed through the PDM Link. Not all these features can be used for the machine learning project. Mainly numeric features will be used.

No real time data is available in the PDM Link. Access to the data works through stored CSV files of the recorded process data that are created after the welding process. To start recording the process data, in the program code of the welding robot a start command had to be implemented ahead every welding seam that is programmed and an end command after the seam. The start, commands were implemented in the code of 3 different products. These products have been chosen because they can be equipped with a RFID Chip.

The PDM link also provides high resolution sensor data for each seam with a resolution of 100-300ms in form of a graph. In the graph, voltage and current curves are shown over time. Unfortunately, at present it is not possible to get access to the raw data of the graph. Cloos does not provide online access to this data. In another application, C-Gate, provided by Cloos, access to this data should be possible by a REST-API. This data could be valuable to detect welding defects. Table 6 gives an overview of data accessible through the PDM Link.

Table 6: Process data accessible through the PDM Link

Power source data		**Robot data**	
Wire diameter Gas Material Voltage: Set value, limit value, actual value, mean value Current: Set value, limit value, actual value, mean value Gas Flow: Set value, limit value, actual value, mean value Wire feed rate: Set value, limit value, actual value, mean value Powder quantity Control Voltage Throttle effect Pulse frequency Base current Pulse time Pulse voltage Pulse current Arc length	Dynamics Duty cycle water shortage Welding data - Malfunction: Current, Voltage Wire feed rate, Gas flow Wire stock, Arc breakage, Porosity Welding time Malfunction: Current, Wire, Gas Arc stationary Welding program Collective malfunction Communication malfunction Main current Temperature control	Date time Component type Serial number Program name Seam number Error text Program line of error Seam position of error Drive active Arc on Interpreter active Orientation drive active Dwell time active Waiting for Operating mode Programming mode List number Editor active Punkteditor active Motion type Seam number User name User-PAKID User-PID User description	User level User status Reason for status Speed

Second source of data is the QS Box where components of the products are measured after the manufacturing process according to a standardized norm. The results of this dimensional measurement, including angles and distances is gathered by means of the software Polyworks. Export of the data is possible and comes in the format of a XSLX file. In this file all the measurements of a particular part is stored and can be observed over time. Data of the QS Box can be accessed through the head of quality control.

Biggest issue is the connection between welding data coming from PDM link and the quality data. At present the quality data a boom produces when measured in the QS Box cannot be mapped to welding data the same boom produces when welded. To give every component an individual mark in order to clearly identify the component and assign the respective data to this component is still a big hurdle. If this hurdle can be overcome supervised learning can be used with the quality data as a label, if not, unsupervised learning can be used to examine the welding data. Quality data could also be examined with unsupervised learning. However, as welding data is easier to access through PDM Link it was decided to examine welding data. Also, the quality control department already monitors continuously the quality data with statistical methods. Nevertheless, unsupervised learning examination of quality data should be borne in mind for future projects.

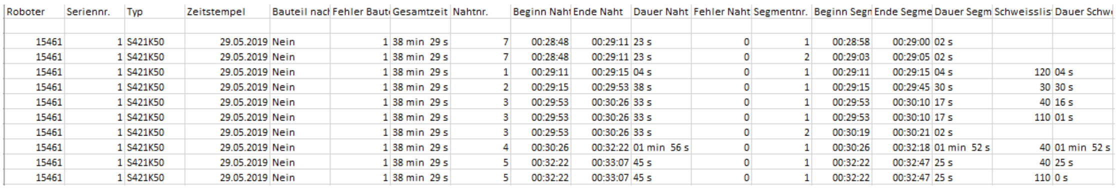
4.5.2 Describe Data

Data Description Report

As mentioned above, data coming from the PDM link comes in the format of exported CSV files of the stored welding data. Every boom represents a separate CSV file. In the file 72 features and depending on the component around 120 instances are represented. Each instance includes welding data of one segment. A segment is the movement of the robot between two coded support points and lasts depending on the segment between a few seconds up to two minutes. Several segments sum up to a seam. Each seam includes between one and eight segments.

Data includes features a variety of formats such as numerical, Boolean (true/false), categorical (string) and a mixture of both. A small extract of the first features and instances of an exported CSV file can be seen in Figure 4-5.

Figure 4-5: Small extract of an exported CSV file from the PDM link

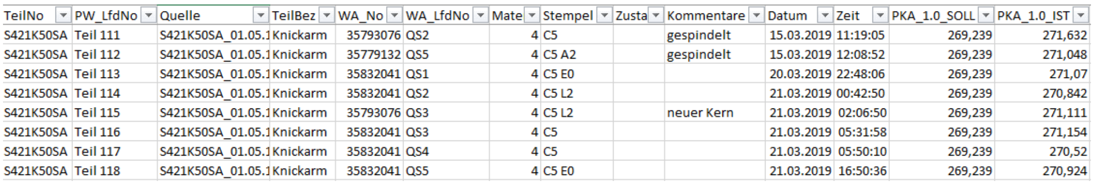


Roboter	Seriennr.	Typ	Zeitstempel	Bauteil nac	Fehler Baut	Gesamtzeit	Nahtnr.	Beginn Naht	Ende Naht	Dauer Naht	Fehler Naht	Segmentnr.	Beginn Segn	Ende Segme	Dauer Segm	Schweisslis	Dauer Schw
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	7	00:28:48	00:29:11	23 s	0	1	00:28:58	00:29:00	02 s		
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	7	00:28:48	00:29:11	23 s	0	2	00:29:03	00:29:05	02 s		
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	1	00:29:11	00:29:15	04 s	0	1	00:29:11	00:29:15	04 s	120	04 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	2	00:29:15	00:29:53	38 s	0	1	00:29:15	00:29:45	30 s	30	30 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	3	00:29:53	00:30:26	33 s	0	1	00:29:53	00:30:10	17 s	40	16 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	3	00:29:53	00:30:26	33 s	0	1	00:29:53	00:30:10	17 s	110	01 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	3	00:29:53	00:30:26	33 s	0	2	00:30:19	00:30:21	02 s		
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	4	00:30:26	00:32:22	01 min 56 s	0	1	00:30:26	00:32:18	01 min 52 s	40	01 min 52 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	5	00:32:22	00:33:07	45 s	0	1	00:32:22	00:32:47	25 s	40	25 s
15461	1	S421K50	29.05.2019	Nein	1	38 min 29 s	5	00:32:22	00:33:07	45 s	0	1	00:32:22	00:32:47	25 s	110	0 s

Every week around 5 booms are manufactured, depending on the order situation. This means per month around 20-25 datasets can be created. In the QS Box all the booms can be measured. Usually only a certain number of booms is measured but in this case every component can be examined. A good share of features does not contain data. This is either because features contain irrelevant information, redundant sensor systems produce zeros or sensor systems are not properly adjusted. However, the most important features such as date time, voltage, current, wire feeder rate, etc. are recorded reliably. In some instances, no data at all is recorded. As this is usually at the end of a seam it is evident that these blank instances are travels to the next starting point of a seam.

Data coming from the QS Box comes in form of a XSLX file. In the file a certain number of angles and distances is recorded according to a standardized process, such as width, height, length of the boom, conicity, twist, different angles and distances between load introduction zones, etc. The file consists of 101 features, thereof 44 set values and 44 actual values of the measurement. The rest are time stamps and other descriptive features of the component. In the file also an overview of the development of every measurement point over time can be found as well as an overview of how critical a measurement point is. As mentioned above, per week around 5 new data points are added to the file. In Figure 4-6 an extract of the quality data can be seen.

Figure 4-6: Small extract of an XSLX file coming from the QS Box



TeilNo	PW_LfdNo	Quelle	TeilBez	WA_No	WA_LfdNo	Mate	Stempel	Zusta	Kommentare	Datum	Zeit	PKA_1.0_SOLL	PKA_1.0_IST
S421K50SA	Teil 111	S421K50SA_01.05.1	Knickarm	35793076	QS2	4	C5		gespindelt	15.03.2019	11:19:05	269,239	271,632
S421K50SA	Teil 112	S421K50SA_01.05.1	Knickarm	35779132	QS5	4	C5 A2		gespindelt	15.03.2019	12:08:52	269,239	271,048
S421K50SA	Teil 113	S421K50SA_01.05.1	Knickarm	35832041	QS1	4	C5 E0			20.03.2019	22:48:06	269,239	271,07
S421K50SA	Teil 114	S421K50SA_01.05.1	Knickarm	35832041	QS2	4	C5 L2			21.03.2019	00:42:50	269,239	270,842
S421K50SA	Teil 115	S421K50SA_01.05.1	Knickarm	35793076	QS3	4	C5 L2		neuer Kern	21.03.2019	02:06:50	269,239	271,111
S421K50SA	Teil 116	S421K50SA_01.05.1	Knickarm	35832041	QS3	4	C5			21.03.2019	05:31:58	269,239	271,154
S421K50SA	Teil 117	S421K50SA_01.05.1	Knickarm	35832041	QS4	4	C5			21.03.2019	05:50:10	269,239	270,52
S421K50SA	Teil 118	S421K50SA_01.05.1	Knickarm	35832041	QS5	4	C5 E0			21.03.2019	16:50:36	269,239	270,924

To create a connection between the QS data and the PDM data several options were discussed. At the beginning it was thought, that this connection does already exist and after this misconception was detected it wasn't clear whether this connection can even be established or not. In a preceding meeting of the interim presentation #1 in April 2019, it was decided that several proposals should be worked out on how a connection could be established.

To interlink the different processes a way must be found to unambiguously identify the component which is currently producing data and to assign the respective data flows to this component. In industry 4.0 the idea of assigning every component the data is has produced throughout its lifecycle is called digital thread [63]. To start with, in the welding data a clear identification of the component must be enabled. The PDM link does not assign a marker to every welding data produced, it only records the component's name which is currently in process. A similar problem occurs in the QS Box. The booms are measured but the data is not assigned to the component and the data is used for statistical quality control.

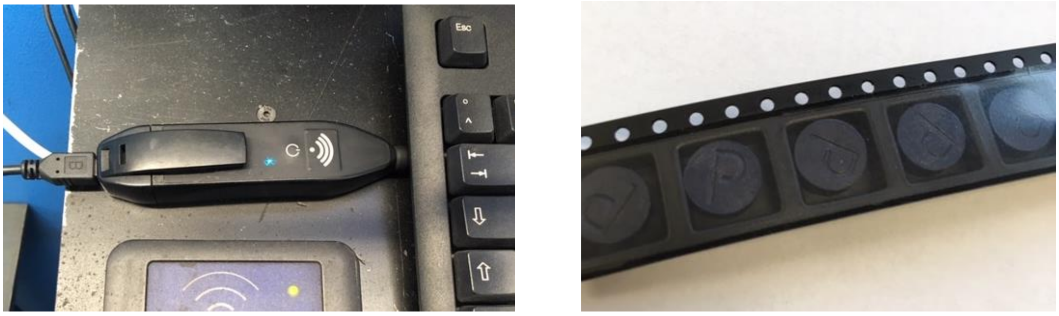
To establish the connection three options seemed to be possible:

- Manual tracking of the component during a certain test period.
- Collecting time stamps before welding and before measuring in the QS Box by scanning the barcodes that are already printed on the booms.
- Scanning the RFID tags that are already integrated in some booms before welding and before measuring in the QS Box.

Manual tracking requires high attention and attendance at the shop floor. Every component must be marked manually and a list has to be maintained were all the components are listed with the respective data. This would be possible, if enough booms would be produced to gather enough data for the machine learning project within 1-2 weeks. Besides, although this approach would deliver fast results in short term, a more long-term solution is favoured.

Collecting time stamps by scanning barcodes on the booms would be an elegant solution as the booms are already marked with a barcode. Unfortunately, this solution is expected to be prone to error as the barcode frequently gets illegible due to high heat exposure during welding or other external influences in manufacturing.

Figure 4-7: RFID tags and the RFID scanning unit



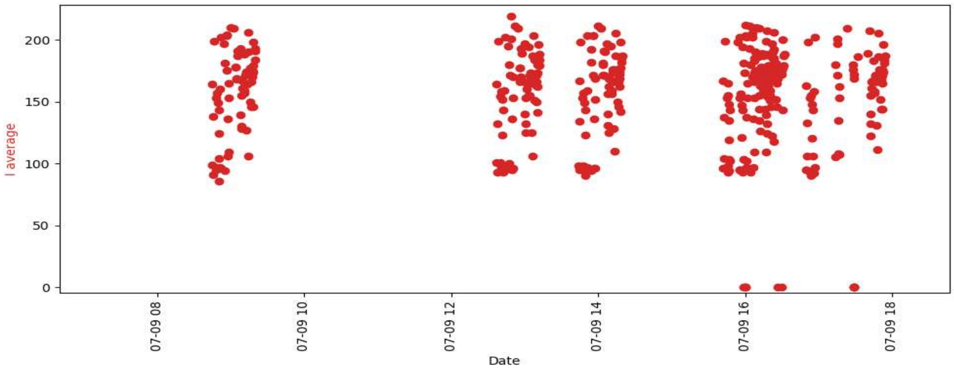
The solution of identifying the component with a RFID Tag that is built in the product is the most promising one although also the most complex one. A project of building in RFID tags in components is already ongoing at the company in the quality department. Every RFID tag is provided with a unique ID number and on each 500 characters can be saved. The idea is to have a scanning unit at each step of the production in order to assign all data flows that occur to the unique ID of the tag. This would provide considerable advantages for the internal shop floor management as well as for after sales.

4.5.3 Explore Data

Data Exploration Report

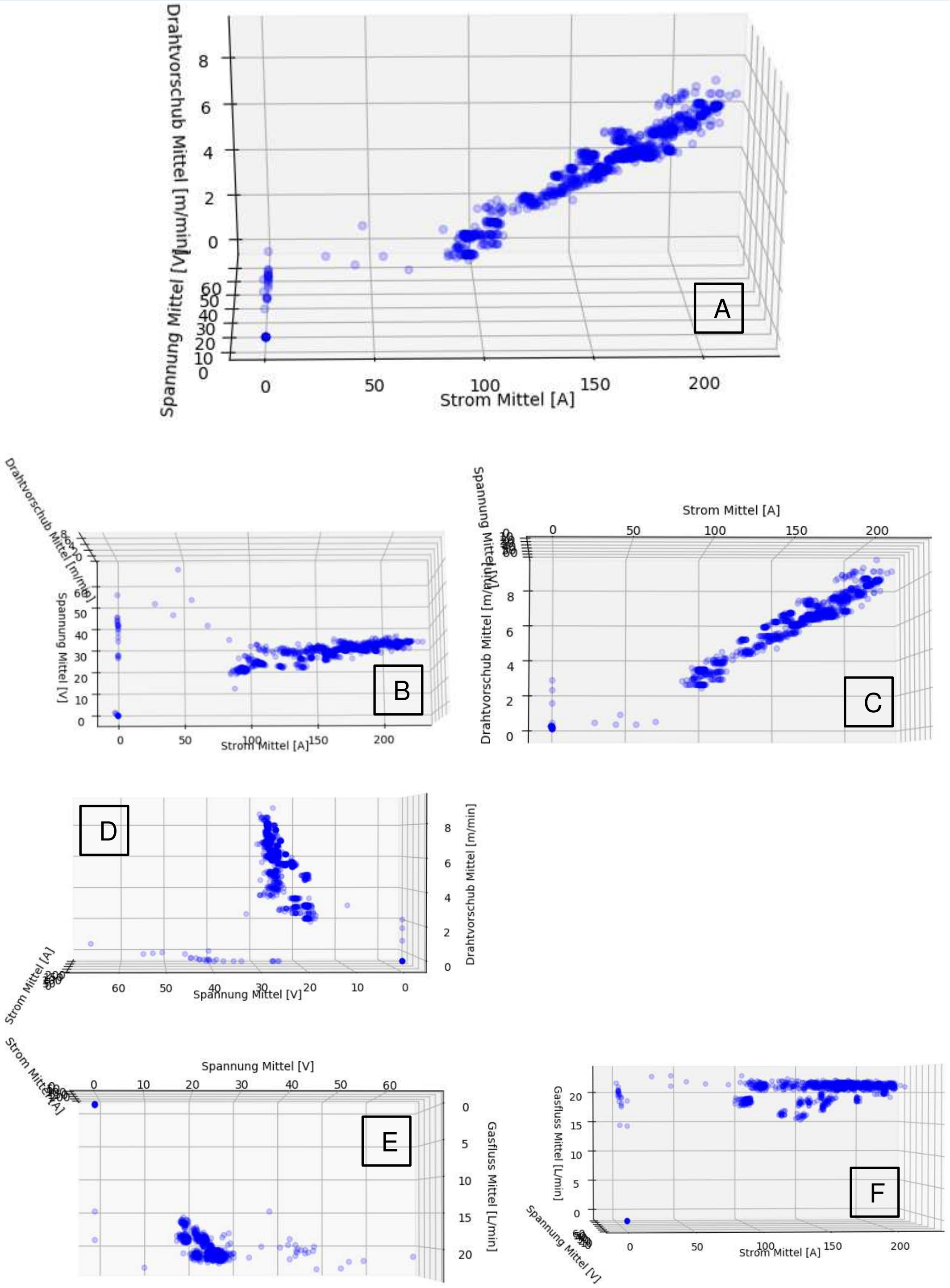
For the Data Exploration Report data from the PDM Link was used dating from 23/05/2019 until the 17/06/2019 including 1366 instances, coming from 18 welded booms. Instances with NaN entries were excluded.

Figure 4-8: Scatter plot of initial welding data (electrical current over time)



First plots of the initial data over time shows, that on the robot the same part was manufactured in series. In Figure 4-8 the plot of the average welding power over time can be seen. Some outliers, mostly zero values, were found. These values correspond to very short welding times, showing either arc breakage during welding or movements of the robot without welding.

Figure 4-9: 3D scatter plots, average values of current, voltage, wire feed speed and gas flow



In Figure 4-9 a 3D scatter plot was used to get better insights in the data. Average current seems to have the broadest spread (B) whereas voltage and gas flow is rather stable (E). Current and wire feed rate correlate positively (C). Low values of current beneath around 70A correlate with higher spread of voltage and could be outliers (B). Voltage values outside the range of 15 to 30 Volt correspond to low wire feed rates which could also indicate outliers (D).

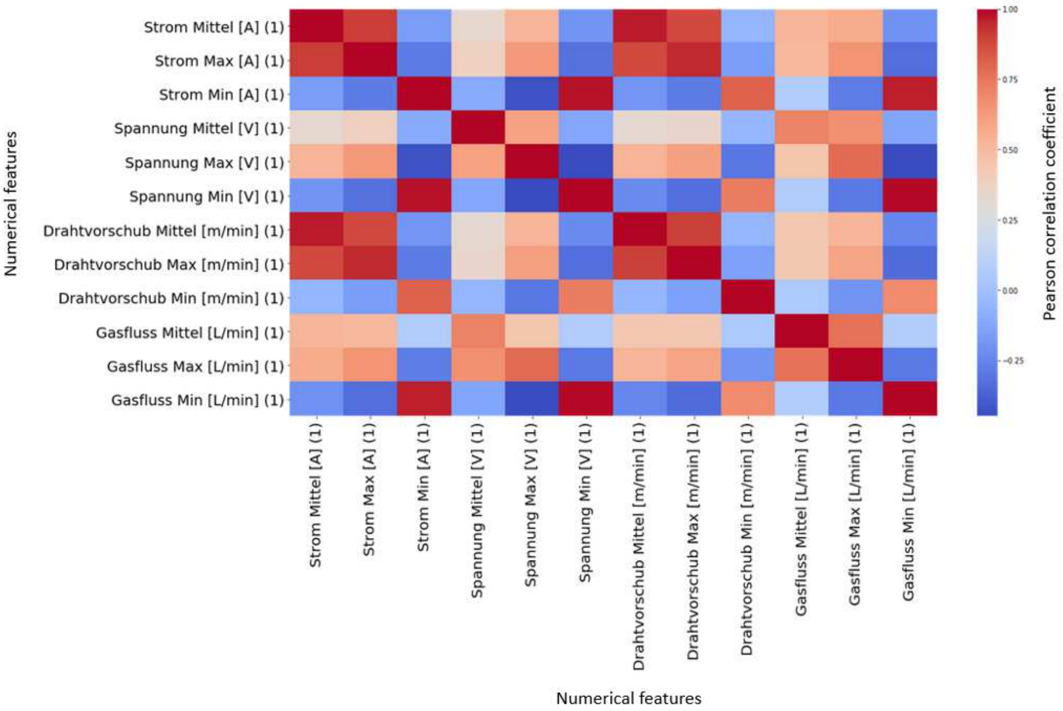
In Table 7 the variance of the numerical features of the initial data set can be seen. As expected from the plots in Figure 4-9 current has the broadest variance. Peak values of current and voltage show high variance which could lead to high noise in the data but nevertheless are good indicators for momentary deviations in the welding process. The feature “Schweissliste” with the highest variance includes the welding parameters of the respected welding seam. As this is no sensor data, this feature can be excluded from the model. All the features with set values do not change at all, except for the set value of speed. Minimum values are close to zero in average and therefore far below the average. This is an indicator, that minimum values should be treated with caution.

Table 7: Variance and average values of selected features in the initial data set

Feature	var	average	Feature	var	average
Seriennr.	0.00	1.00	Spannung Mittel [V] (1)	21.94	24.14
Fehler Bauteil	5.54	1.65	Spannung Max [V] (1)	632.52	69.38
Nahtnr.	239.14	24.29	Spannung Min [V] (1)	22.24	1.11
Fehler Naht	0.26	0.11	Drahtvorschub Soll [m/min] (1)	0.00	8.00
Segmentnr.	0.56	1.20	Drahtvorschub Mittel [m/min] (1)	3.30	5.50
Dauer Segment	972.23	30.32	Drahtvorschub Max [m/min] (1)	4.20	6.84
Schweissliste	4793.86	57.02	Drahtvorschub Min [m/min] (1)	0.61	0.13
Strom Soll [A] (1)	0.00	300.00	Gasfluss Soll [L/min] (1)	0.00	13.00
Strom Mittel [A] (1)	1864.57	152.66	Gasfluss Mittel [L/min] (1)	6.80	20.24
Strom Max [A] (1)	2774.88	190.88	Gasfluss Max [L/min] (1)	13.97	25.74
Strom Min [A] (1)	660.20	5.88	Gasfluss Min [L/min] (1)	22.16	1.12
Spannung Soll [V] (1)	0.00	37.50	Geschw. (Soll) [cm/min] (1)	244.46	42.44

To get a better understanding of the correlations in the data, a heatmap (see Figure 4-10), visualizing positive or negative correlation within the data has been made. The heatmap shows the Pearson correlation coefficient for each combination of features. This coefficient has a value between -1 and 1. 1 for total positive correlation and -1 for total negative correlation. 0 means no linear correlation between two features exists. As expected, average current and wire feed rate correlate heavily. Also, minimum values of voltage and gas flow correlate with minimum values of current, indicating as described above, that if the current falls below a certain threshold the instance is likely to be faulty. Minimum values are more likely to have negative correlations with other features then maximum or mean values. This could be owed to the fact that minimum values are mostly zero, even if average and peak values are high. Another indicator that minimum values must treated with caution.

Figure 4-10: Correlation Heatmap of selected features in the initial data set



4.5.4 Verify Data Quality

Data Quality Report

For the evaluation of the data quality, the evaluation process as shown in 3.4.1 was used. The data was rated and evaluated according to the matrix provided by [41] in Table 2. In the following a short explication on the rating will be given.

Data collection mode was rated 3, as the PDM Link collects automated specifying which features and components shall be recorded. The data is then stored in CSV files. No real-time access and no online access is possible in the current solution.

Completeness was rated 2, because some relevant characteristics such as set values are recorded in the data set. Likewise, fine tuning parameters (Pulsfrequenz, Pulszeit, etc.) are not captured.

The sample size of the data set varies. Quality data is available in higher amounts then welding data. Depending on the component, currently around 80 instances for quality data are available and 18 instances of welding data. Both numbers can be considered as small and a data base of more than 100 instances should be pursued. As the number of instances increases weekly, this aim can be reached in foreseeable future.

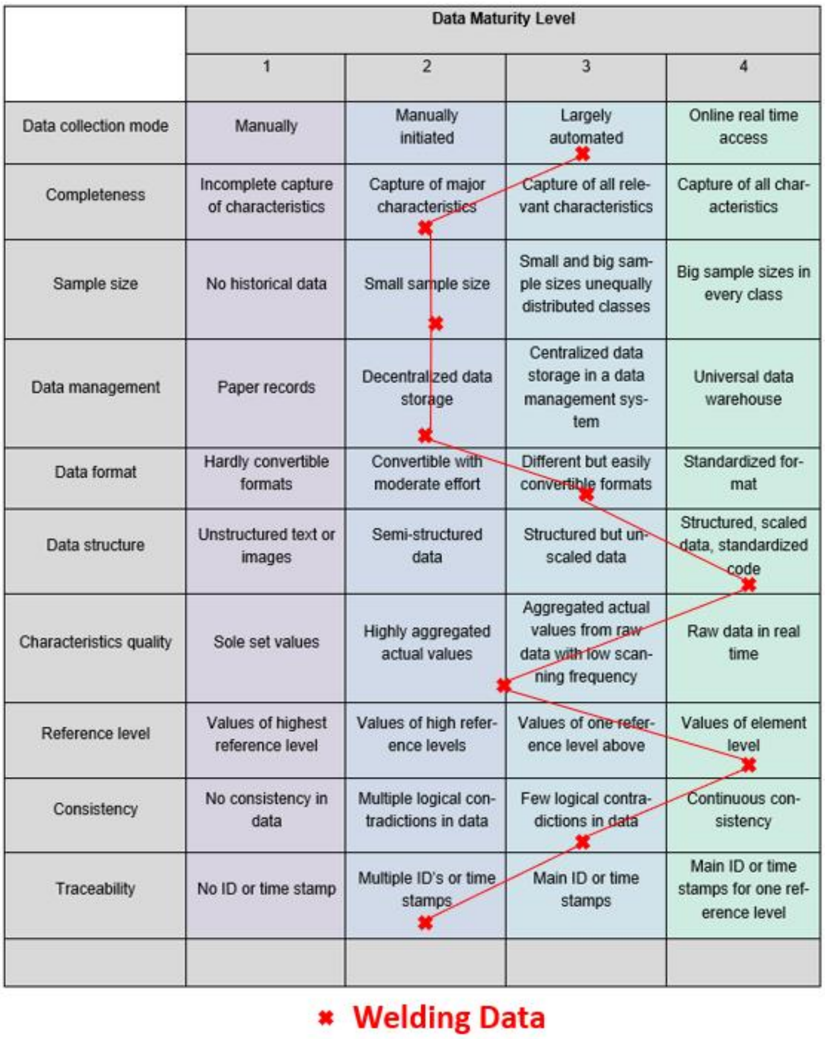
Currently, data storage is file based, hence no central data base is available. Welding data is stored on site without online access and quality data is only accessible through the key person of the QS Box. However, quality data will soon also be accessible through the data management system of the RFID Tags.

Data format comes in 2 different formats. Welding data comes in CSV files provided by the PDM Link and quality data in form of XLSX files. In the files several features must be converted from strings to datetimes, as well as from numbers to datetimes. Conversions to a standardized format can be done easily with Python Pandas.

Data structure is well organized and standardized as both data sets are exports of data management systems.

The current welding data set consists of aggregated actual values from the PDM Link. The PDM Link is able to record high-resolution actual values but cannot provide the raw data of this measurements. In the export files only highly aggregated mean maximum and minimum values are

Figure 4-11: Welding data evaluated as described in 3.4.1



	Data Maturity Level			
	1	2	3	4
Data collection mode	Manually	Manually initiated	Largely automated	Online real time access
Completeness	Incomplete capture of characteristics	Capture of major characteristics	Capture of all relevant characteristics	Capture of all characteristics
Sample size	No historical data	Small sample size	Small and big sample sizes unequally distributed classes	Big sample sizes in every class
Data management	Paper records	Decentralized data storage	Centralized data storage in a data management system	Universal data warehouse
Data format	Hardly convertible formats	Convertible with moderate effort	Different but easily convertible formats	Standardized format
Data structure	Unstructured text or images	Semi-structured data	Structured but unscaled data	Structured, scaled data, standardized code
Characteristics quality	Sole set values	Highly aggregated actual values	Aggregated actual values from raw data with low scanning frequency	Raw data in real time
Reference level	Values of highest reference level	Values of high reference levels	Values of one reference level above	Values of element level
Consistency	No consistency in data	Multiple logical contradictions in data	Few logical contradictions in data	Continuous consistency
Traceability	No ID or time stamp	Multiple ID's or time stamps	Main ID or time stamps	Main ID or time stamps for one reference level

provided. For that reason, completeness of welding data rated 2. The quality data with 44 measurement points can be rated 3 as this is a very in-depth measurement.

Reference level describes at which position the data is captured. In the case of this project, data is captured directly at the component and therefore element level which is optimal (rating 4).

Consistency in data can be rated 3, as only little contradictions were found. Contradictory are minimum values that tend to be very close to zero and set values that do not change throughout the data set which makes them not useable. Although these small contradictions a rating of 3 seems appropriate.

As already described in chapter 3.2, traceability is a big issue in the industrial environment. In the case of this project, the problem is to trace back quality data to a certain set of welding data. This task has turned out to be more complicated than expected as time stamps in the process are not sufficient to trace back components. For the moment, no main time stamp or ID can be provided to trace back components.

In summary, the data provided by the PDM Link and QS Box can be described as sufficient to get interesting new insights into the data and to apply unsupervised learning. Completeness, characteristics quality and traceability should be considered as open task and can be improved with reasonable effort.

4.6 Data Preparation

A description of this task can be found in 2.4. This task includes everything that is necessary to provide and refine the data that is finally used for the model.

Data Preparation is divided in the following subtasks:

- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

At this point of the project it had been decided, that the initial plan of extracting previously unknown knowledge out of welding data and identifying influencing factors of dimensional deviations as described in 4.3.3 can not be carried out as intended. This is owed to the fact that not enough data is available and that the connection between welding data and quality data is not yet stable enough. So far only one single boom was scanned in the Roy55 as well as in the QS box, whereby the welding data in this case was not properly recorded. As determined in Table 5, in this case the modelling of an outlier detection for the welding data will be carried out.

4.6.1 Select Data

Rationale for Inclusion or Exclusion of Data

As mentioned in the data discription report in 4.5.2 the data set coming from the PDM Link consists of 72 features. The features can be found in Table 8. The PDM Link is able to provide data for up to two welding units (power sources) on one robot. All features with a “(2)” in the name are not information-bearing, as only power source 1 delivers data. All data coming from power source 2 can therefore be excluded. All features with set values can be excluded as well, as they don’t change throughout the whole data set. Properly adjusted, set values would change with the “Schweissliste” feature and could be used to calculate an additional feature showing the discrepancy between set and actual value. Although considerable time has been spent on it, it was not possible to fix this issue. Also excluded were all columns with minimum values due to high noise. As shown in 4.5.3 minimum values cause a lot of noise within the data as these values always are very close to zero. As said in 4.5.4, features between “Pulsfrequenz” and “Grundstrom” should give information about the fine tuning of the welding process. Unfortunalty no information is captured. It was not possible to fix this issue. As a result these features must also be excluded.

This first exclusion of non information bearing features results in a much slender data set with high information content. Of 72 features at the beginning 41 can be excluded. Table 8 lists all features coming from the PDM Link whereby features in red/italic were excluded.

Table 8: Features provided by the PDM Link (excluded features in red and italic)

Roboter	*Strom Min [A] (1)*	*Gasfluss Min [L/min] (1)*
Seriennr.	*Strom Min [A] (2)*	*Gasfluss Min [L/min] (2)*
Typ	*Spannung Soll [V] (1)*	Drahtdurchmesser [mm] (1)
Zeitstempel	*Spannung Soll [V] (2)*	*Drahtdurchmesser [mm] (2)*
Bauteil nachbearbeitet	Spannung Mittel [V] (1)	Gas (1)
Fehler Bauteil	*Spannung Mittel [V] (2)*	*Gas (2)*
Gesamtzeit Bauteil	Spannung Max [V] (1)	Werkstoff (1)
Nahtnr.	*Spannung Max [V] (2)*	*Werkstoff (2)*
Beginn Naht	*Spannung Min [V] (1)*	Regelung (1)
Ende Naht	*Spannung Min [V] (2)*	*Regelung (2)*
Dauer Naht	*Drahtvorschub Soll [m/min] (1)*	*Pulsfrequenz [Hz] (1)*
Fehler Naht	*Drahtvorschub Soll [m/min] (2)*	*Pulsfrequenz [Hz] (2)*
Segmentnr.	Drahtvorschub Mittel [m/min] (1)	*Pulszeit [ms] / Fokuslagenverstellung [V] (1)*
Beginn Segment	*Drahtvorschub Mittel [m/min] (2)*	*Pulszeit [ms] / Fokuslagenverstellung [V] (2)*
Ende Segment	Drahtvorschub Max [m/min] (1)	*Pulsspannung [V] / Laserleistung [%] (1)*
Dauer Segment	*Drahtvorschub Max [m/min] (2)*	*Pulsspannung [V] / Laserleistung [%] (2)*
Schweissliste	*Drahtvorschub Min [m/min] (1)*	*Pulsstrom [A] (1)*
Dauer Schweissliste	*Drahtvorschub Min [m/min] (2)*	*Pulsstrom [A] (2)*

Strom Soll [A] (1)	*Gasfluss Soll [L/min] (1)*	*Spannung [V] (1)*
Strom Soll [A] (2)	*Gasfluss Soll [L/min] (2)*	*Spannung [V] (2)*
Strom Mittel [A] (1)	Gasfluss Mittel [L/min] (1)	*Grundstrom [A] (1)*
Strom Mittel [A] (2)	*Gasfluss Mittel [L/min] (2)*	*Grundstrom [A] (2)*
Strom Max [A] (1)	Gasfluss Max [L/min] (1)	Geschw. (Soll) [cm/min] (1)
Strom Max [A] (2)	*Gasfluss Max [L/min] (2)*	*Geschw. (Soll) [cm/min] (2)*

For the outlier detection machine learning model only numeric values and values with a variance greater than zero are of interest. Features providing information such as time stamps, material, robot number, etc. can therefore be excluded as well. The 10 features written in green were used in the model for outlier detection along with four other derived features.

4.6.2 Clean Data

Data cleaning report

Some rows must be dropped as they don't bear information. For example if the robot is moving without welding, this movement is captured but without collecting data. In this case only zero values or NaN's are captured. Rows with an entry "NaN" (not a number) were dropped as they bear no additional information. Similary, rows with mean values either for current or voltage equal to zero were dropped as they also bear no additional information. The feature "Fehler Naht" indicates errors that occurred during welding a seam. This can include robot errors or power source errors, e.g. if the welding program is paused or interrupted. This results in longer welding times or other considerable different welding data. Consequently all rows of this features with values not equal to 0 were excluded in order to reduce noise.

4.6.3 Construct Data

Derived Features

Generated Records

Originating from the initial features three new features were derived. The features "Leistung Mittel [W] (1)" and "Leistung Max [W] (1)" coming from power source 1 provide information about the electric power. The resulting unit is Watt [W].

$$P = U * I \qquad \begin{aligned} &P \ldots el.power \\ &U \ldots Voltage \\ &I \ldots Current \end{aligned} \tag{3}$$

Another feature derived from the initial data is the feature "Energie Mittel [kJ] (1)", giving information about the Energy that was used for one segment. It can be calculated by multiplying the electric power and welding time of the segment. The resulting unit is Kilojoule [kJ].

$$E = \frac{P * t_{seg}}{1000} \qquad \begin{aligned} &E \ldots Energy \\ &t_{seg} \ldots welding\ time\ of\ a\ segment \end{aligned} \tag{4}$$

The third feature derived is named “Streckenenergie Mittel [kJ/mm] (1)”. This feature provides information about the energy that was introduced into the welding seam per unit of length which in turn gives information about the temperature cycle and the size of the heat affected zone. It is assumed that the velocity of the robot is constant throughout the welding process.

$$E_{Temp} = \frac{P * 60}{v * 10000} \qquad \begin{array}{l} E_{Temp} \ldots Energy\ per\ unit\ length \\ v \ldots Velocity\ of\ the\ robot\ [cm/min] \end{array} \qquad (5)$$

Besides the features described above, no other records were generated.

4.6.4 Integrate Data

Merged Data

As described in 4.5.2, every component generates one or more data sets that can be exported after the welding process from the PDM Link in form of a CSV file. These files are stored in a separate folder and then merged together in chronological order. The output file contains all instances of the individual CSV files.

4.6.5 Format data

Reformatted Data

All records in the data set come in form of string’s or floats. In order to proceed the data in practical ways the strings containing the time stamps and durations of welding seams had to be converted to Pandas datetimes and timedeltas. Pandas, the software library used to pre-process and manipulate the data, provides usefull functions to convert strings or numbers to datetimes. Welding times were converted to timedeltas with seconds as unit.

4.7 Modeling

This task encompasses everything from selecting the right algorithm to adjusting and assessing the model. Modeling is divided in the following subtasks:

- Select modeling technique
- Generate test design
- Build model
- Assess model

4.7.1 Select modeling technique

Modeling technique

Modeling assumptions

With the selected features as shown in 4.6, a Principle Component Analysis (PCA) will be carried out. PCA is a popular tool to reduce the dimension of the data set, to structure and to visualize it [64]. In our case the fourteen-dimensional data set (14 features) can be reduced to a two- or three-dimensional data set in order to visualize the data while still retaining most of the information. The resulting features are called components of the PCA. The components are a mixture of the original features and are orthogonal to each other. By visualizing data this way, clusters, patterns, or outliers can be made visible.

Besides a better visualization the PCA has even more advantages. Firstly, two- or three-dimensional data can be processed much faster than higher dimensional data, which leads to reduced computing time. Secondly, by calculating the covariance matrix, which contains the loading vectors of the individual components of the PCA, an even deeper insight in the data can be achieved. In the loading vectors the underlying factors of the individual components are listed. In other words, in the loading vectors the features containing most information related to a specific problem can be found [64].

Finally, the results of the PCA can be used to implement an outlier detection. If the PCA shows patterns in the plot, there are similarities within the data. If these patterns are well separated, a model can be made were new instances are classified according to their match or mismatch with the existing patterns. To do so it is essential to eliminate outliers before fitting the model to the patterns found by the PCA. Otherwise the model would be negatively affected by outliers in the training set and could not effectively detect outliers in the test set.

Besides eliminating outliers by the measures taken as described in 4.6.2 another possibility to clean the dataset is the application of an algorithm to find previously undetected outliers. Two different algorithms will be tested. One algorithm is DBSCAN, a density-based clustering algorithm introduced in 1996. The underlying key idea of DBSCAN is “that for each point of a cluster the neighbourhood of a given radius has to contain at least a minimum number of points, i.e. the density in the neighbourhood has to exceed some threshold” [65]. Its advantage in contrary to other clustering algorithms such as Kmeans is, that it can handle clusters with arbitrary shape and detect outliers. However, if local densities vary in different regions of the plot an algorithm with a single density level such as DBSCAN might encounter problems with describing well all clusters [66].

Local outlier factor (LOF) is the second algorithm that will be applied on the training data set. It is also density-based, but in contrast to DBSCAN, where being outlying is a binary property, LOF assigns to each instance an outlier factor indicating the degree of being outlying [67]. The user can influence the model by adjusting several parameters such as the k-nearest neighbours that are used to calculate the density and a threshold for the outlier factor can be set, as of which an instance will be considered as outlying. The advantage of LOF to other outlier detection algorithms is, that it can handle varying densities throughout the dataset [68]. LOF has already been applied successfully in outlier detection projects such as network intrusion detection [68] and is therefore expected to deliver better results than DBSCAN.

Finally, another model using “one-class SVM” algorithm will be applied and compared to the results of LOF. One-class SVM is a support vector machine based algorithm that can be used to detect outliers and learn boundaries and can then be used to classify new instances based on these boundaries.

4.7.2 Generate test design

Test design

This chapter is dedicated to the test design of the model, or on how to evaluate the model and assess its quality. For supervised learning a common test design is the k-fold cross validation. For this evaluation method the data set is split into k similar sized parts. In the following one part is always used to test the model and the other parts to train the model. A common practice is to split the set into $k = 5$ parts, meaning 80% of the data is used for training and 20% for testing [69]. Every data point must be used for testing once. For $k = 5$ the evaluation is carried out 5 times with different training and test sets. The best performing data set is then used for training the model. K-fold cross validation will not be used in this project, as unsupervised learning is used.

Another possibility to assess the outlier detection model is to deploy a confusion matrix (see Figure 4-12). In the case of outlier detection the matrix is a $2\ x\ 2$ Matrix as there are only two classes. Outlying instances and inlying ones. If an outlier is predicted correctly, it is rated *True positive.* If the model predicts an outlier to be an inlier this is rated *False negative.* Actual inliers that are classified as outliers are *False positives* and inliers that are correctly classified are *True negatives.* Ideally, if everything is correctly classified the off-diagonale entries would be zero [5].

Figure 4-12: Confusion Matrix for two classes inspired by [5]

		Predicted Class		
		Positiv	**Negative**	**Total**
Actual Class	**Positive**	*True positive*	*False negative*	n
	Negative	*False positive*	*True negative*	m
	Total	n’	m’	

As the model is build on unsupervised learning, no class labels are available, meaning, to use the above described evaluation methods, labels have to be generated. To do this, the results of the model will be analysed by reviewing all instances of the test data set. If welding data of the instances that are classified as outlying show a percentage difference greater than 15% from the average value in at least two different features, the classification will be counted as true positiv. If the percentage difference of instances that are classified as outlying does not exceed 15% from the average in more than one feature, these instances will be counted as false positives. False negatives will be instances that show a percentag difference of more than 15% from the average in at least two different features and that are classified as inlying. True negatives are instances classified as inlying that show no percentag difference of more than 15% from the average in more that one feature.

4.7.3 Build model

For the PCA only one parameter setting was changed from default settings. The parameter “n_components” sets the number of components to be kept. This parameter was set to 3 components. The same applies for the Kmeans algorithm. The parameter “n_clusters” is the only parameter not on default setting and sets the number of clusters to form. This parameter was set on 20 clusters as the number of unique welding parameters that can be found in the “Schweissliste” feature is 20. It is hoped that Kmeans can identify the different welding parameters.

Most important parameter for DBSCAN is the eps parameter. It is “the maximum distance between two samples for one to be considered as in the neighbourhood of the other” [70] Default value for this parameter is 0.5, which is also the value that gives the best results in this model. Another parameter is “min_samples”. This Parameter influences “the number of samples (or total weight) in a neighbourhood for a point to be considered as a core point” [70]. As the sample size for the training set is 18 this parameter was set 15, slightly beneath the sample size. The parameter must be set below the sample size as otherwise, in case of outliers, the whole cluster would be classified as outlying. The parameter “n_jobs” is the parameter for how many processors are used to calculate the model. This parameter was set -1 which means all processors are used for calculation.

For the algorithm “Local outlier factor” two parameters were set. “n_neighbors” is a parameter similar to DBSCAN’s “min_samples”. It was set 15 as this is slightly below the training sample size of 18. The parameter “contamination” influences the threshold of the decision function, if a point is considered as outlier or not. This parameter was set ‘auto’ which gives the best results.

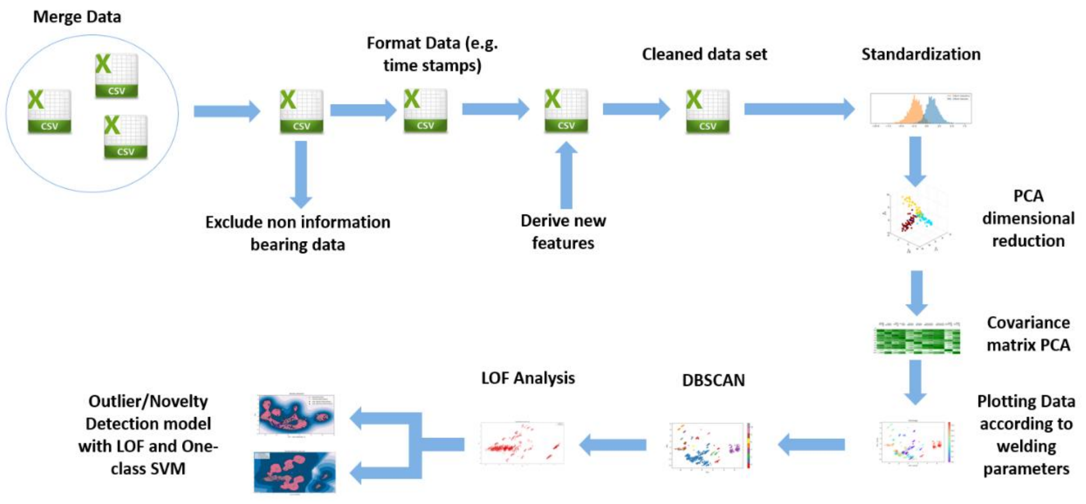
The only parameter set differently from default for the “One-class SVM” algorithm is the “nu-parameter”. This parameter sets an upper bound for error probability in the training set and a lower bound for the number of support vectors [71]. As the training set has already been cleaned from outliers the “nu-parameter” can be set at a low value to keep the rate of misclassified training examples low. All other parameters for this algorithm were left set on default.

Table 9: Parameter settings

Algorithm	Parameters	Set value	Explanation
PCA	Number of components	3	3 components represent around 85% of information, which can be considered as sufficient. Furthermore, 3 components can be visualized in informative plots.
Kmeans	Number of clusters to form	20	According to 20 existing unique entries in the “Schweissliste” features.
DBSCAN	Eps	0.5	Default value gives the best results.
	min-samples	15	The sample size of welded booms in the training set is 18. The set value is slightly below.
	n_jobs	-1	Parameter setting to use all processors to calculate the model and not just one as default.
Local outlier factor	n_neighbors	15	Slightly below the training sample size of 18
	contamination	‘auto’	Gives back the best results.
One-class SVM	nu	0.01	Sets a narrow boundary

The procedure of the project has been described in the previous chapters. In Figure 4-13 the project steps can be seen in an overview.

Figure 4-13: Overview of project steps


Merge Data
CSV
CSV
CSV
Format Data (e.g. time stamps)
Cleaned data set
Standardization
Exclude non information bearing data
Derive new features
PCA dimensional reduction
Covariance matrix PCA
Plotting Data according to welding parameters
DBSCAN
LOF Analysis
Outlier/Novelty Detection model with LOF and One-class SVM

After cleaning and constructing the data set the data has to be standardized using Scikit-learn’s “StandardScaler” tool. Many machine learning tools need standardized or normalized data to work correctly [72]. Standardizing the features means to rescale the data, remove the mean value (mean value is zero) and scaling the variance to a unit variance (standard deviation is 1). Normalization would be to rescale the data to fit it into a value between 0 and 1.

After standardizing the data, the PCA can be carried out, reducing the dimension of the training set from a $14\ x\ 1274$ matrix to a $3\ x\ 1274$ matrix. The explained variance is an indicator for the information that has been kept after the dimensional reduction. The first three components account for 84,5% of the total explained variance, which means that around 85% of the information has been kept although the number of features has been reduced by almost four-fifths.

Figure 4-15: Explained variance, PC1: 60.4%, PC2: 14.2%, PC3: 9.9%

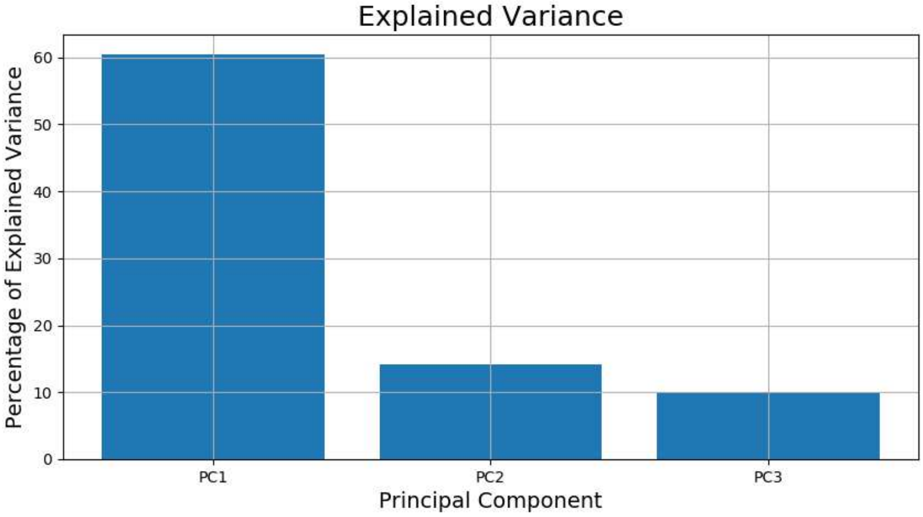
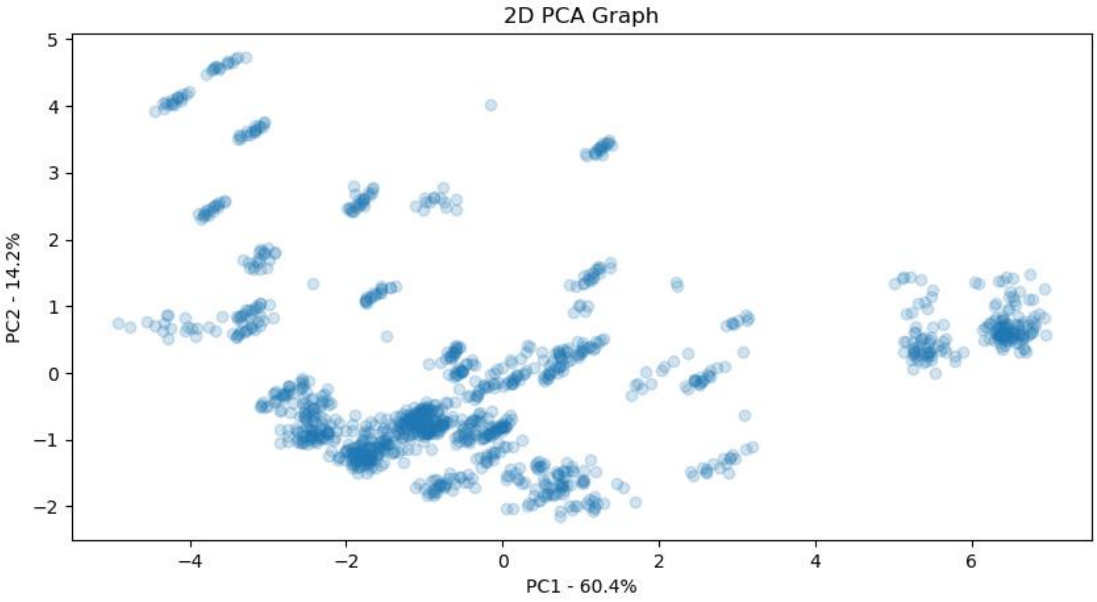


Figure 4-14: 2D density graph of the training data set after dimensional reduction

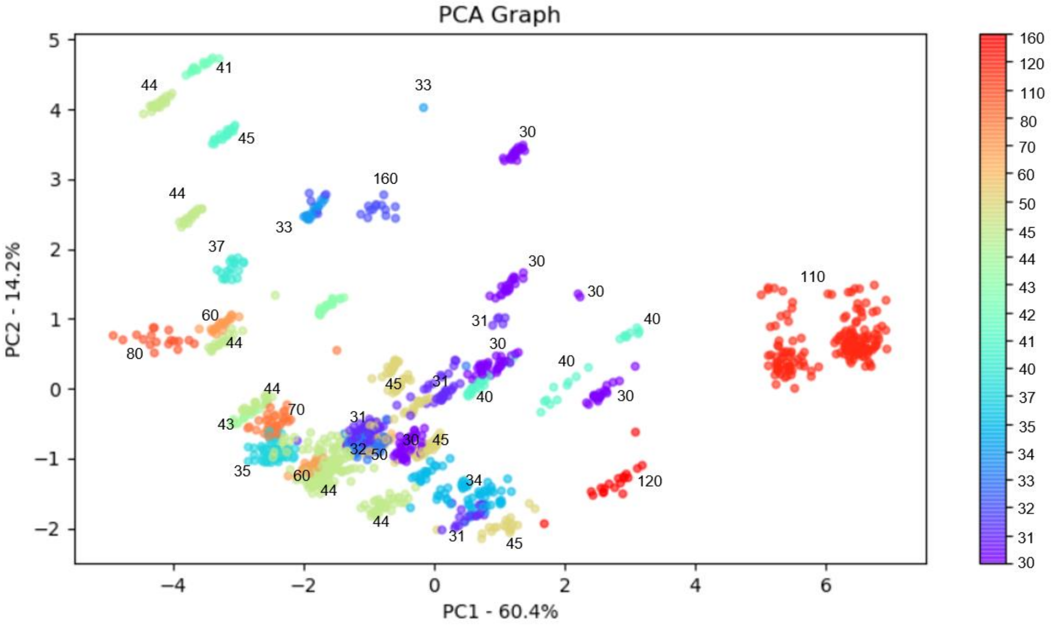


In [73] the author explains that the explained variance for further processing should be at least 60%. As can be seen in Figure 4-15. already the first two components account for 74,6% of explained variance and should therefore be valid for further modeling. If the third component is used, the explained variance rises to 84,5%.

In Figure 4-14 a plot of the first two PCA components can be seen. Every instance (one segment welded by the welding robot) represents a data point in the scatter plot. In the plot several clusters can be found which indicates that patterns exist within the data. It is assumed that the cluster match well with the welding parameters as listed in the feature, as these parameters are set for each segment.

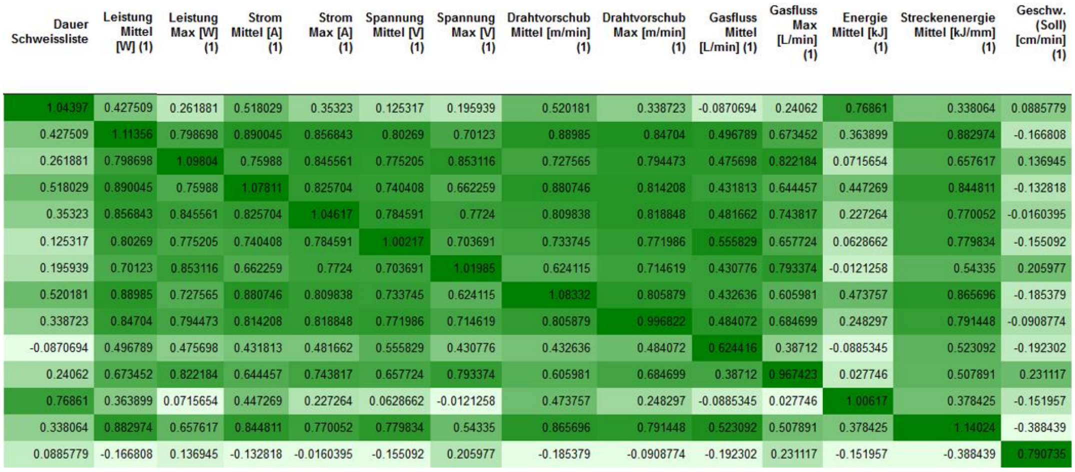
Figure 4-16 shows that the pattern in the plot match well with the welding parameters. Some welding parameters form more than one cluster, which could be related to the welding time as the same welding parameters are used for different segments.

Figure 4-16: 2D graph of the training data set with colours according to welding parameters

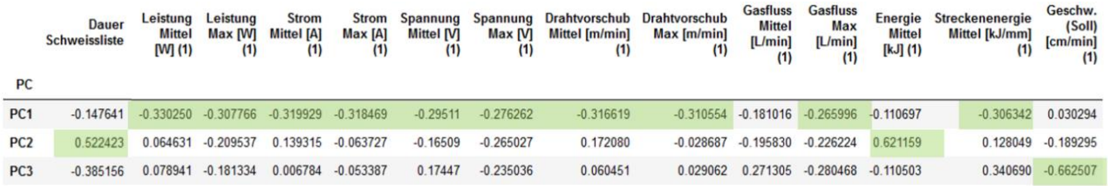


In Figure 4-17 the covariance matrix and the loading matrix of the PCA are shown. The covariance matrix with a size of $p\ x\ p$ is, where p is the number of features, is the first step to calculate the principal components of the PCA [74]. The elements in the diagonal of the covariance matrix are the variances, the entries outside the diagonal are the correlations. Similar to the heatmap, that was used to explore the data in 4.5.3, correlations within the data can be seen. For example, in the first row, the welding time correlates with total energy consumption that was used. Likewise, it can be seen that energy per unit length correlates positively with the electric power and negatively with the welding speed.

Figure 4-17: Covariance matrix (above) and loading matrix (below) of the PCA



Dauer Schweissliste	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie Mittel [kJ/mm] (1)	Geschw. (Soll) [cm/min] (1)
1.04397	0.427509	0.261881	0.518029	0.35323	0.125317	0.195939	0.520181	0.338723	-0.0870694	0.24062	0.76861	0.338064	0.0885779
0.427509	1.11356	0.798698	0.890045	0.856843	0.80269	0.70123	0.88985	0.84704	0.496789	0.673452	0.363899	0.882974	-0.166808
0.261881	0.798698	1.09804	0.75988	0.845561	0.775205	0.853116	0.727565	0.794473	0.475698	0.822184	0.0715654	0.657617	0.136945
0.518029	0.890045	0.75988	1.07811	0.825704	0.740408	0.662259	0.880746	0.814208	0.431813	0.644457	0.447269	0.844811	-0.132818
0.35323	0.856843	0.845561	0.825704	1.04617	0.784591	0.7724	0.809838	0.818848	0.481662	0.743817	0.227264	0.770052	-0.0160395
0.125317	0.80269	0.775205	0.740408	0.784591	1.00217	0.703691	0.733745	0.771986	0.555829	0.657724	0.0628662	0.779834	-0.155092
0.195939	0.70123	0.853116	0.662259	0.7724	0.703691	1.01985	0.624115	0.714619	0.430776	0.793374	-0.0121258	0.54335	0.205977
0.520181	0.88985	0.727565	0.880746	0.809838	0.733745	0.624115	1.08332	0.805879	0.432636	0.605981	0.473757	0.865696	-0.185379
0.338723	0.84704	0.794473	0.814208	0.818848	0.771986	0.714619	0.805879	0.996822	0.484072	0.684699	0.248297	0.791448	-0.0908774
-0.0870694	0.496789	0.475698	0.431813	0.481662	0.555829	0.430776	0.432636	0.484072	0.624416	0.38712	-0.0885345	0.523092	-0.192302
0.24062	0.673452	0.822184	0.644457	0.743817	0.657724	0.793374	0.605981	0.684699	0.38712	0.967423	0.027746	0.507891	0.231117
0.76861	0.363899	0.0715654	0.447269	0.227264	0.0628662	-0.0121258	0.473757	0.248297	-0.0885345	0.027746	1.00617	0.378425	-0.151957
0.338064	0.882974	0.657617	0.844811	0.770052	0.779834	0.54335	0.865696	0.791448	0.523092	0.507891	0.378425	1.14024	-0.388439
0.0885779	-0.166808	0.136945	-0.132818	-0.0160395	-0.155092	0.205977	-0.185379	-0.0908774	-0.192302	0.231117	-0.151957	-0.388439	0.790735



	Dauer Schweissliste	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie Mittel [kJ/mm] (1)	Geschw. (Soll) [cm/min] (1)
PC														
PC1	-0.147641	-0.330250	-0.307766	-0.319929	-0.318469	-0.29511	-0.276262	-0.316619	-0.310554	-0.181016	-0.265996	-0.110697	-0.306342	0.030294
PC2	0.522423	0.064631	-0.209537	0.139315	-0.063727	-0.16509	-0.265027	0.172080	-0.028687	-0.195830	-0.226224	0.621159	0.128049	-0.189295
PC3	-0.385156	0.078941	-0.181334	0.006784	-0.053387	0.17447	-0.235036	0.060451	0.029062	0.271305	-0.280468	-0.110503	0.340690	-0.662507

The loading matrix below is calculated by the means of the covariance matrix. This matrix contains the factors that load on the individual principal components of the PCA. Hence, by analysing the loading matrix, the features that have the most impact on the individual principal components can be found. For instance, mean electric power has the highest impact on the first component, followed by the mean and maximum values of the current and the max maximum value of wire feeding rate. The first component is therefore mainly influenced by the factor current energy consumption, including several features that affect this factor. The second principal component is mainly loaded by welding time and total energy consumption. The underlying factor for PC2 can therefore be called total energy consumption. The third principal component is loaded by speed. This is the third factor that influences the PCA. Together these three factors account for almost 85% of the information.

Finding the underlying factors, also called latent variables [75], can help to interpret correlations between data and provide information about the most striking features in the dataset. Especially if the data set has a high number of features finding latent variables that sum up several features can reveal valuable new insights.

In the plots Figure 4-14 and Figure 4-16 besides clusters in the welding data also individual points that lie apart from the clusters. These instances apparently show deviations in their features from average values. For the implementation of an outlier detection model it is crucial to find and exclude outliers before training the model to the training set [74], as otherwise the boundary in the vector space between inlying and outlying points would be negatively affected.

Several criteria have been invented in order to formalize the classification of outliers. Two criteria were used on the training data set and compared. The DBSCAN algorithm and local outlier factor (LOF).

Figure 4-18: DBSCAN Outlier detection with outliers in red colour

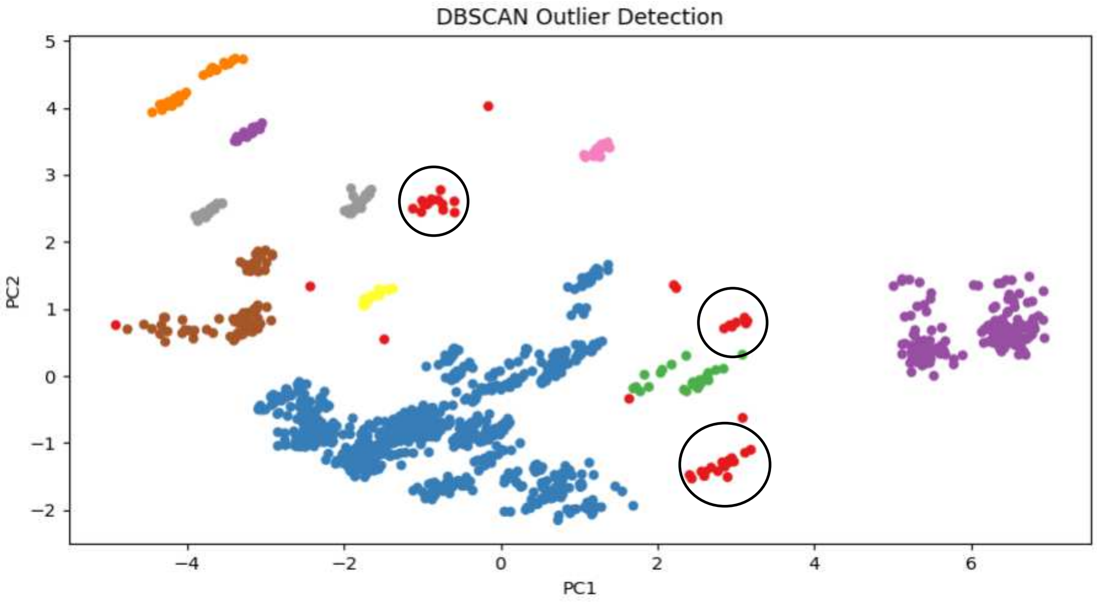
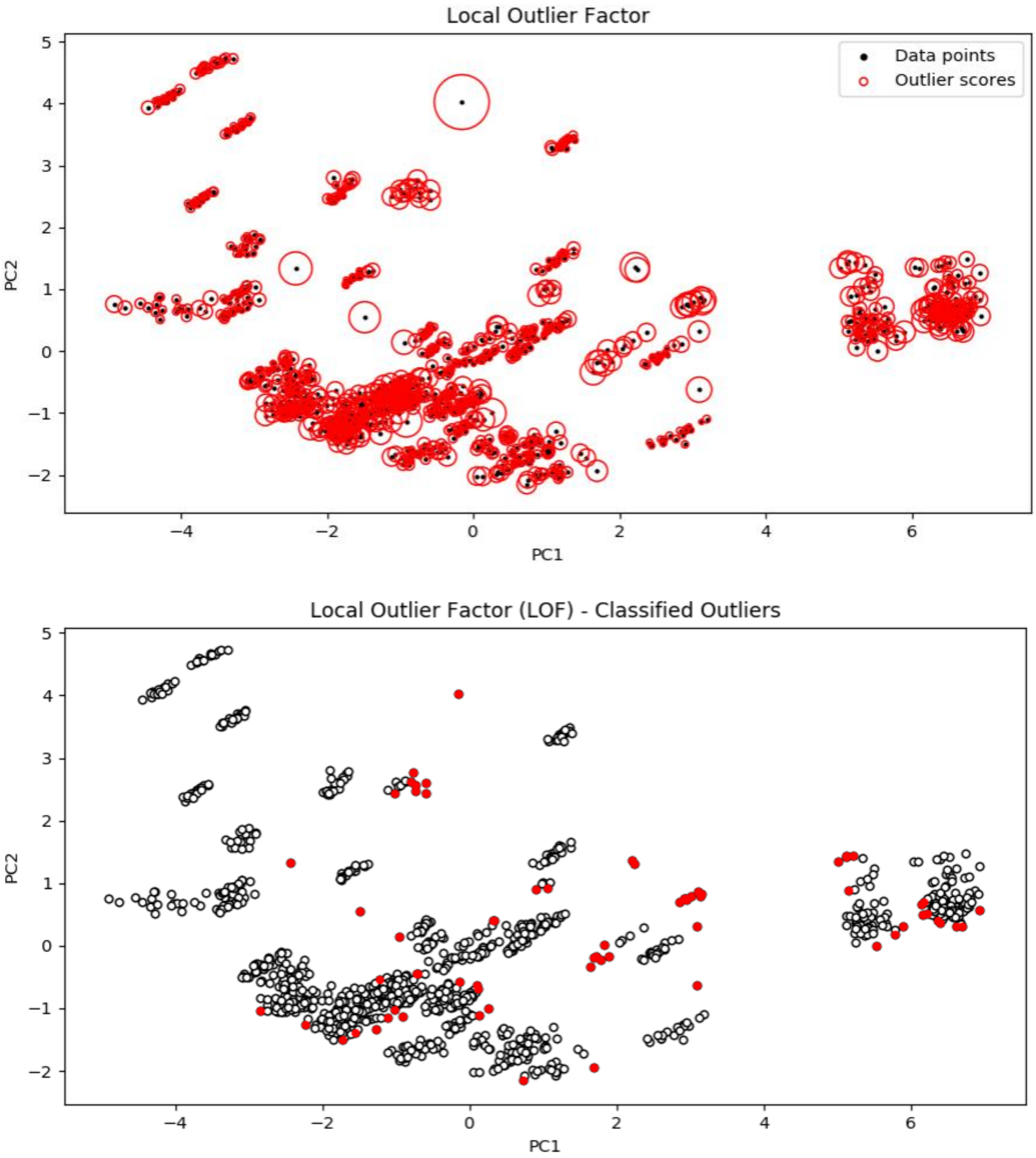


Figure 4-18 shows the results of the outlier detection with DBSCAN. Instances labelled as outlying are coloured in red. Several clusters that do not seem to be outliers were classified so. The respective clusters are encircled in the plot. Two reasons seem to play a decisive role. Firstly, as mentioned above the application of DBSCAN can lead to problems in the case that densities of the clusters differ strongly. In Figure 4-14 the density of the cluster is shown. The encircled clusters show low density. Another problem is that the number of currently 18 booms that have been captured so far is probably not enough and should be increased in order to properly classify outliers.

To get a second local outlier factor algorithm has also been applied. As already mentioned, this algorithm shows two main differences to DBSCAN. Its classification is not a binary one, outlying or not outlying. In the first step it assigns a factor to each point according to its probability of being outlying. After adjusting a threshold, it is decided whether a point is outlying or not. The other difference is its ability to handle different local densities.

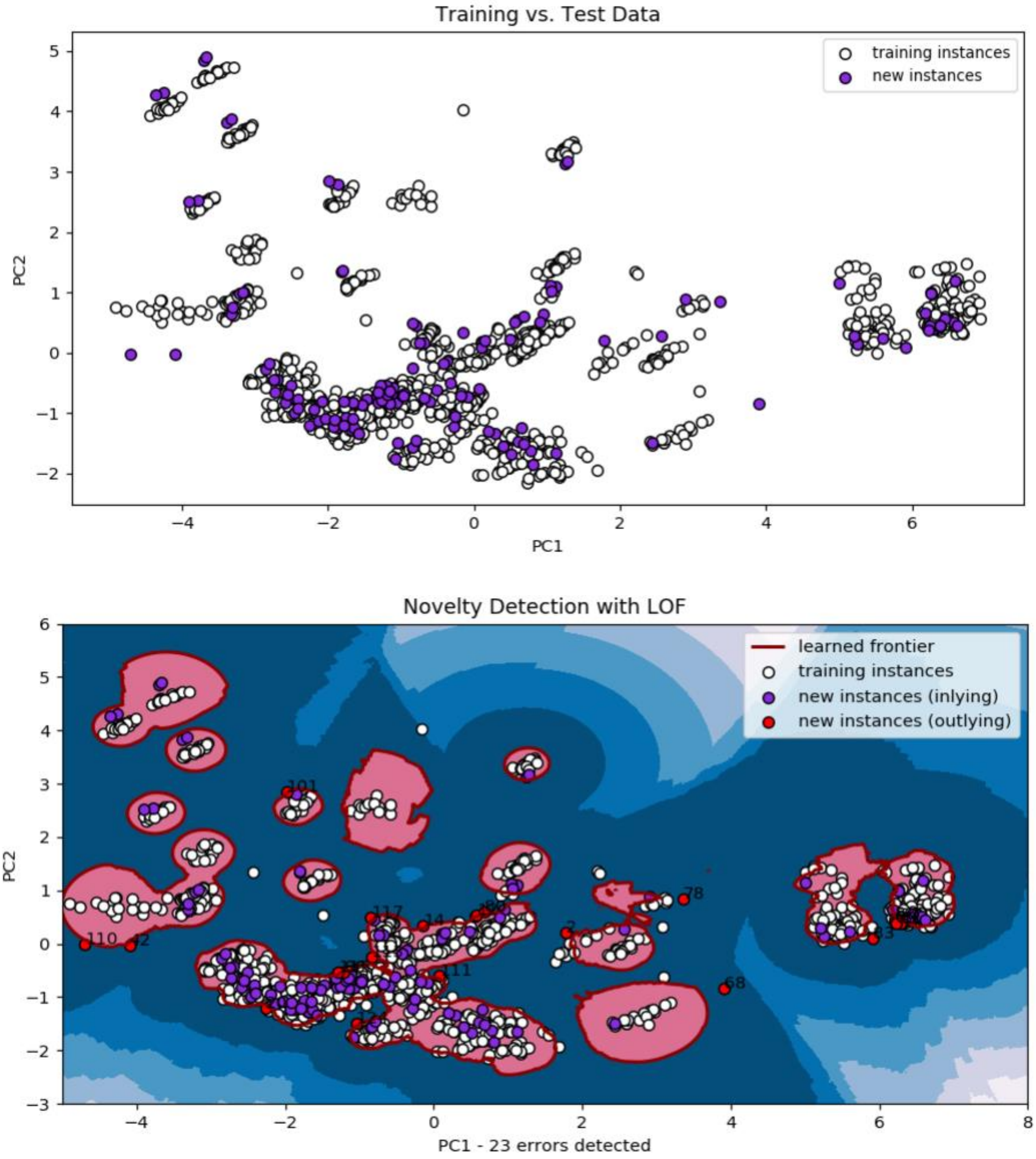
In Figure 4-19 on the upper plot the local outlier factor of each instance is visualized by the size of the circle surrounding the instance. Big circles indicate a high outlier factor. The plot below shows the instances that have been classified as outlying. The results differ considerable from the results found by DBSCAN. More instances have been classified as outlying and the regions of outliers differ. Especially on the side-lines of clusters many instances have been classified as outlying. Outliers are also more evenly spread over the whole plot. Unlike DBSCAN, LOF is less prone to class whole clusters as outlying but might be too sensitive to local differences in density which could lead to misclassified points.

Figure 4-19: Local outlier factor (LOF)



After using LOF to calculate the local outlier factor of each instance in the training set, Scikit Learn provides a method called “decision function” which returns information about the threshold that classifies a point as inlying or outlying. New, previously unseen data can then be classified according to the boundaries learned from the training data. The new data comes in form of welding data of two new booms. It acts as the test data set in order to evaluate the model learned from the training data. As the training data set consists of welding data from 18 booms, the test data set comprises data accounting for around 10% of the size of the training data set. In the upper plot of Figure 4-20 both data sets can be seen.

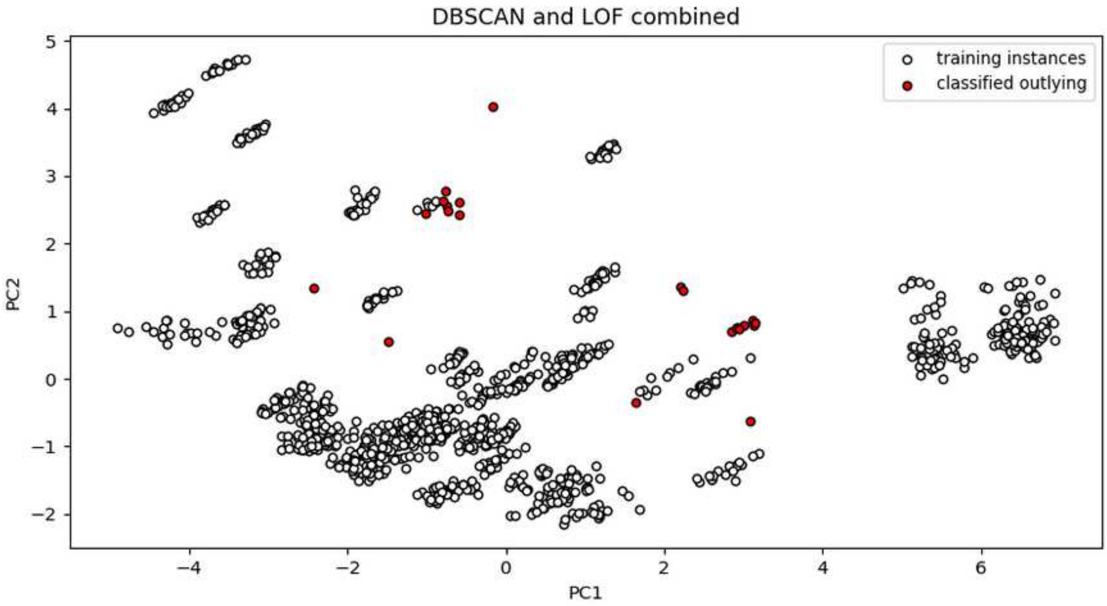
Figure 4-20: Training and test data instances (above), Novelty Detection with LOF (below)



In the lower plot of Figure 4-20 the classification of the test data according to the boundaries trained from the training data can be seen. The model has been trained with LOF. This is the reason why clusters with lower density tend to have larger boundaries and dense clusters have more narrow boundaries. From 136 instances in the test data set, the model classifies 23 instances as outlying. This looks quite high but can be attributed to the fact, that the model would need more training data to improve classification task and set the boundaries more accurately. Nevertheless, instances that seem outlying on a first glance, such as number 110, 42 or 68 have been classified outlying, showing an overall functioning of the model.

In order to improve the model another approach has been tested. As mentioned above, the exclusion of outliers is crucial for a proper implementation of outlier detection. For this reason, both outlier detection strategies, DBSCAN and LOF, have been used and instances classified as outlying from both algorithms have been excluded. In Figure 4-21 the results are shown. Two clusters have been classified as outlying although it can be assumed that if of more instances would be available, these clusters would be classified as inlaying.

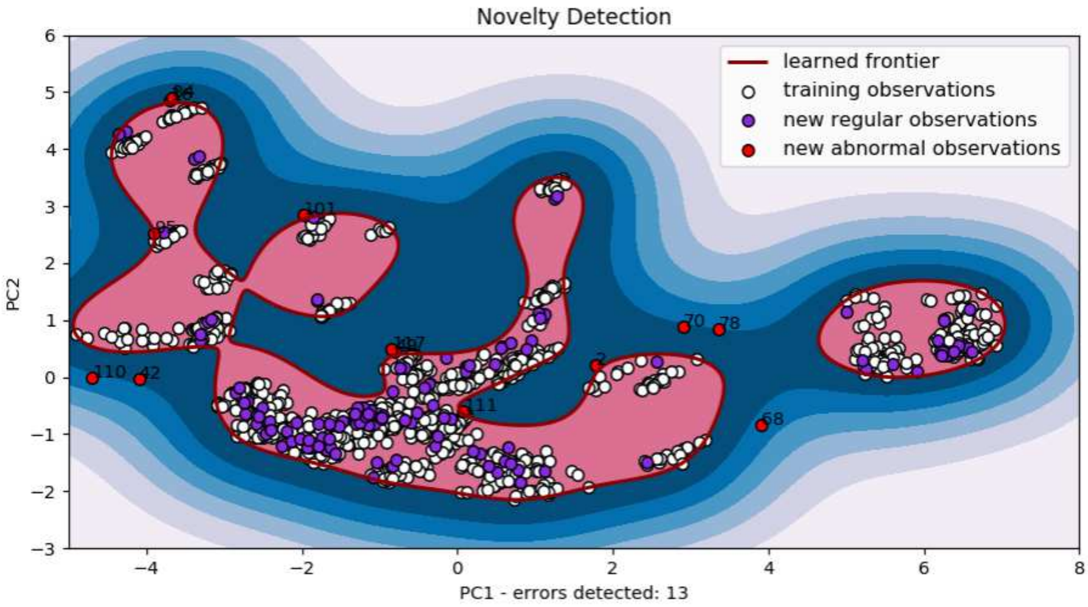
Figure 4-21: Combination of DBSCAN and LOF to find outliers



After excluding outliers classified by both algorithms, another algorithm called “One-class SVM” is applied to learn boundaries in order classify the test data set. In Figure 4-22 the results are shown. With this model, from 136 instances in the test data set, 13 have been labelled as outlying.

Both models show functioning when classifying outliers and can be used to detect outliers in the welding process. More instances are needed to improve accuracy of classification. If more training data can be provided, accuracy should improve considerable for both models.

Figure 4-22: *Novelty detection with One-class classification*



4.7.4 Assess model

In 4.7.2 a test design to assess the models has been explained. For both models a confusion matrix will be created to compare them. As outlier detection is an unsupervised learning method, labels must be generated to evaluate the models.

For the whole test data set every feature in every instance has been compared to the average value. The test data was compared to average values of instances with the same welding list number and the same welding time. Average values for each welding list and each welding time can be found in Figure 9-1 and Figure 9-2. All values of the test data including the percentage difference for each feature and instance from the average value can be found in the tables shown between Figure 9-3 and Figure 9-10.

By applying the evaluation method proposed in 4.7.2, 23 instances have been labelled as outlying. Hence, 23 instances of the test data show a percentage difference from the average value of more than 15% in at least 2 features.

The model carried out with LOF (Figure 4-20) classifies 23 instances as outlying. 4 of them show differences in their welding data from comparable instances of the same welding parameters. These are the instances with the index 2, 68, 78 and 113. Hence, out of the 23 instances classified as outlying 4 instances are true positives and 19 are false positives. 94 instances classified as inlying are also labelled inlying by the evaluation method proposed in 4.7.2 and therefore true negatives. The 19 instances classified as inlying by the algorithm show percentage difference from the average of more than 15% in at least two features and therefore false negatives.

In the one-class SVM model 13 instances are classified as outlying. 4 of them show differences in their welding data from comparable instances of the same welding parameters, namely the instances with the index 2, 68, 70 and 78. Hence, out of the 13 instances classified as outlying, 4 instances are true positives and 9 are false positives. 104 instances were correctly classified as inlying and therefore true negatives. 19 instances are counted as false negatives as they show a percentage difference from the average of more than 15% in at least two features.

Table 10: Confusion matrix for outlier detection with LOF

		Predicted Class			
		Positive	**Negative**	**Total**	
Actual	**Positive**	4	19	23	
	Negative	19	94	113	
	Total	23	113		
	Sensitivity	$\text{True positiv rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$			0.174
	Specificity	$\text{True negativ rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}$			0.832
	False positive ratio	$\text{False positive rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$			0.168
	Positive predictive value	$\text{Positiv predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}$			0.174

Table 11: Confusion matrix for outlier detection with one-class SVM

		Predicted Class			
		Positive	**Negative**	**Total**	
Actual	**Positive**	4 TP	19 FN	23	
	Negative	9 FP	104 TN	113	
	Total	13	123		
	Sensitivity	$\text{True positiv rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$			0.174
	Specificity	$\text{True negativ rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}$			0.920
	False positive ratio	$\text{False positive rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$			0.080
	Positive predictive value	$\text{Positiv predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}$			0.308

As apparent in the two matrices Table 10 and Table 11 the one-class SVM model and the LOF model achieve equal results in terms of sensitivity. However, the one-class SVM generalizes better than the LOF model and shows therefore less false positive classifications and a higher positive predictive value. With increasing training data these numbers are assumed to improve considerable. Higher sensitivity of both models and a lower false positive ratio (false alarm ratio) of the LOF model would be need before being implemented in the production process. At large, the one-class SVM model shows better results than the LOF model and is therefore recommended for further use.

4.8 Evaluation

In 2.4 a more detailed description of this task can be found. Evaluation encompasses the evaluation of the project according to the previously set business criteria as well as a reflection on the project.

Evaluation is divided in the following subtasks:

- Evaluate results
- Review process
- Determine next steps

4.8.1 Evaluate results

In 4.4.1 the business objectives have been identified. It has been said that due to the scientific character of this project it is not possible to express the achievements of this work in economic numbers. The objective was to improve field experience in machine learning, to provide a blueprint on how to carry out machine learning projects in form of a transparent documentation and to increase sensitivity to this field.

As can be seen in Figure 4-3 different departments were involved and contributed to the project. This has led to lively debates in meetings on how machine learning could be applied and which applications would have the biggest potential. Two interim presentations and the final presentation have pointed up possibilities of machine learning in manufacturing and the challenges which have to be taken to implement machine learning projects. Results of the theoretical part provide a practicable overview of the current developments in machine learning as well as an assistance on how to select algorithms or assess the quality of data. State-of-the-art research results give a comprehensive overview on applications in manufacturing that have been carried out so far and can serve as a pool of ideas for future projects.

Results of the theoretical part were adopted to carry out the practical part to show their practical usability. In the practical part several hurdles were encountered such as the traceability between welding data and quality data or setting the set values for in the PDM Link. Likewise, it was problematic to access sufficient data for the use cases partly because no sensor systems are installed and partly because OEM's do not give access to machine data. To some extend it was possible to overcome the hurdles such as the traceability problem where good progress has been made. These are useful experience values that will enrich future projects and should therefore not be underestimated. Finally, by using a well-established industrial standard for data mining such as CRISP-DM it was possible to provide a blueprint for future machine learning projects.

Although the business success criteria cannot be numerically quantified it can be said that the business objectives have been met in the sense of improving field experience and in providing experience value for future projects.

Two models were tested and assessed. One-class SVM shows better performance as can be seen in Table 10 and Table 11. Hence, the one-class SVM model is recommended for future outlier detection. Nevertheless, both models would profit from a considerable increase in training data but unfortunately, within the timeline of this project, it was not possible to generate more data.

4.8.2 Review process

At this step the project can be reviewed. This involves a thorough re-thinking of the approach that has been chosen. In future machine learning projects, the application of the CRISP-DM standard process should be considered from the very beginning to provide a consistent framework for the whole project. Hence, already in the kick-off meeting the manual of CRISP-DM should be used to discuss more effectively business objectives as well as data mining objectives. In the case of this project, business objectives and data mining goals have been derived from the findings of the kick-off meeting but could have been defined more precisely.

In the run-up to future projects, the findings of the theoretical part, especially the methods to assess data quality of use cases can provide valuable tools to pre-assess data quality in order to set data mining goals more precisely.

4.8.3 Determine next steps

The projects business success criteria were determined among other things as improving field experience and in providing experience value for future projects. Now findings of this project should be used to realize other further projects in this direction and to deepen ever more the company's experience in this field. As findings from chapter 3.2 and 3.5 suggest, machine learning is likely to gain ground in the industrial environment in the coming years. It can therefore be assumed, that expertise developed in this field today now will later pay off.

To boost expertise in machine learning the author suggests the following measures:

- Devoting resources to this field to improve internal expertise on the field of machine learning.
- Building up a personnel pool in this field to develop internal expert knowledge.
- Enhancing cooperation with research institutions such as universities to keep the state of knowledge up to date in this rapidly developing field.
    - A possible option would be to work out a research project that can then be executed together. Carrying out a potential analysis for machine learning in manufacturing of the company could be an enriching experience for both parties.
    - Another option is to work out projects in the size of a master thesis or bachelor thesis and to assign these projects to students for execution as happened in the case of this project.
- Enhancing cooperation with other companies active in this field.
- The now established connection of welding data and quality data can be used to examine the relation of the two data sets.

5 Conclusion and Outlook

5.1 Conclusion

The theoretical part comprises fundamentals and a solid overview of the latest developments and progress made in the field of machine learning. In chapter 2 common terminology that is used in this field has been summarized. Also, types of machine learning have been explained such as supervised, unsupervised and reinforcement learning. Finally, chapter 2 gives a summary of the “cross-industry standard process for data mining”, a standardized process for machine learning and data mining projects that was later used to carry out the practical part.

Chapter 3 is dedicated to machine learning in a manufacturing environment. After a short historical review, challenges of machine learning in manufacturing have been collected and depicted in an illustrative summary (Figure 3-1). Opportunities of machine learning in manufacturing have also been examined in 3.2 and a list of popular frameworks which provide a variety of machine learning algorithms for implementation has been given in Table 1. In 3.3 an outline of machine learning algorithms is given where the algorithms are classified according to the way they operate. In Figure 3-7 the findings are summarized in an illustrative way. Chapter 3.4 proposes a scheme on how to assess data quality and a template was developed that can be used to assess data quality in an illustrative way (Figure 3-8). Also, a table can be found providing a comparison of machine learning algorithms. Finally, chapter 3.5 provides an overview of the current state-of-the-art, starting with the findings of a survey over advantages, costs and the most important applications of data analytics in an industrial environment. Subsequently, a variety of applications already carried out in industry, respectively including the source, are listed. For a better overview the results of the state-of-the-art analysis were divided into different fields of machine learning applications.

Chapter 4 comprises the practical part of the project carried out according to the cross-industry standard process for data mining (CRISP-DM). In this part, findings of the theoretical part were used and applied to a real-life use case to test their applicability. Instruments and findings from chapter 3, such as the overview over open-source machine learning frameworks in Table 1 and the model to assess data quality in Figure 3-8 were used to define basic factors like programming framework and use case. The CRISP-DM process as described in chapter 2.4 was used as a red thread going through all typical steps of a machine learning project, namely business understanding, data understanding, data preparation, modeling and evaluation. An unsupervised outlier detection model has been implemented, using welding data provided by the company to detect irregular data points. Two different approaches have been compared. One model using the “local outlier factor” algorithm and the other on a “one-class SVM” algorithm. The project has been carried out using the framework Scikit-learn. At the end a solid evaluation of the project is given in terms of performance of the models and to what extent the business objectives set in 4.4.1 have been met. Chapter 4 concludes with a review of the project as well as a recommendation for the company of further steps in the field of machine learning.

In chapter 1.1 the research issues for this work were defined. The first two questions about opportunities and challenges that arise with the use of machine learning have been answered in chapter 3.1 and chapter 3.2. Key challenges for manufacturing industry in general, challenges for the implementation of machine learning and challenges that arise from data were identified, depicted and summarized in Figure 3-1. The findings are manifold and range from data generation and accessibility of data to cyber security and the mistrust of personnel towards digitalization. Subsequently, opportunities that arise from the use of machine learning are listed, such as the possibility of handling high-dimensional data or the possibility of handling a variety of different data formats. With the findings of the state-of-the-art research in chapter 3.5 a variety of opportunities that arise from the use of machine learning we identified in fields like research and development, customer service or production. The third research issue about concepts and methods that show particular suitability for an industrial environment was covered in several chapters. In chapter 2.4 the cross-industry standard process for data mining (CRISP-DM) was introduced. This process was found suitable for future machine learning projects and applied for the execution of the practical part. Besides a model was presented to assess data quality in an industrial environment. The template developed in Figure 3-8 is hoped to serve as an additional tool for meetings in the run-up to future machine learning projects.

In chapter 1.2. the objective of the work has been defined as providing the company an experience value in the field of machine learning in manufacturing by giving an overview about the fundamentals and state of the art as well as by identifying useful instruments and concepts that can be used in future projects. Furthermore, the objective has been divided into the following subtasks:

- Identifying the current state-of-the-art of machine learning in manufacturing.
- Providing an application-oriented overview of the available algorithms and methods.
- Providing a roadmap or overview on how to proceed when implementing a machine learning project in manufacturing (problem description, data acquisition, data analysis and pre-processing, etc.)
- Selection of a real-life use case in the production of the company
- Prototypical implementation with Python and the selection of Python based framework (e.g. Scikit-learn, etc.)
- Analysis and evaluation of the results.

It was possible to meet all requested objectives in the course of this work. In the theoretical part the outline of machine learning fundamentals will give useful insights into the field of machine learning in manufacturing and the state-of-the-art analysis can serve as a pool of ideas for future applications in the production of the company. Concepts such as the model to assess data in an industrial environment can be used not only in machine learning projects but also in other applications. The CRISP-DM process as described in chapter 2.4 and executed in the practical part can assure an efficient implementation of machine learning projects and is therefore hoped to proof especially valuable for the company. By implementing a prototypical use-case in the production of the company, the concepts found in the theoretical were able to be proofed on their applicability in a real-life environment. In conclusion the project was evaluated according to criteria set at the start, the findings were discussed and a recommendation for further steps has been given.

5.2 Outlook

A recommendation for the next steps to take to further improve the company's expertise in the field of machine learning has been given in the evaluation of the practical part in chapter 4.8.3.

In the course of this work some interesting points that could be covered came up which would be worth dedicating more research efforts. Quality data coming from the QS Box could be submitted to an unsupervised learning examination. It is not improbable that previously unknown patterns can be found in the data that could be used to assess sources of errors. Furthermore, the now established connection between welding data and quality data can be used to relaunch the attempt of examining the correlation between welding data and quality data. Also, the use-cases proposed in 4.3.2 should still be born in mind for future machine learning projects.

This work is hoped to be a valuable source of information for the company on their way to industry 4.0. New concepts and insights were found that will consolidate the company's experience on the field of machine learning. In the course of the implementation of the use-case interesting discussions led to new ideas were machine learning could be applied and the state-of-the-art analysis can serve as a pool of ideas for new projects and innovations.

6 List of Reference

[1] S. W. T. W. Klaus-Dieter Thoben, „“Industrie 4.0” and Smart Manufacturing – A Review of Research Issues and Application Examples,“ 2016.

[2] D. W. C. I. K.-D. T. Thorsten Wuest, „Machine learning in manufacturing: advantages, challenges and applications,“ 2016.

[3] T. M. Mitchell, „Machine Learning,“ 1997.

[4] A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, 1959.

[5] E. Alppaydin, „Introduction to machine learning third edition,“ 2014.

[6] S. G. Andreas C. Müller, Introduction to Machine Learning with Python, 2016.

[7] G. Loupes, Understanding Random Forests, University of Liège, 2014.

[8] M. Ivan, „Types of machine learning algorithms,“ [Online]. Available: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/. [Zugriff am 27 March 2019].

[9] V. J. H. &. J. AUSTIN, „A Survey of Outlier Detection Methodologies,“ 2004.

[10] J. Brownlee, „Machine Learning Mastery,“ 2013. [Online]. Available: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

[11] R. S. S. a. A. G. Barto, Reinforcement Learning an Introduction, The MIT Press, 2017.

[12] M. F. S. Ana Azevedo, „KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW,“ 2008.

[13] IBM Corporation, „IBM SPSS Modeler CRISP-DM Guide,“ 2011.

[14] K. Jensen, 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png. [Zugriff am 01 April 2019].

[15] C. Shearer, „The CRISP-DM Model: The New Blueprint for Data Mining,“ *Journal of Data Warehousing,* 2000.

[16] R. W. a. J. Hipp, „CRISP-DM: Towards a Standard Process Model for Data,“ 2000.

[17] VDMA, „Quick Guide Machine Learning im Maschinen- und Anlagenbau,“ Frankfurt am Main, 2018.

[18] D. Evans, „The Internet of Things,“ 2011.

[19] D. L. e. al., Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning.

[20] L. Z. Keliang Zhou & Taigang Liu, „Industry 4.0: Towards Future Industrial Opportunities and Challenges,“ 2015.

[21] D. Bennett, „Future challenges for manufacturing,“ *emerald insight,* 2014.

[22] T. W. E. W. Juergen Lenz, „Holistic approach to machine tool data analytics,“ 2018.

[23] A. R. W. Z. F. F. P. L. X. F. J. T. Dazhong Wu, „Cybersecurity for digital manufacturing,“ 2017.

[24] P. L. a. R. G. &. S. Srinivasan, „Manufacturing Analytics and Industrial Internet of Things,“ 2017.

[25] T. Wuest, Identifying Product and Process State Drivers in Manufacturing Systems Using Supervised Machine Learning, 2015.

[26] G. B. Team, „TensorFlow: A system for large-scale machine learning,“ 2016.

[27] T. D. Team, „Theano: A Python framework for fast computation of mathematical expressions,“ 2016.

[28] [Online]. Available: https://en.wikipedia.org/wiki/Keras. [Zugriff am 20 05 2019].

[29] E. F. M. A. H. C. J. P. Ian H. Witten, Data Mining - Practical Machine Learning Tools and Techniques, Fourth Edition, 2017.

[30] Wikipedia, „Wikipedia,“ 12 04 2019. [Online]. Available: https://en.wikipedia.org/wiki/Bayes%27_theorem.

[31] S. B. Kotsiantis, „Supervised Machine Learning: A Review of Classification Techniques,“ *Informatica,* 2007.

[32] S. S. &. M. Giri, „Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey,“ 2014.

[33] Wikipedia, „Wikipedia,“ [Online]. Available: https://en.wikipedia.org/wiki/Instance-based_learning. [Zugriff am 15 04 2019].

[34] F. Gagliardi, „Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction,“ 2011.

[35] J. P. &. A. Gibson, Deep Learning, O'Reilly, 2017.

[36] R. M. David Opitz, „Popular Ensemble Methods: An Empirical Study,“ 1999.

[37] P. Gupta, „TowardsDataScience," 2017. [Online]. Available: https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a. [Zugriff am 16 04 2019].

[38] N. Guaro. [Online]. Available: https://en.wikipedia.org/wiki/Regularization_(mathematics). [Zugriff am 21 05 2019].

[39] Scikit-learn, „scikit-learn," [Online]. Available: https://scikit-learn.org/stable/modules/svm.html. [Zugriff am 16 04 2019].

[40] M. G. H. H. Knut Hildebrand, Daten- und Informationsqualität, 2008.

[41] M. W. J. D. R. B. Michel Eickelmann, „Bewertungsmodell zur Analyse der Datenreife," *ZWF,* 2019.

[42] D. E. M. e. al., „Machine Learning 2030," VDMA, 2016.

[43] C. P. Z. D. W. a. Z. Z. K. Knud Lasse Lueth, „INDUSTRIAL ANALYTICS 2016/2017 - The current state of data analytics usage in industrial companies," 2016.

[44] R. A. T. H. J. Michael Sharp, „A survey of the advancing use and development of machine learning in smart manufacturing," 2018.

[45] A. K. C. ·. J. A. H. ·. M. K. Tiwari, „Data mining in manufacturing: a review based on the kind of knowledge," 2008.

[46] D. U. Waltinger, „Artificial Intelligence & Deep Learning: Unlock the Potential with AI," 2018.

[47] A. B. e. al., „Cyber Physical Systems for Life Cycle Continuous Technical Documentation of Manufacturing Facilities," 2014.

[48] A. S. e. al., „Exploring data-driven innovations for manufacturing in a lightweight living," 2017.

[49] T.-H. Y. S. Denga, „Using least squares support vector machines for the airframe structuresmanufacturing cost estimation," 2011.

[50] G. A. S. e. al., „Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," 2015.

[51] H. M. a. Y. Teranishi, „Development of Automated Data Mining System for Quality Control in Manufacturing," 2001.

[52] X. G. a. S. K. Deyong You, „WPD-PCA-Based Laser Welding ProcessMonitoring and Defects Diagnosis byUsing FNN and SVM," 2015.

[53] M. U. C. W. Carsten Steger, „Machine Vision Algorithms and Applications," 2018.

[54] K. Walmsley, „keanw," [Online]. Available: https://www.keanw.com/2018/08/brickbot-an-autodesk-research-project-exploring-the-future-of-manufacturing.html. [Zugriff am 13 05 2019].

[55] S. U. Daniel Schneegaß, „Method for computer-aided control and/or regulation using two neural networks wherein the second neural network models a quality function and can be used to control a gas turbine". 2013.

[56] „Siemens," [Online]. Available: https://new.siemens.com/global/de/produkte/mobilitaet/schienenverkehr/services/spare-part-services/easy-spares-idea.html. [Zugriff am 2019 05 2019].

[57] J. W. e. al., „Deep learning for smart manufacturing: Methods and applications," 2018.

[58] T. H. a. K. C. T. Ruoxu Ren, „A Generic Deep-Learning-Based Approach for Automated Surface Inspection," 2018.

[59] D. Schall, „Maschinen, die wir nicht verstehen: Was kann KI heute wirklich?," derStandard, Wien, 2019.

[60] J. C. (. R. K. (. T. K. (. T. R. (. C. S. (. a. R. W. (. Pete Chapman (NCR), „CRISP-DM 1.0," 2000.

[61] P. J. Benner, „Cloos Prozessdaten Monitoring (PDM) - Bedienungsanleitung," 2014.

[62] F. P. e. al., „Scikit-learn: Machine Learning in Python," 2011.

[63] T. H. J. A. B. F. Moneer Helu, „Reference architecture to integrate heterogeneous manufacturing systems for the digital thread," 2017.

[64] S. WOLD, „Principal Component Analysis," 1997.

[65] H.-P. K. J. S. X. X. Martin Ester, „A Density-Based Algorithmfor Discovering Clusters," 1996.

[66] H.-P. K. e. al., „Density-based clustering," 2011.

[67] H.-P. K. R. T. N. J. S. Markus M. Breunig, „LOF: Identifying Density-Based Local Outliers," 2000.

[68] L. E. V. K. A. O. J. S. Aleksandar Lazarevic, „A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," 2003.

[69] R. Sacher, „Verbesserung der Ladeprognose von batteriebetriebenen Fahrzeugen durch maschinelles Lernen / Künstliche Intelligenz," 2019.

[70] s.-l. developers, „scikit-learn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html. [Zugriff am 2019 07 18].

[71] s.-l. developers, „Scikit Learn,“ [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html. [Zugriff am 24 07 2019].

[72] s.-l. developers, „Scikit-learn,“ [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler.fit_transform. [Zugriff am 18 07 2019].

[73] J. F. H. e. a., „Multivariate Data Analysis,“ Pearson, 2012.

[74] M. G. B. e. al., „Principal component analysis in sensory analysis: covariance or correlation matrix?,“ 2000.

[75] W. J. Krzanowski, „Statistical Principles and Techniques in Scientific and Social Investigations,“ 2007.

[76] „Wikipedia,“ [Online]. Available: https://en.wikipedia.org/wiki/Outline_of_machine_learning#Machine_learning_methods.

[77] „Towards data science,“ 2017. [Online]. Available: https://towardsdatascience.com/machine-learning-types-2-c1291d4f04b1. [Zugriff am 27 March 2019].

[78] S. Raschka, Python Machine Learning, Packed Publishing, 2015.

[79] T. G. Dietterich, „Ensemble Learning,“ *MITPress,* 2002.

[80] Wikipedia, „Wikipedia,“ [Online]. Available: https://en.wikipedia.org/wiki/Outline_of_machine_learning. [Zugriff am 17 04 2019].

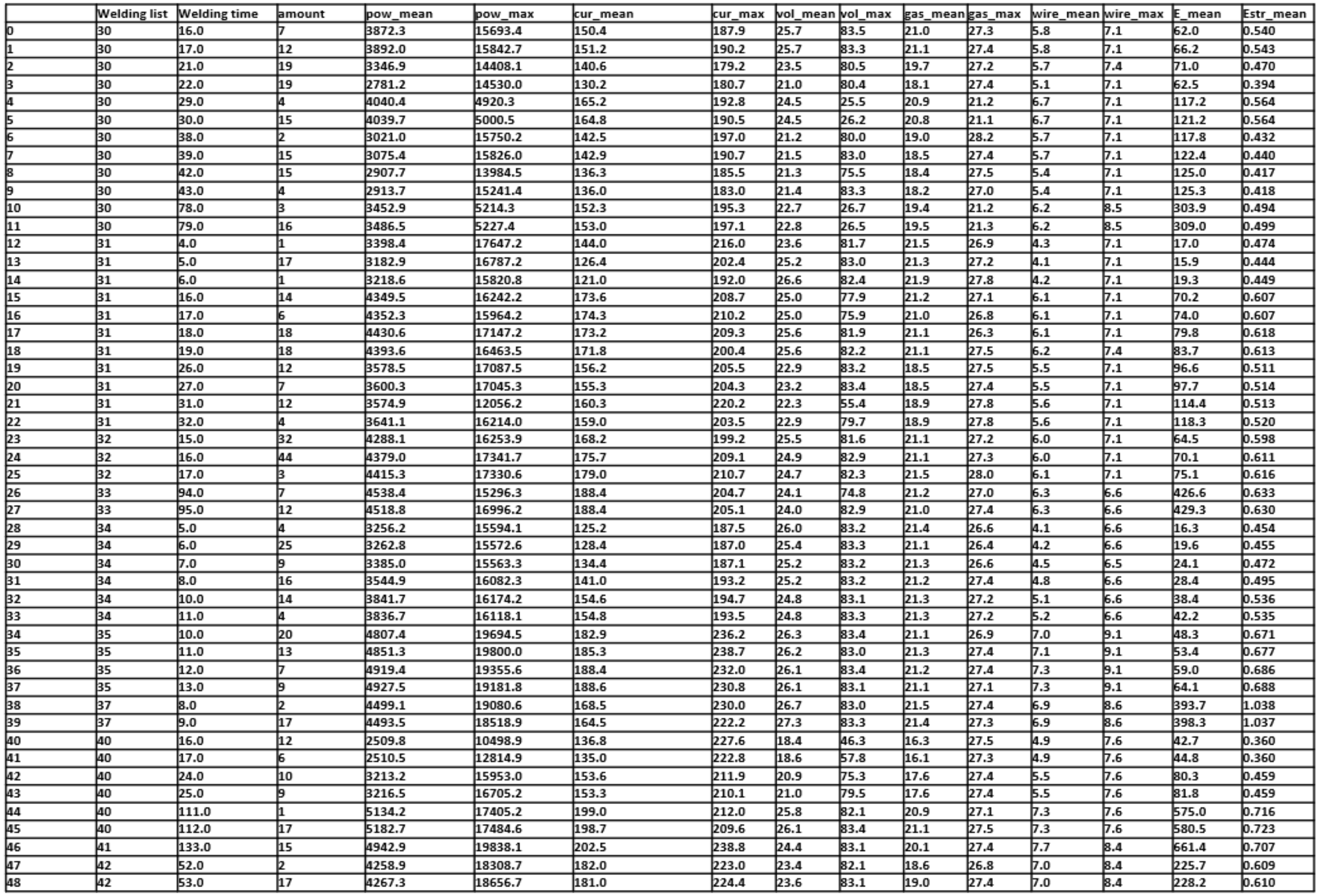
[81] TensorFLow, [Online]. Available: https://www.tensorflow.org/. [Zugriff am 20 05 2019].

7 List of Figures

8 List of Tables

9 Annex

Figure 9-1: Average values of different welding lists with different welding times (1/2)

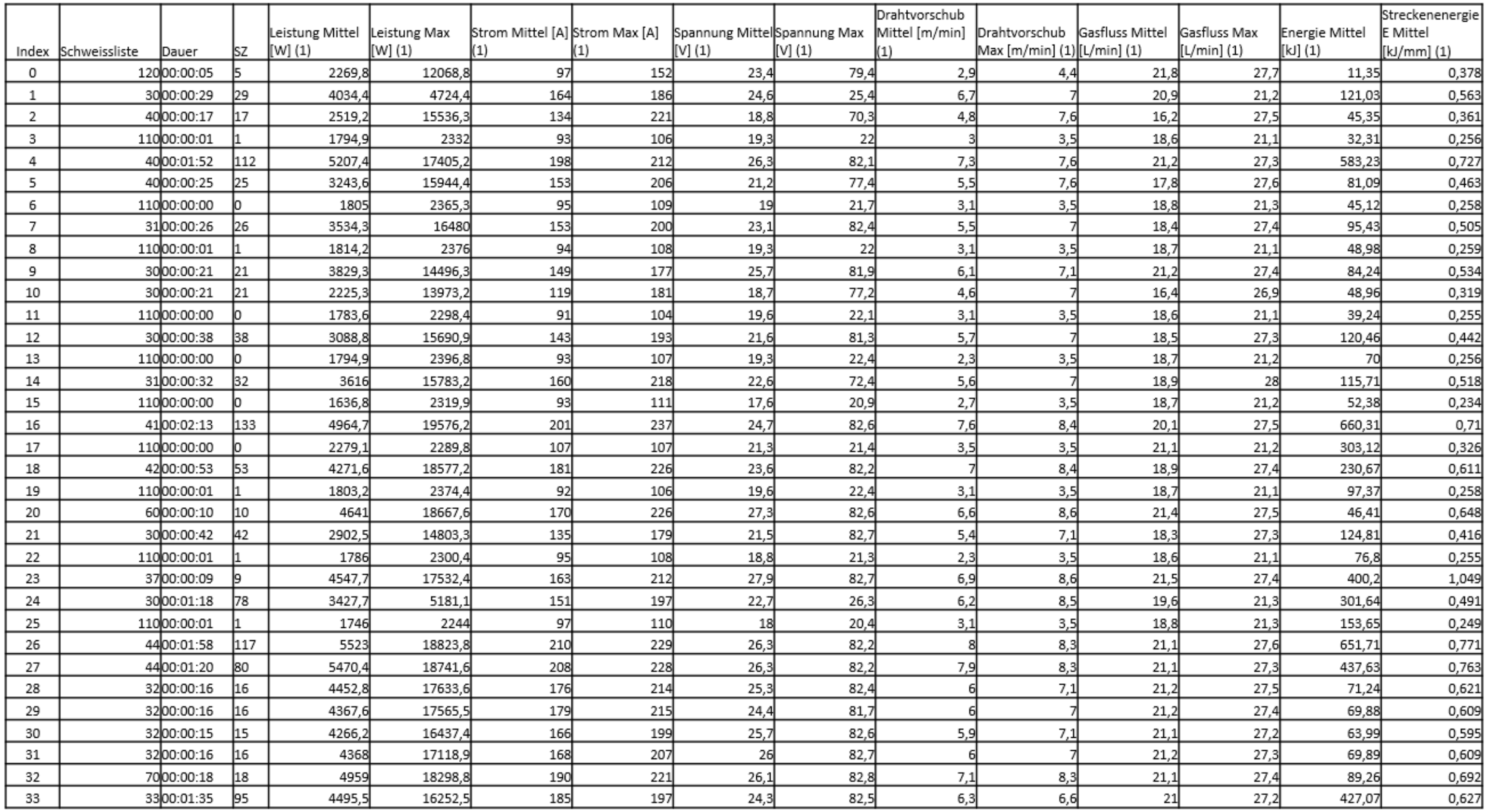


	Welding list	Welding time	amount	pow_mean	pow_max	cur_mean	cur_max	vol_mean	vol_max	gas_mean	gas_max	wire_mean	wire_max	E_mean	Estr_mean
0	30	16.0	7	3872.3	15693.4	150.4	187.9	25.7	83.5	21.0	27.3	5.8	7.1	62.0	0.540
1	30	17.0	12	3892.0	15842.7	151.2	190.2	25.7	83.3	21.1	27.4	5.8	7.1	66.2	0.543
2	30	21.0	19	3346.9	14408.1	140.6	179.2	23.5	80.5	19.7	27.2	5.7	7.4	71.0	0.470
3	30	22.0	19	2781.2	14530.0	130.2	180.7	21.0	80.4	18.1	27.4	5.1	7.1	62.5	0.394
4	30	29.0	4	4040.4	4920.3	165.2	192.8	24.5	25.5	20.9	21.2	6.7	7.1	117.2	0.564
5	30	30.0	15	4039.7	5000.5	164.8	190.5	24.5	26.2	20.8	21.1	6.7	7.1	121.2	0.564
6	30	38.0	2	3021.0	15750.2	142.5	197.0	21.2	80.0	19.0	28.2	5.7	7.1	117.8	0.432
7	30	39.0	15	3075.4	15826.0	142.9	190.7	21.5	83.0	18.5	27.4	5.7	7.1	122.4	0.440
8	30	42.0	15	2907.7	13984.5	136.3	185.5	21.3	75.5	18.4	27.5	5.4	7.1	125.0	0.417
9	30	43.0	4	2913.7	15241.4	136.0	183.0	21.4	83.3	18.2	27.0	5.4	7.1	125.3	0.418
10	30	78.0	3	3452.9	5214.3	152.3	195.3	22.7	26.7	19.4	21.2	6.2	8.5	303.9	0.494
11	30	79.0	16	3486.5	5227.4	153.0	197.1	22.8	26.5	19.5	21.3	6.2	8.5	309.0	0.499
12	31	4.0	1	3398.4	17647.2	144.0	216.0	23.6	81.7	21.5	26.9	4.3	7.1	17.0	0.474
13	31	5.0	17	3182.9	16787.2	126.4	202.4	25.2	83.0	21.3	27.2	4.1	7.1	15.9	0.444
14	31	6.0	1	3218.6	15820.8	121.0	192.0	26.6	82.4	21.9	27.8	4.2	7.1	19.3	0.449
15	31	16.0	14	4349.5	16242.2	173.6	208.7	25.0	77.9	21.2	27.1	6.1	7.1	70.2	0.607
16	31	17.0	6	4352.3	15964.2	174.3	210.2	25.0	75.9	21.0	26.8	6.1	7.1	74.0	0.607
17	31	18.0	18	4430.6	17147.2	173.2	209.3	25.6	81.9	21.1	26.3	6.1	7.1	79.8	0.618
18	31	19.0	18	4393.6	16463.5	171.8	200.4	25.6	82.2	21.1	27.5	6.2	7.4	83.7	0.613
19	31	26.0	12	3578.5	17087.5	156.2	205.5	22.9	83.2	18.5	27.5	5.5	7.1	96.6	0.511
20	31	27.0	7	3600.3	17045.3	155.3	204.3	23.2	83.4	18.5	27.4	5.5	7.1	97.7	0.514
21	31	31.0	12	3574.9	12056.2	160.3	220.2	22.3	55.4	18.9	27.8	5.6	7.1	114.4	0.513
22	31	32.0	4	3641.1	16214.0	159.0	203.5	22.9	79.7	18.9	27.8	5.6	7.1	118.3	0.520
23	32	15.0	32	4288.1	16253.9	168.2	199.2	25.5	81.6	21.1	27.2	6.0	7.1	64.5	0.598
24	32	16.0	44	4379.0	17341.7	175.7	209.1	24.9	82.9	21.1	27.3	6.0	7.1	70.1	0.611
25	32	17.0	3	4415.3	17330.6	179.0	210.7	24.7	82.3	21.5	28.0	6.1	7.1	75.1	0.616
26	33	94.0	7	4538.4	15296.3	188.4	204.7	24.1	74.8	21.2	27.0	6.3	6.6	426.6	0.633
27	33	95.0	12	4518.8	16996.2	188.4	205.1	24.0	82.9	21.0	27.4	6.3	6.6	429.3	0.630
28	34	5.0	4	3256.2	15594.1	125.2	187.5	26.0	83.2	21.4	26.6	4.1	6.6	16.3	0.454
29	34	6.0	25	3262.8	15572.6	128.4	187.0	25.4	83.3	21.1	26.4	4.2	6.6	19.6	0.455
30	34	7.0	9	3385.0	15563.3	134.4	187.1	25.2	83.2	21.3	26.6	4.5	6.5	24.1	0.472
31	34	8.0	16	3544.9	16082.3	141.0	193.2	25.2	83.2	21.2	27.4	4.8	6.6	28.4	0.495
32	34	10.0	14	3841.7	16174.2	154.6	194.7	24.8	83.1	21.3	27.2	5.1	6.6	38.4	0.536
33	34	11.0	4	3836.7	16118.1	154.8	193.5	24.8	83.3	21.3	27.2	5.2	6.6	42.2	0.535
34	35	10.0	20	4807.4	19694.5	182.9	236.2	26.3	83.4	21.1	26.9	7.0	9.1	48.3	0.671
35	35	11.0	13	4851.3	19800.0	185.3	238.7	26.2	83.0	21.3	27.4	7.1	9.1	53.4	0.677
36	35	12.0	7	4919.4	19355.6	188.4	232.0	26.1	83.4	21.2	27.4	7.3	9.1	59.0	0.686
37	35	13.0	9	4927.5	19181.8	188.6	230.8	26.1	83.1	21.1	27.1	7.3	9.1	64.1	0.688
38	37	8.0	2	4499.1	19080.6	168.5	230.0	26.7	83.0	21.5	27.4	6.9	8.6	393.7	1.038
39	37	9.0	17	4493.5	18518.9	164.5	222.2	27.3	83.3	21.4	27.3	6.9	8.6	398.3	1.037
40	40	16.0	12	2509.8	10498.9	136.8	227.6	18.4	46.3	16.3	27.5	4.9	7.6	42.7	0.360
41	40	17.0	6	2510.5	12814.9	135.0	222.8	18.6	57.8	16.1	27.3	4.9	7.6	44.8	0.360
42	40	24.0	10	3213.2	15953.0	153.6	211.9	20.9	75.3	17.6	27.4	5.5	7.6	80.3	0.459
43	40	25.0	9	3216.5	16705.2	153.3	210.1	21.0	79.5	17.6	27.4	5.5	7.6	81.8	0.459
44	40	111.0	1	5134.2	17405.2	199.0	212.0	25.8	82.1	20.9	27.1	7.3	7.6	575.0	0.716
45	40	112.0	17	5182.7	17484.6	198.7	209.6	26.1	83.4	21.1	27.5	7.3	7.6	580.5	0.723
46	41	133.0	15	4942.9	19838.1	202.5	238.8	24.4	83.1	20.1	27.4	7.7	8.4	661.4	0.707
47	42	52.0	2	4258.9	18308.7	182.0	223.0	23.4	82.1	18.6	26.8	7.0	8.4	225.7	0.609
48	42	53.0	17	4267.3	18656.7	181.0	224.4	23.6	83.1	19.0	27.4	7.0	8.4	228.2	0.610

Figure 9-2: Average values of different welding lists with different welding times (2/2)

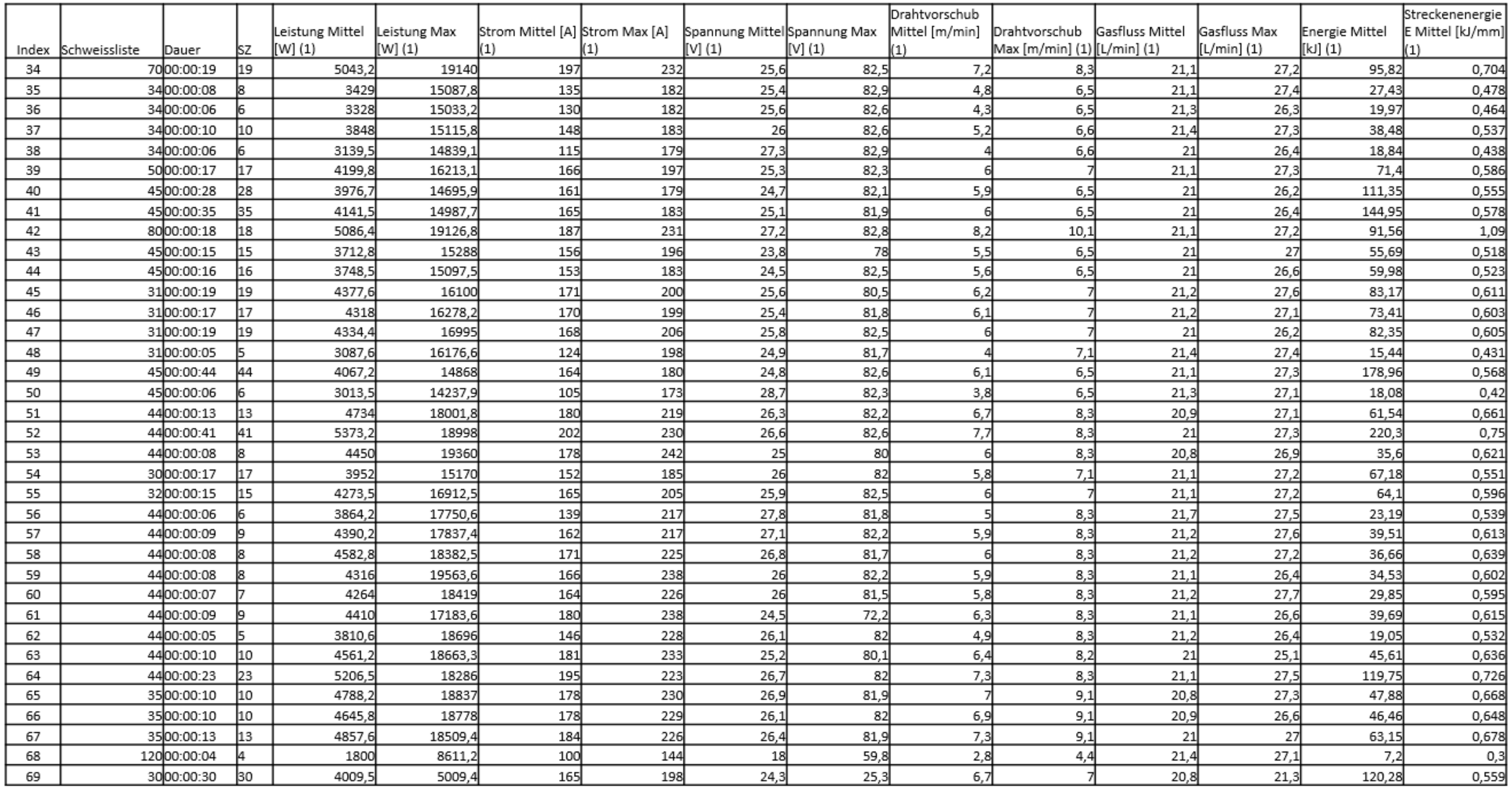
	Welding list	Welding time	amount	pow_mean	pow_max	cur_mean	cur_max	vol_mean	vol_max	gas_mean	gas_max	wire_mean	wire_max	E_mean	Estr_mean
49	43	20.0	9	5198.2	19667.7	194.8	236.2	26.7	83.3	21.2	27.4	7.6	8.6	104.0	0.725
50	43	21.0	6	5158.5	19276.4	193.3	231.8	26.7	83.1	21.2	27.4	7.5	8.6	108.3	0.720
51	44	5.0	19	3800.6	18882.1	145.6	226.6	26.1	83.3	21.3	26.8	5.0	8.3	19.0	0.530
52	44	6.0	16	3830.8	18958.5	146.4	228.0	26.2	83.1	21.3	26.8	5.0	8.3	23.0	0.535
53	44	7.0	6	4195.9	16942.8	168.2	234.7	25.0	72.4	21.2	27.6	6.0	8.3	29.4	0.586
54	44	8.0	66	4453.3	19348.4	173.1	234.1	25.7	82.7	21.3	27.2	6.1	8.3	35.6	0.621
55	44	9.0	36	4460.3	19095.9	174.6	233.7	25.6	81.7	21.2	26.9	6.2	8.3	40.1	0.622
56	44	10.0	17	4640.3	19187.5	179.5	231.0	25.9	83.1	21.0	24.8	6.4	8.3	46.4	0.647
57	44	12.0	7	4837.3	19427.9	186.6	235.0	25.9	82.7	21.2	27.5	6.9	8.3	58.7	0.675
58	44	13.0	12	4885.4	19335.0	188.1	235.8	26.0	82.0	21.2	27.3	6.8	8.3	63.5	0.682
59	44	22.0	2	5148.3	18289.6	196.5	222.5	26.2	82.2	21.0	27.1	7.4	8.3	113.3	0.718
60	44	23.0	14	5213.9	18853.8	197.9	225.9	26.3	83.4	21.2	27.4	7.4	8.3	119.9	0.727
61	44	41.0	14	5364.2	18993.1	205.4	239.0	26.1	79.6	21.1	27.5	7.8	8.3	219.9	0.748
62	44	42.0	5	5369.2	19445.0	205.4	236.0	26.1	82.4	21.0	27.3	7.8	8.3	225.5	0.749
63	44	79.0	1	5496.7	18500.5	209.0	227.0	26.3	81.5	20.8	27.0	8.0	8.3	439.7	0.767
64	44	80.0	17	5504.4	19062.2	209.3	230.6	26.3	82.6	21.2	27.5	8.0	8.3	440.4	0.768
65	44	117.0	18	5557.5	19220.5	210.8	231.2	26.4	83.1	21.1	27.8	8.0	8.3	650.2	0.775
66	45	5.0	10	2930.6	15030.2	110.2	180.4	26.6	83.3	21.2	27.3	3.8	7.1	14.7	0.409
67	45	6.0	9	2946.4	14812.5	107.8	177.3	27.4	83.5	21.2	27.2	3.8	6.6	17.7	0.411
68	45	14.0	1	3760.4	15132.0	158.0	194.0	23.8	78.0	21.3	27.3	5.6	6.6	52.6	0.525
69	45	15.0	17	3784.9	15934.8	158.5	197.6	23.9	80.7	21.2	27.1	5.6	6.6	56.8	0.528
70	45	16.0	19	3800.9	15873.8	156.2	193.2	24.3	82.2	21.1	26.8	5.6	6.6	60.8	0.530
71	45	27.0	14	4003.8	15733.8	161.2	188.8	24.8	83.3	21.0	26.3	5.9	6.6	108.4	0.559
72	45	28.0	5	3999.4	15860.4	161.4	191.0	24.8	83.0	21.3	26.6	5.9	6.6	112.0	0.558
73	45	34.0	7	4077.1	15312.8	167.0	185.1	24.4	82.7	20.9	26.7	6.1	6.6	138.6	0.569
74	45	35.0	12	4079.6	15421.8	166.8	186.0	24.5	82.9	21.2	26.9	6.0	6.6	142.8	0.569
75	45	43.0	15	4062.1	15343.0	165.4	184.1	24.6	83.3	21.1	27.4	6.1	6.6	174.7	0.567
76	45	44.0	4	4054.9	15261.8	164.0	183.8	24.7	83.0	21.0	27.2	6.1	6.6	178.4	0.566
77	50	16.0	10	4180.6	16660.2	167.5	199.9	25.0	83.3	21.1	27.4	6.0	7.1	66.9	0.583
78	50	17.0	9	4212.8	16526.9	167.8	198.6	25.1	83.2	21.2	27.6	6.0	7.1	71.6	0.588
79	60	9.0	1	4655.0	18791.0	175.0	230.0	26.6	81.7	22.0	27.9	6.8	8.6	46.6	0.650
80	60	10.0	16	4555.1	18720.4	171.1	224.6	26.6	83.4	21.3	27.5	6.6	8.6	45.6	0.636
81	60	11.0	2	4486.5	18870.2	169.0	226.0	26.6	83.5	20.7	26.9	6.6	8.6	49.3	0.626
82	60	38.0	1	5286.3	18709.7	201.0	223.0	26.3	83.9	21.1	27.2	8.1	8.6	211.4	0.793
83	60	39.0	18	5331.3	18308.6	202.5	222.3	26.3	82.3	21.1	27.2	8.1	8.7	216.8	0.800
84	70	17.0	1	5035.0	18586.1	190.0	221.0	26.5	84.1	21.5	27.9	7.1	8.3	85.6	0.703
85	70	18.0	23	5010.6	18336.0	193.3	229.7	25.9	80.2	21.2	27.4	7.1	8.3	90.2	0.699
86	70	19.0	10	5010.4	19025.4	196.9	232.0	25.4	82.0	21.0	27.4	7.2	8.3	95.2	0.699
87	70	20.0	1	5356.8	19327.6	186.0	229.0	28.8	84.4	20.9	27.1	6.8	8.3	107.1	0.747
88	80	16.0	2	5530.7	20165.4	197.5	241.5	28.0	83.5	20.9	27.2	8.6	10.1	99.6	1.186
89	80	17.0	4	5256.9	20007.6	192.0	238.8	27.4	83.8	21.2	27.5	8.4	10.1	93.3	1.126
90	80	18.0	12	5118.8	19906.4	190.8	239.2	26.8	83.2	21.2	27.4	8.3	10.2	92.1	1.097
91	110	0.0	62	1952.3	2331.4	98.6	108.2	19.7	21.6	19.6	21.1	2.9	3.5	85.1	0.279
92	110	1.0	118	1909.8	2320.3	97.8	108.2	19.5	21.4	19.4	21.2	2.9	3.5	93.2	0.273
93	120	4.0	12	2242.2	11397.4	98.8	143.8	22.7	79.2	21.4	27.2	2.9	4.9	9.2	0.374
94	120	5.0	7	2373.2	11843.5	96.6	145.9	24.7	81.2	21.5	27.4	2.8	4.4	11.9	0.396
95	160	1.0	8	5067.2	5851.2	202.8	219.9	25.0	26.6	20.1	21.3	7.9	8.6	203.4	0.949
96	160	2.0	11	5249.1	5813.1	206.7	219.0	25.4	26.5	20.4	21.2	8.1	8.6	215.2	0.984

Figure 9-3: Test set data. Including 136 instances (1/4)



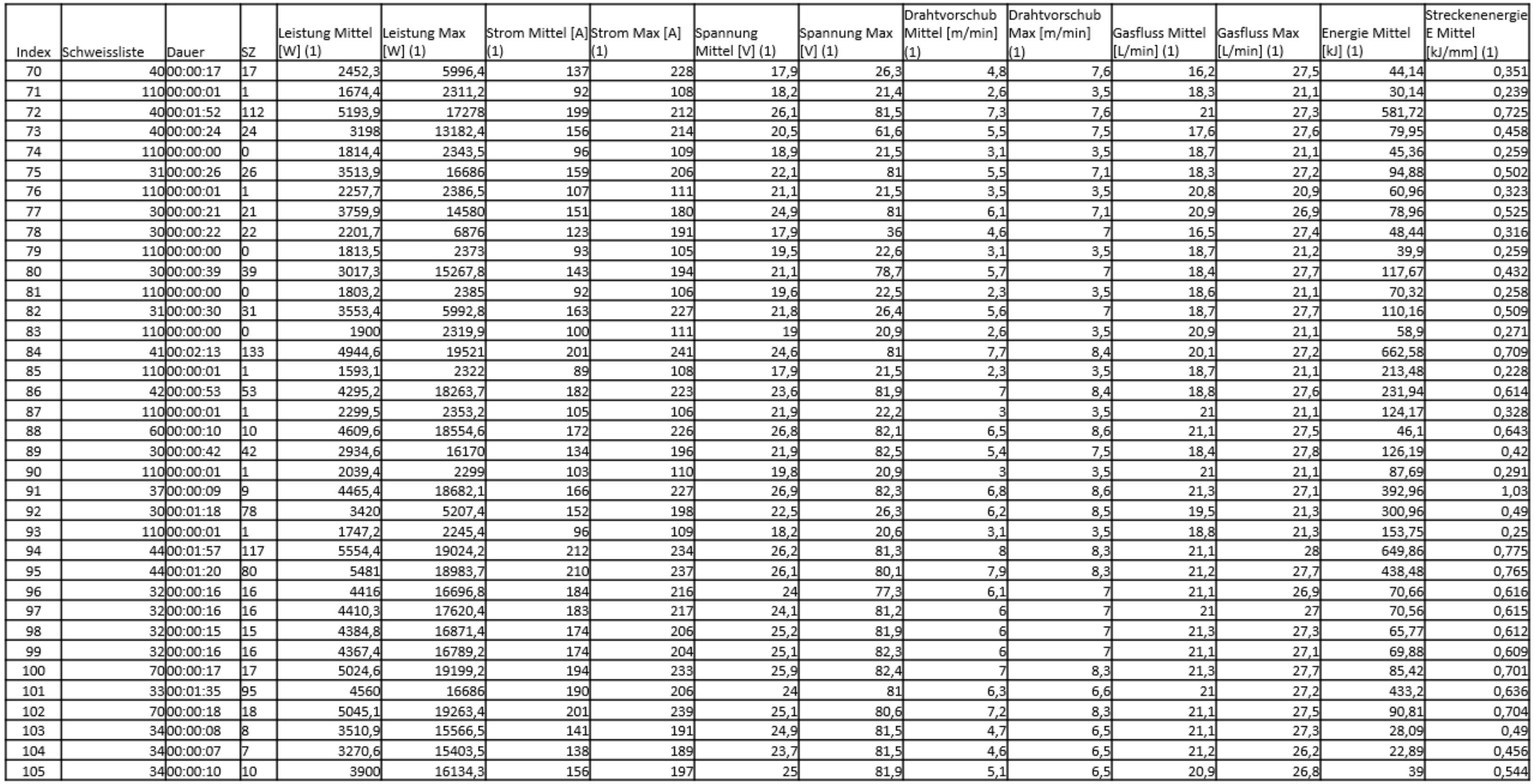
Index	Schweissliste	Dauer	SZ	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie E Mittel [kJ/mm] (1)
0	120	00:00:05	5	2269,8	12068,8	97	152	23,4	79,4	2,9	4,4	21,8	27,7	11,35	0,378
1	30	00:00:29	29	4034,4	4724,4	164	186	24,6	25,4	6,7	7	20,9	21,2	121,03	0,563
2	40	00:00:17	17	2519,2	15536,3	134	221	18,8	70,3	4,8	7,6	16,2	27,5	45,35	0,361
3	110	00:00:01	1	1794,9	2332	93	106	19,3	22	3	3,5	18,6	21,1	32,31	0,256
4	40	00:01:52	112	5207,4	17405,2	198	212	26,3	82,1	7,3	7,6	21,2	27,3	583,23	0,727
5	40	00:00:25	25	3243,6	15944,4	153	206	21,2	77,4	5,5	7,6	17,8	27,6	81,09	0,463
6	110	00:00:00	0	1805	2365,3	95	109	19	21,7	3,1	3,5	18,8	21,3	45,12	0,258
7	31	00:00:26	26	3534,3	16480	153	200	23,1	82,4	5,5	7	18,4	27,4	95,43	0,505
8	110	00:00:01	1	1814,2	2376	94	108	19,3	22	3,1	3,5	18,7	21,1	48,98	0,259
9	30	00:00:21	21	3829,3	14496,3	149	177	25,7	81,9	6,1	7,1	21,2	27,4	84,24	0,534
10	30	00:00:21	21	2225,3	13973,2	119	181	18,7	77,2	4,6	7	16,4	26,9	48,96	0,319
11	110	00:00:00	0	1783,6	2298,4	91	104	19,6	22,1	3,1	3,5	18,6	21,1	39,24	0,255
12	30	00:00:38	38	3088,8	15690,9	143	193	21,6	81,3	5,7	7	18,5	27,3	120,46	0,442
13	110	00:00:00	0	1794,9	2396,8	93	107	19,3	22,4	2,3	3,5	18,7	21,2	70	0,256
14	31	00:00:32	32	3616	15783,2	160	218	22,6	72,4	5,6	7	18,9	28	115,71	0,518
15	110	00:00:00	0	1636,8	2319,9	93	111	17,6	20,9	2,7	3,5	18,7	21,2	52,38	0,234
16	41	00:02:13	133	4964,7	19576,2	201	237	24,7	82,6	7,6	8,4	20,1	27,5	660,31	0,71
17	110	00:00:00	0	2279,1	2289,8	107	107	21,3	21,4	3,5	3,5	21,1	21,2	303,12	0,326
18	42	00:00:53	53	4271,6	18577,2	181	226	23,6	82,2	7	8,4	18,9	27,4	230,67	0,611
19	110	00:00:01	1	1803,2	2374,4	92	106	19,6	22,4	3,1	3,5	18,7	21,1	97,37	0,258
20	60	00:00:10	10	4641	18667,6	170	226	27,3	82,6	6,6	8,6	21,4	27,5	46,41	0,648
21	30	00:00:42	42	2902,5	14803,3	135	179	21,5	82,7	5,4	7,1	18,3	27,3	124,81	0,416
22	110	00:00:01	1	1786	2300,4	95	108	18,8	21,3	2,3	3,5	18,6	21,1	76,8	0,255
23	37	00:00:09	9	4547,7	17532,4	163	212	27,9	82,7	6,9	8,6	21,5	27,4	400,2	1,049
24	30	00:01:18	78	3427,7	5181,1	151	197	22,7	26,3	6,2	8,5	19,6	21,3	301,64	0,491
25	110	00:00:01	1	1746	2244	97	110	18	20,4	3,1	3,5	18,8	21,3	153,65	0,249
26	44	00:01:58	117	5523	18823,8	210	229	26,3	82,2	8	8,3	21,1	27,6	651,71	0,771
27	44	00:01:20	80	5470,4	18741,6	208	228	26,3	82,2	7,9	8,3	21,1	27,3	437,63	0,763
28	32	00:00:16	16	4452,8	17633,6	176	214	25,3	82,4	6	7,1	21,2	27,5	71,24	0,621
29	32	00:00:16	16	4367,6	17565,5	179	215	24,4	81,7	6	7	21,2	27,4	69,88	0,609
30	32	00:00:15	15	4266,2	16437,4	166	199	25,7	82,6	5,9	7,1	21,1	27,2	63,99	0,595
31	32	00:00:16	16	4368	17118,9	168	207	26	82,7	6	7	21,2	27,3	69,89	0,609
32	70	00:00:18	18	4959	18298,8	190	221	26,1	82,8	7,1	8,3	21,1	27,4	89,26	0,692
33	33	00:01:35	95	4495,5	16252,5	185	197	24,3	82,5	6,3	6,6	21	27,2	427,07	0,627

Figure 9-4: Test set data. Including 136 instances (2/4)



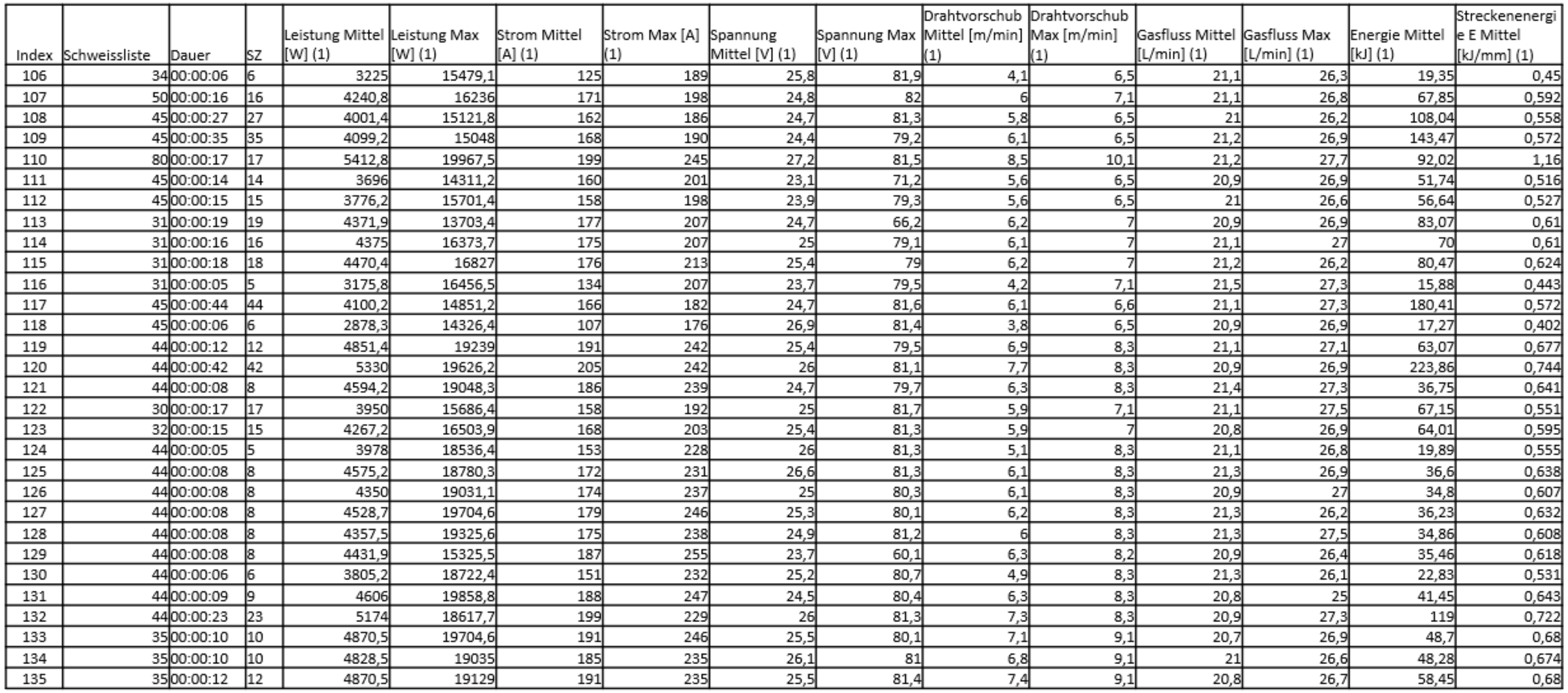
Index	Schweissliste	Dauer	SZ	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie E Mittel [kJ/mm] (1)
34	70	00:00:19	19	5043,2	19140	197	232	25,6	82,5	7,2	8,3	21,1	27,2	95,82	0,704
35	34	00:00:08	8	3429	15087,8	135	182	25,4	82,9	4,8	6,5	21,1	27,4	27,43	0,478
36	34	00:00:06	6	3328	15033,2	130	182	25,6	82,6	4,3	6,5	21,3	26,3	19,97	0,464
37	34	00:00:10	10	3848	15115,8	148	183	26	82,6	5,2	6,6	21,4	27,3	38,48	0,537
38	34	00:00:06	6	3139,5	14839,1	115	179	27,3	82,9	4	6,6	21	26,4	18,84	0,438
39	50	00:00:17	17	4199,8	16213,1	166	197	25,3	82,3	6	7	21,1	27,3	71,4	0,586
40	45	00:00:28	28	3976,7	14695,9	161	179	24,7	82,1	5,9	6,5	21	26,2	111,35	0,555
41	45	00:00:35	35	4141,5	14987,7	165	183	25,1	81,9	6	6,5	21	26,4	144,95	0,578
42	80	00:00:18	18	5086,4	19126,8	187	231	27,2	82,8	8,2	10,1	21,1	27,2	91,56	1,09
43	45	00:00:15	15	3712,8	15288	156	196	23,8	78	5,5	6,5	21	27	55,69	0,518
44	45	00:00:16	16	3748,5	15097,5	153	183	24,5	82,5	5,6	6,5	21	26,6	59,98	0,523
45	31	00:00:19	19	4377,6	16100	171	200	25,6	80,5	6,2	7	21,2	27,6	83,17	0,611
46	31	00:00:17	17	4318	16278,2	170	199	25,4	81,8	6,1	7	21,2	27,1	73,41	0,603
47	31	00:00:19	19	4334,4	16995	168	206	25,8	82,5	6	7	21	26,2	82,35	0,605
48	31	00:00:05	5	3087,6	16176,6	124	198	24,9	81,7	4	7,1	21,4	27,4	15,44	0,431
49	45	00:00:44	44	4067,2	14868	164	180	24,8	82,6	6,1	6,5	21,1	27,3	178,96	0,568
50	45	00:00:06	6	3013,5	14237,9	105	173	28,7	82,3	3,8	6,5	21,3	27,1	18,08	0,42
51	44	00:00:13	13	4734	18001,8	180	219	26,3	82,2	6,7	8,3	20,9	27,1	61,54	0,661
52	44	00:00:41	41	5373,2	18998	202	230	26,6	82,6	7,7	8,3	21	27,3	220,3	0,75
53	44	00:00:08	8	4450	19360	178	242	25	80	6	8,3	20,8	26,9	35,6	0,621
54	30	00:00:17	17	3952	15170	152	185	26	82	5,8	7,1	21,1	27,2	67,18	0,551
55	32	00:00:15	15	4273,5	16912,5	165	205	25,9	82,5	6	7	21,1	27,2	64,1	0,596
56	44	00:00:06	6	3864,2	17750,6	139	217	27,8	81,8	5	8,3	21,7	27,5	23,19	0,539
57	44	00:00:09	9	4390,2	17837,4	162	217	27,1	82,2	5,9	8,3	21,2	27,6	39,51	0,613
58	44	00:00:08	8	4582,8	18382,5	171	225	26,8	81,7	6	8,3	21,2	27,2	36,66	0,639
59	44	00:00:08	8	4316	19563,6	166	238	26	82,2	5,9	8,3	21,1	26,4	34,53	0,602
60	44	00:00:07	7	4264	18419	164	226	26	81,5	5,8	8,3	21,2	27,7	29,85	0,595
61	44	00:00:09	9	4410	17183,6	180	238	24,5	72,2	6,3	8,3	21,1	26,6	39,69	0,615
62	44	00:00:05	5	3810,6	18696	146	228	26,1	82	4,9	8,3	21,2	26,4	19,05	0,532
63	44	00:00:10	10	4561,2	18663,3	181	233	25,2	80,1	6,4	8,2	21	25,1	45,61	0,636
64	44	00:00:23	23	5206,5	18286	195	223	26,7	82	7,3	8,3	21,1	27,5	119,75	0,726
65	35	00:00:10	10	4788,2	18837	178	230	26,9	81,9	7	9,1	20,8	27,3	47,88	0,668
66	35	00:00:10	10	4645,8	18778	178	229	26,1	82	6,9	9,1	20,9	26,6	46,46	0,648
67	35	00:00:13	13	4857,6	18509,4	184	226	26,4	81,9	7,3	9,1	21	27	63,15	0,678
68	120	00:00:04	4	1800	8611,2	100	144	18	59,8	2,8	4,4	21,4	27,1	7,2	0,3
69	30	00:00:30	30	4009,5	5009,4	165	198	24,3	25,3	6,7	7	20,8	21,3	120,28	0,559

Figure 9-5: Test set data. Including 136 instances (3/4)



Index	Schweissliste	Dauer	SZ	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie E Mittel [kJ/mm] (1)
70	40	00:00:17	17	2452,3	5996,4	137	228	17,9	26,3	4,8	7,6	16,2	27,5	44,14	0,351
71	110	00:00:01	1	1674,4	2311,2	92	108	18,2	21,4	2,6	3,5	18,3	21,1	30,14	0,239
72	40	00:01:52	112	5193,9	17278	199	212	26,1	81,5	7,3	7,6	21	27,3	581,72	0,725
73	40	00:00:24	24	3198	13182,4	156	214	20,5	61,6	5,5	7,5	17,6	27,6	79,95	0,458
74	110	00:00:00	0	1814,4	2343,5	96	109	18,9	21,5	3,1	3,5	18,7	21,1	45,36	0,259
75	31	00:00:26	26	3513,9	16686	159	206	22,1	81	5,5	7,1	18,3	27,2	94,88	0,502
76	110	00:00:01	1	2257,7	2386,5	107	111	21,1	21,5	3,5	3,5	20,8	20,9	60,96	0,323
77	30	00:00:21	21	3759,9	14580	151	180	24,9	81	6,1	7,1	20,9	26,9	78,96	0,525
78	30	00:00:22	22	2201,7	6876	123	191	17,9	36	4,6	7	16,5	27,4	48,44	0,316
79	110	00:00:00	0	1813,5	2373	93	105	19,5	22,6	3,1	3,5	18,7	21,2	39,9	0,259
80	30	00:00:39	39	3017,3	15267,8	143	194	21,1	78,7	5,7	7	18,4	27,7	117,67	0,432
81	110	00:00:00	0	1803,2	2385	92	106	19,6	22,5	2,3	3,5	18,6	21,1	70,32	0,258
82	31	00:00:30	31	3553,4	5992,8	163	227	21,8	26,4	5,6	7	18,7	27,7	110,16	0,509
83	110	00:00:00	0	1900	2319,9	100	111	19	20,9	2,6	3,5	20,9	21,1	58,9	0,271
84	41	00:02:13	133	4944,6	19521	201	241	24,6	81	7,7	8,4	20,1	27,2	662,58	0,709
85	110	00:00:01	1	1593,1	2322	89	108	17,9	21,5	2,3	3,5	18,7	21,1	213,48	0,228
86	42	00:00:53	53	4295,2	18263,7	182	223	23,6	81,9	7	8,4	18,8	27,6	231,94	0,614
87	110	00:00:01	1	2299,5	2353,2	105	106	21,9	22,2	3	3,5	21	21,1	124,17	0,328
88	60	00:00:10	10	4609,6	18554,6	172	226	26,8	82,1	6,5	8,6	21,1	27,5	46,1	0,643
89	30	00:00:42	42	2934,6	16170	134	196	21,9	82,5	5,4	7,5	18,4	27,8	126,19	0,42
90	110	00:00:01	1	2039,4	2299	103	110	19,8	20,9	3	3,5	21	21,1	87,69	0,291
91	37	00:00:09	9	4465,4	18682,1	166	227	26,9	82,3	6,8	8,6	21,3	27,1	392,96	1,03
92	30	00:01:18	78	3420	5207,4	152	198	22,5	26,3	6,2	8,5	19,5	21,3	300,96	0,49
93	110	00:00:01	1	1747,2	2245,4	96	109	18,2	20,6	3,1	3,5	18,8	21,3	153,75	0,25
94	44	00:01:57	117	5554,4	19024,2	212	234	26,2	81,3	8	8,3	21,1	28	649,86	0,775
95	44	00:01:20	80	5481	18983,7	210	237	26,1	80,1	7,9	8,3	21,2	27,7	438,48	0,765
96	32	00:00:16	16	4416	16696,8	184	216	24	77,3	6,1	7	21,1	26,9	70,66	0,616
97	32	00:00:16	16	4410,3	17620,4	183	217	24,1	81,2	6	7	21	27	70,56	0,615
98	32	00:00:15	15	4384,8	16871,4	174	206	25,2	81,9	6	7	21,3	27,3	65,77	0,612
99	32	00:00:16	16	4367,4	16789,2	174	204	25,1	82,3	6	7	21,1	27,1	69,88	0,609
100	70	00:00:17	17	5024,6	19199,2	194	233	25,9	82,4	7	8,3	21,3	27,7	85,42	0,701
101	33	00:01:35	95	4560	16686	190	206	24	81	6,3	6,6	21	27,2	433,2	0,636
102	70	00:00:18	18	5045,1	19263,4	201	239	25,1	80,6	7,2	8,3	21,1	27,5	90,81	0,704
103	34	00:00:08	8	3510,9	15566,5	141	191	24,9	81,5	4,7	6,5	21,1	27,3	28,09	0,49
104	34	00:00:07	7	3270,6	15403,5	138	189	23,7	81,5	4,6	6,5	21,2	26,2	22,89	0,456
105	34	00:00:10	10	3900	16134,3	156	197	25	81,9	5,1	6,5	20,9	26,8	39	0,544

Figure 9-6: Test set data. Including 136 instances (4/4)

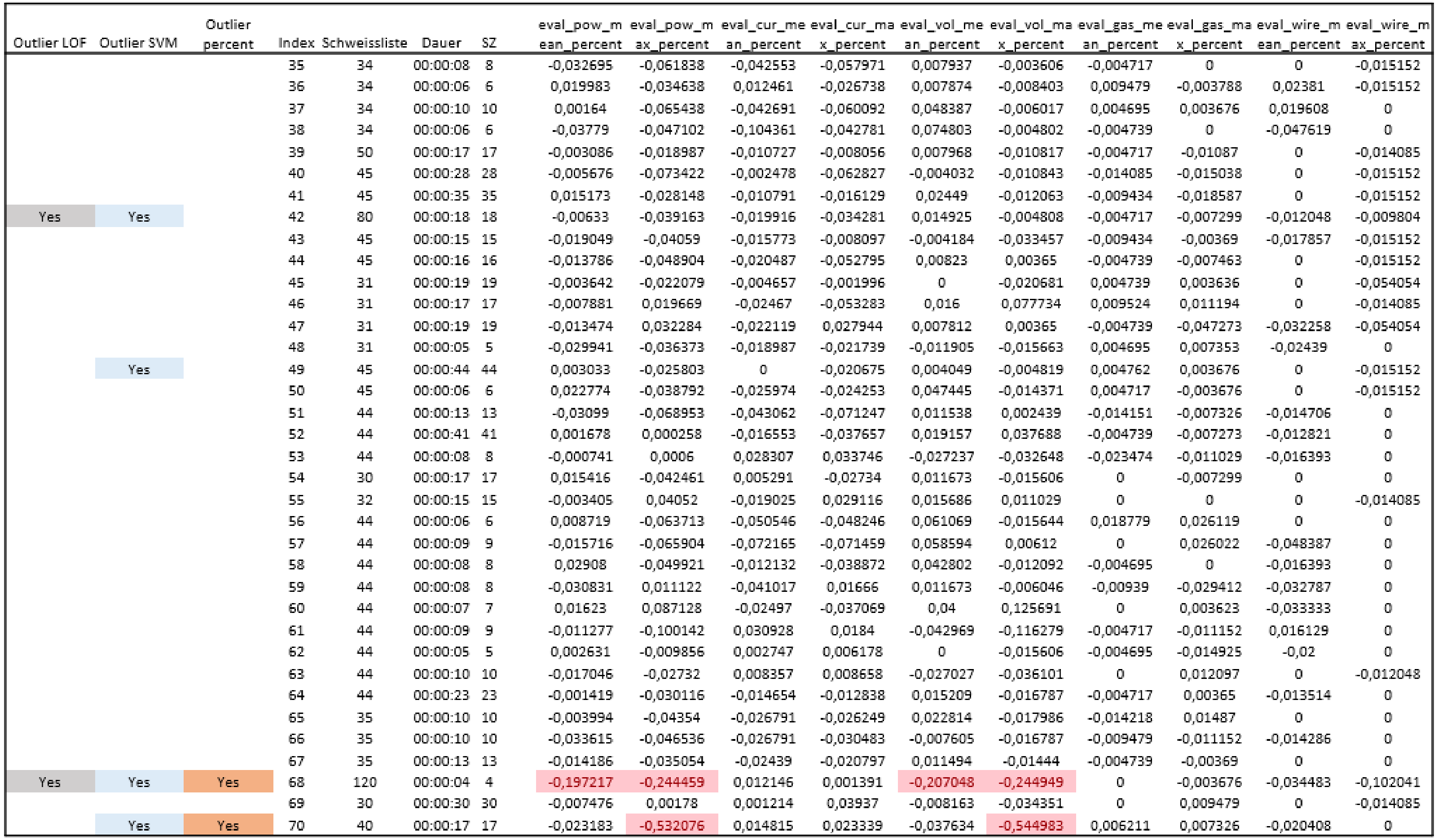


Index	Schweissliste	Dauer	SZ	Leistung Mittel [W] (1)	Leistung Max [W] (1)	Strom Mittel [A] (1)	Strom Max [A] (1)	Spannung Mittel [V] (1)	Spannung Max [V] (1)	Drahtvorschub Mittel [m/min] (1)	Drahtvorschub Max [m/min] (1)	Gasfluss Mittel [L/min] (1)	Gasfluss Max [L/min] (1)	Energie Mittel [kJ] (1)	Streckenenergie E Mittel [kJ/mm] (1)
106	34	00:00:06	6	3225	15479,1	125	189	25,8	81,9	4,1	6,5	21,1	26,3	19,35	0,45
107	50	00:00:16	16	4240,8	16236	171	198	24,8	82	6	7,1	21,1	26,8	67,85	0,592
108	45	00:00:27	27	4001,4	15121,8	162	186	24,7	81,3	5,8	6,5	21	26,2	108,04	0,558
109	45	00:00:35	35	4099,2	15048	168	190	24,4	79,2	6,1	6,5	21,2	26,9	143,47	0,572
110	80	00:00:17	17	5412,8	19967,5	199	245	27,2	81,5	8,5	10,1	21,2	27,7	92,02	1,16
111	45	00:00:14	14	3696	14311,2	160	201	23,1	71,2	5,6	6,5	20,9	26,9	51,74	0,516
112	45	00:00:15	15	3776,2	15701,4	158	198	23,9	79,3	5,6	6,5	21	26,6	56,64	0,527
113	31	00:00:19	19	4371,9	13703,4	177	207	24,7	66,2	6,2	7	20,9	26,9	83,07	0,61
114	31	00:00:16	16	4375	16373,7	175	207	25	79,1	6,1	7	21,1	27	70	0,61
115	31	00:00:18	18	4470,4	16827	176	213	25,4	79	6,2	7	21,2	26,2	80,47	0,624
116	31	00:00:05	5	3175,8	16456,5	134	207	23,7	79,5	4,2	7,1	21,5	27,3	15,88	0,443
117	45	00:00:44	44	4100,2	14851,2	166	182	24,7	81,6	6,1	6,6	21,1	27,3	180,41	0,572
118	45	00:00:06	6	2878,3	14326,4	107	176	26,9	81,4	3,8	6,5	20,9	26,9	17,27	0,402
119	44	00:00:12	12	4851,4	19239	191	242	25,4	79,5	6,9	8,3	21,1	27,1	63,07	0,677
120	44	00:00:42	42	5330	19626,2	205	242	26	81,1	7,7	8,3	20,9	26,9	223,86	0,744
121	44	00:00:08	8	4594,2	19048,3	186	239	24,7	79,7	6,3	8,3	21,4	27,3	36,75	0,641
122	30	00:00:17	17	3950	15686,4	158	192	25	81,7	5,9	7,1	21,1	27,5	67,15	0,551
123	32	00:00:15	15	4267,2	16503,9	168	203	25,4	81,3	5,9	7	20,8	26,9	64,01	0,595
124	44	00:00:05	5	3978	18536,4	153	228	26	81,3	5,1	8,3	21,1	26,8	19,89	0,555
125	44	00:00:08	8	4575,2	18780,3	172	231	26,6	81,3	6,1	8,3	21,3	26,9	36,6	0,638
126	44	00:00:08	8	4350	19031,1	174	237	25	80,3	6,1	8,3	20,9	27	34,8	0,607
127	44	00:00:08	8	4528,7	19704,6	179	246	25,3	80,1	6,2	8,3	21,3	26,2	36,23	0,632
128	44	00:00:08	8	4357,5	19325,6	175	238	24,9	81,2	6	8,3	21,3	27,5	34,86	0,608
129	44	00:00:08	8	4431,9	15325,5	187	255	23,7	60,1	6,3	8,2	20,9	26,4	35,46	0,618
130	44	00:00:06	6	3805,2	18722,4	151	232	25,2	80,7	4,9	8,3	21,3	26,1	22,83	0,531
131	44	00:00:09	9	4606	19858,8	188	247	24,5	80,4	6,3	8,3	20,8	25	41,45	0,643
132	44	00:00:23	23	5174	18617,7	199	229	26	81,3	7,3	8,3	20,9	27,3	119	0,722
133	35	00:00:10	10	4870,5	19704,6	191	246	25,5	80,1	7,1	9,1	20,7	26,9	48,7	0,68
134	35	00:00:10	10	4828,5	19035	185	235	26,1	81	6,8	9,1	21	26,6	48,28	0,674
135	35	00:00:12	12	4870,5	19129	191	235	25,5	81,4	7,4	9,1	20,8	26,7	58,45	0,68

Figure 9-7: Percentage difference of test data to average values of the same welding list and time. Including outlier labels. (1/4)

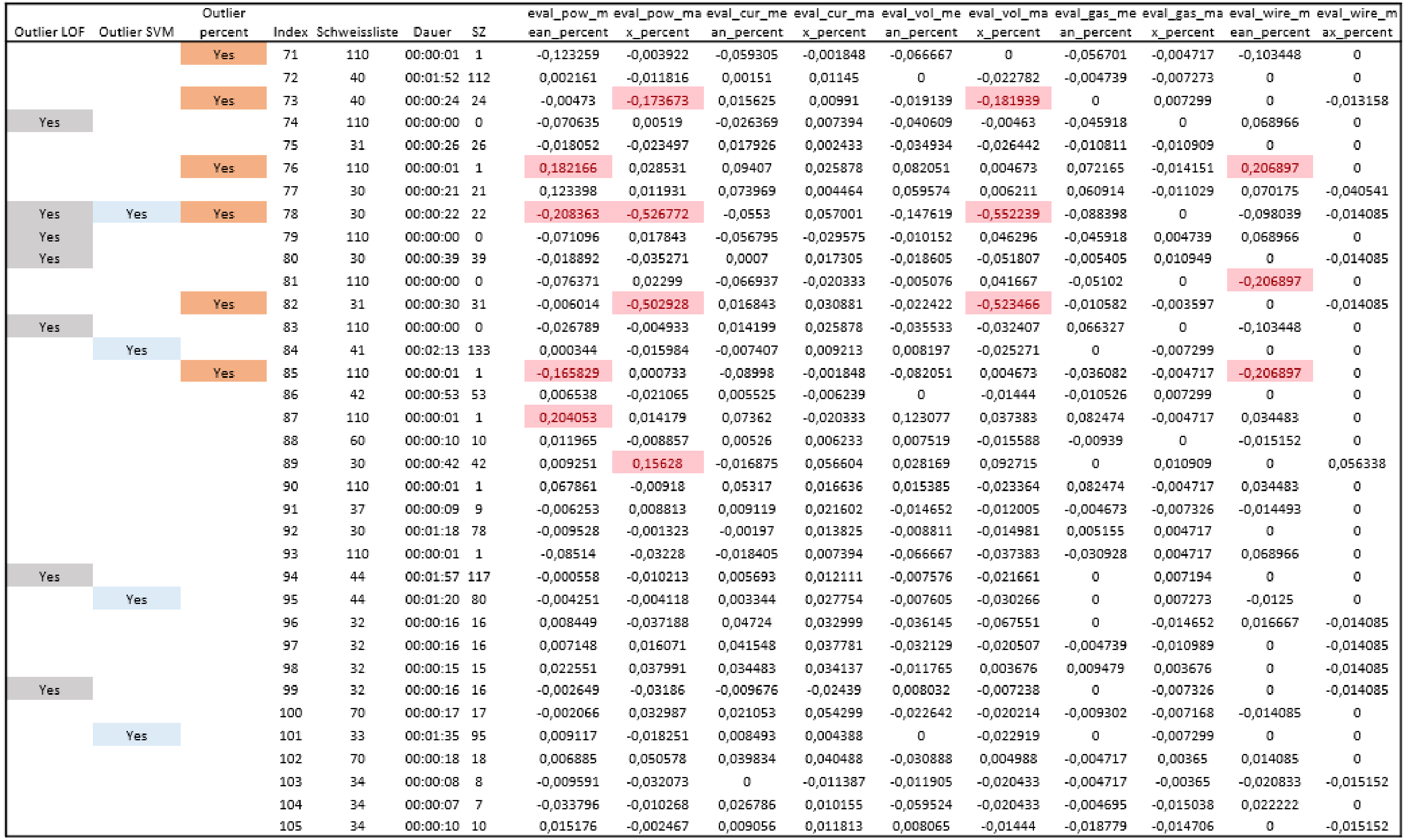
Outlier LOF	Outlier SVM	Outlier percent	Index	Schweissliste	Dauer	SZ	eval_pow_mean_percent	eval_pow_max_percent	eval_cur_mean_percent	eval_cur_max_percent	eval_vol_mean_percent	eval_vol_max_percent	eval_gas_mean_percent	eval_gas_max_percent	eval_wire_mean_percent	eval_wire_max_percent
			0	120	00:00:05	5	-0,04357	0,019023	0,004141	0,041809	-0,052632	-0,022167	0,013953	0,010949	0,035714	0
			1	30	00:00:29	29	-0,001485	-0,039815	-0,007264	-0,03527	0,004082	-0,003922	0	0	0	-0,014085
Yes	Yes	Yes	2	40	00:00:17	17	0,003465	0,212362	-0,007407	-0,008079	0,010753	0,216263	0,006211	0,007326	-0,020408	0
Yes			3	110	00:00:01	1	-0,060163	0,005042	-0,04908	-0,020333	-0,010256	0,028037	-0,041237	-0,004717	0,034483	0
			4	40	00:01:52	112	0,004766	-0,004541	-0,003523	0,01145	0,007663	-0,015588	0,004739	-0,007273	0	0
			5	40	00:00:25	25	0,008425	-0,045543	-0,001957	-0,019515	0,009524	-0,026415	0,011364	0,007299	0	0
Yes			6	110	00:00:00	0	-0,075449	0,014541	-0,036511	0,007394	-0,035533	0,00463	-0,040816	0,009479	0,068966	0
			7	31	00:00:26	26	-0,012352	-0,035552	-0,020487	-0,026764	0,008734	-0,009615	-0,005405	-0,003636	0	-0,014085
Yes			8	110	00:00:01	1	-0,050058	0,024006	-0,038855	-0,001848	-0,010256	0,028037	-0,036082	-0,004717	0,068966	0
			9	30	00:00:21	21	0,144133	0,006122	0,059744	-0,012277	0,093617	0,017391	0,076142	0,007353	0,070175	-0,040541
		Yes	10	30	00:00:21	21	-0,335116	-0,030184	-0,153627	0,010045	-0,204255	-0,040994	-0,167513	-0,011029	-0,192982	-0,054054
Yes			11	110	00:00:00	0	-0,086411	-0,014155	-0,077079	-0,038817	-0,005076	0,023148	-0,05102	0	0,068966	0
Yes			12	30	00:00:38	38	0,022443	-0,003765	0,003509	-0,020305	0,018868	0,01625	-0,026316	-0,031915	0	-0,014085
			13	110	00:00:00	0	-0,080623	0,028052	-0,056795	-0,011091	-0,020305	0,037037	-0,045918	0,004739	-0,206897	0
Yes			14	31	00:00:32	32	-0,006894	-0,02657	0,006289	0,071253	-0,0131	-0,091593	0	0,007194	0	-0,014085
			15	110	00:00:00	0	-0,161604	-0,004933	-0,056795	0,025878	-0,106599	-0,032407	-0,045918	0,004739	-0,068966	0
	Yes		16	41	00:02:13	133	0,00441	-0,013202	-0,007407	-0,007538	0,012295	-0,006017	0	0,00365	-0,012987	0
		Yes	17	110	00:00:00	0	0,167392	-0,017843	0,085193	-0,011091	0,081218	-0,009259	0,076531	0,004739	0,206897	0
			18	42	00:00:53	53	0,001008	-0,004261	0	0,00713	0	-0,01083	-0,005263	0	0	0
			19	110	00:00:01	1	-0,055817	0,023316	-0,059305	-0,020333	0,005128	0,046729	-0,036082	-0,004717	0,068966	0
Yes			20	60	00:00:10	10	0,018858	-0,00282	-0,006429	0,006233	0,026316	-0,009592	0,004695	0	0	0
			21	30	00:00:42	42	-0,001788	0,058551	-0,009538	-0,03504	0,00939	0,095364	-0,005435	-0,007273	0	0
			22	110	00:00:01	1	-0,064824	-0,008576	-0,02863	-0,001848	-0,035897	-0,004673	-0,041237	-0,004717	-0,206897	0
			23	37	00:00:09	9	0,012062	-0,05327	-0,009119	-0,045905	0,021978	-0,007203	0,004673	0,003663	0	0
			24	30	00:01:18	78	-0,007298	-0,006367	-0,008536	0,008705	0	-0,014981	0,010309	0,004717	0	0
			25	110	00:00:01	1	-0,085768	-0,032884	-0,00818	0,016636	-0,076923	-0,046729	-0,030928	0,004717	0,068966	0
			26	44	00:01:58	117	-0,006208	-0,020639	-0,003795	-0,009516	-0,003788	-0,01083	0	-0,007194	0	0
			27	44	00:01:20	80	-0,006177	-0,016819	-0,006211	-0,011275	0	-0,004843	-0,004717	-0,007273	-0,0125	0
			28	32	00:00:16	16	0,016853	0,016832	0,001707	0,023434	0,016064	-0,006031	0,004739	0,007326	0	0
			29	32	00:00:16	16	-0,002603	0,012905	0,018782	0,028216	-0,02008	-0,014475	0,004739	0,003663	0	-0,014085
			30	32	00:00:15	15	-0,005107	0,01129	-0,01308	-0,001004	0,007843	0,012255	0	0	-0,016667	0
			31	32	00:00:16	16	-0,002512	-0,012848	-0,043825	-0,010043	0,044177	-0,002413	0,004739	0	0	-0,014085
			32	70	00:00:18	18	-0,010298	-0,002029	-0,017072	-0,037875	0,007722	0,032419	-0,004717	0	0	0
			33	33	00:01:35	95	-0,005156	-0,043757	-0,018047	-0,039493	0,0125	-0,004825	0	-0,007299	0	0
			34	70	00:00:19	19	0,006546	0,006024	0,000508	0	0,007874	0,006098	0,004762	-0,007299	0	0

Figure 9-8: Percentage difference of test data to average values of the same welding list and time. Including outlier labels. (2/4)



Outlier LOF	Outlier SVM	Outlier percent	Index	Schweissliste	Dauer	SZ	eval_pow_mean_percent	eval_pow_max_percent	eval_cur_mean_percent	eval_cur_max_percent	eval_vol_mean_percent	eval_vol_max_percent	eval_gas_mean_percent	eval_gas_max_percent	eval_wire_mean_percent	eval_wire_max_percent
			35	34	00:00:08	8	-0,032695	-0,061838	-0,042553	-0,057971	0,007937	-0,003606	-0,004717	0	0	-0,015152
			36	34	00:00:06	6	0,019983	-0,034638	0,012461	-0,026738	0,007874	-0,008403	0,009479	-0,003788	0,02381	-0,015152
			37	34	00:00:10	10	0,00164	-0,065438	-0,042691	-0,060092	0,048387	-0,006017	0,004695	0,003676	0,019608	0
			38	34	00:00:06	6	-0,03779	-0,047102	-0,104361	-0,042781	0,074803	-0,004802	-0,004739	0	-0,047619	0
			39	50	00:00:17	17	-0,003086	-0,018987	-0,010727	-0,008056	0,007968	-0,010817	-0,004717	-0,01087	0	-0,014085
			40	45	00:00:28	28	-0,005676	-0,073422	-0,002478	-0,062827	-0,004032	-0,010843	-0,014085	-0,015038	0	-0,015152
			41	45	00:00:35	35	0,015173	-0,028148	-0,010791	-0,016129	0,02449	-0,012063	-0,009434	-0,018587	0	-0,015152
Yes	Yes		42	80	00:00:18	18	-0,00633	-0,039163	-0,019916	-0,034281	0,014925	-0,004808	-0,004717	-0,007299	-0,012048	-0,009804
			43	45	00:00:15	15	-0,019049	-0,04059	-0,015773	-0,008097	-0,004184	-0,033457	-0,009434	-0,00369	-0,017857	-0,015152
			44	45	00:00:16	16	-0,013786	-0,048904	-0,020487	-0,052795	0,00823	0,00365	-0,004739	-0,007463	0	-0,015152
			45	31	00:00:19	19	-0,003642	-0,022079	-0,004657	-0,001996	0	-0,020681	0,004739	0,003636	0	-0,054054
			46	31	00:00:17	17	-0,007881	0,019669	-0,02467	-0,053283	0,016	0,077734	0,009524	0,011194	0	-0,014085
			47	31	00:00:19	19	-0,013474	0,032284	-0,022119	0,027944	0,007812	0,00365	-0,004739	-0,047273	-0,032258	-0,054054
			48	31	00:00:05	5	-0,029941	-0,036373	-0,018987	-0,021739	-0,011905	-0,015663	0,004695	0,007353	-0,02439	0
	Yes		49	45	00:00:44	44	0,003033	-0,025803	0	-0,020675	0,004049	-0,004819	0,004762	0,003676	0	-0,015152
			50	45	00:00:06	6	0,022774	-0,038792	-0,025974	-0,024253	0,047445	-0,014371	0,004717	-0,003676	0	-0,015152
			51	44	00:00:13	13	-0,03099	-0,068953	-0,043062	-0,071247	0,011538	0,002439	-0,014151	-0,007326	-0,014706	0
			52	44	00:00:41	41	0,001678	0,000258	-0,016553	-0,037657	0,019157	0,037688	-0,004739	-0,007273	-0,012821	0
			53	44	00:00:08	8	-0,000741	0,0006	0,028307	0,033746	-0,027237	-0,032648	-0,023474	-0,011029	-0,016393	0
			54	30	00:00:17	17	0,015416	-0,042461	0,005291	-0,02734	0,011673	-0,015606	0	-0,007299	0	0
			55	32	00:00:15	15	-0,003405	0,04052	-0,019025	0,029116	0,015686	0,011029	0	0	0	-0,014085
			56	44	00:00:06	6	0,008719	-0,063713	-0,050546	-0,048246	0,061069	-0,015644	0,018779	0,026119	0	0
			57	44	00:00:09	9	-0,015716	-0,065904	-0,072165	-0,071459	0,058594	0,00612	0	0,026022	-0,048387	0
			58	44	00:00:08	8	0,02908	-0,049921	-0,012132	-0,038872	0,042802	-0,012092	-0,004695	0	-0,016393	0
			59	44	00:00:08	8	-0,030831	0,011122	-0,041017	0,01666	0,011673	-0,006046	-0,00939	-0,029412	-0,032787	0
			60	44	00:00:07	7	0,01623	0,087128	-0,02497	-0,037069	0,04	0,125691	0	0,003623	-0,033333	0
			61	44	00:00:09	9	-0,011277	-0,100142	0,030928	0,0184	-0,042969	-0,116279	-0,004717	-0,011152	0,016129	0
			62	44	00:00:05	5	0,002631	-0,009856	0,002747	0,006178	0	-0,015606	-0,004695	-0,014925	-0,02	0
			63	44	00:00:10	10	-0,017046	-0,02732	0,008357	0,008658	-0,027027	-0,036101	0	0,012097	0	-0,012048
			64	44	00:00:23	23	-0,001419	-0,030116	-0,014654	-0,012838	0,015209	-0,016787	-0,004717	0,00365	-0,013514	0
			65	35	00:00:10	10	-0,003994	-0,04354	-0,026791	-0,026249	0,022814	-0,017986	-0,014218	0,01487	0	0
			66	35	00:00:10	10	-0,033615	-0,046536	-0,026791	-0,030483	-0,007605	-0,016787	-0,009479	-0,011152	-0,014286	0
			67	35	00:00:13	13	-0,014186	-0,035054	-0,02439	-0,020797	0,011494	-0,01444	-0,004739	-0,00369	0	0
Yes	Yes	Yes	68	120	00:00:04	4	-0,197217	-0,244459	0,012146	0,001391	-0,207048	-0,244949	0	-0,003676	-0,034483	-0,102041
			69	30	00:00:30	30	-0,007476	0,00178	0,001214	0,03937	-0,008163	-0,034351	0	0,009479	0	-0,014085
	Yes	Yes	70	40	00:00:17	17	-0,023183	-0,532076	0,014815	0,023339	-0,037634	-0,544983	0,006211	0,007326	-0,020408	0

Figure 9-9: Percentage difference of test data to average values of the same welding list and time. Including outlier labels. (3/4)



Outlier LOF	Outlier SVM	Outlier percent	Index	Schweissliste	Dauer	SZ	eval_pow_mean_percent	eval_pow_max_percent	eval_cur_mean_percent	eval_cur_max_percent	eval_vol_mean_percent	eval_vol_max_percent	eval_gas_mean_percent	eval_gas_max_percent	eval_wire_mean_percent	eval_wire_max_percent
		Yes	71	110	00:00:01	1	-0,123259	-0,003922	-0,059305	-0,001848	-0,066667	0	-0,056701	-0,004717	-0,103448	0
			72	40	00:01:52	112	0,002161	-0,011816	0,00151	0,01145	0	-0,022782	-0,004739	-0,007273	0	0
		Yes	73	40	00:00:24	24	-0,00473	-0,173673	0,015625	0,00991	-0,019139	-0,181939	0	0,007299	0	-0,013158
Yes			74	110	00:00:00	0	-0,070635	0,00519	-0,026369	0,007394	-0,040609	-0,00463	-0,045918	0	0,068966	0
			75	31	00:00:26	26	-0,018052	-0,023497	0,017926	0,002433	-0,034934	-0,026442	-0,010811	-0,010909	0	0
		Yes	76	110	00:00:01	1	0,182166	0,028531	0,09407	0,025878	0,082051	0,004673	0,072165	-0,014151	0,206897	0
			77	30	00:00:21	21	0,123398	0,011931	0,073969	0,004464	0,059574	0,006211	0,060914	-0,011029	0,070175	-0,040541
Yes	Yes	Yes	78	30	00:00:22	22	-0,208363	-0,526772	-0,0553	0,057001	-0,147619	-0,552239	-0,088398	0	-0,098039	-0,014085
Yes			79	110	00:00:00	0	-0,071096	0,017843	-0,056795	-0,029575	-0,010152	0,046296	-0,045918	0,004739	0,068966	0
Yes			80	30	00:00:39	39	-0,018892	-0,035271	0,0007	0,017305	-0,018605	-0,051807	-0,005405	0,010949	0	-0,014085
			81	110	00:00:00	0	-0,076371	0,02299	-0,066937	-0,020333	-0,005076	0,041667	-0,05102	0	-0,206897	0
		Yes	82	31	00:00:30	31	-0,006014	-0,502928	0,016843	0,030881	-0,022422	-0,523466	-0,010582	-0,003597	0	-0,014085
Yes			83	110	00:00:00	0	-0,026789	-0,004933	0,014199	0,025878	-0,035533	-0,032407	0,066327	0	-0,103448	0
	Yes		84	41	00:02:13	133	0,000344	-0,015984	-0,007407	0,009213	0,008197	-0,025271	0	-0,007299	0	0
		Yes	85	110	00:00:01	1	-0,165829	0,000733	-0,08998	-0,001848	-0,082051	0,004673	-0,036082	-0,004717	-0,206897	0
			86	42	00:00:53	53	0,006538	-0,021065	0,005525	-0,006239	0	-0,01444	-0,010526	0,007299	0	0
			87	110	00:00:01	1	0,204053	0,014179	0,07362	-0,020333	0,123077	0,037383	0,082474	-0,004717	0,034483	0
			88	60	00:00:10	10	0,011965	-0,008857	0,00526	0,006233	0,007519	-0,015588	-0,00939	0	-0,015152	0
			89	30	00:00:42	42	0,009251	0,15628	-0,016875	0,056604	0,028169	0,092715	0	0,010909	0	0,056338
			90	110	00:00:01	1	0,067861	-0,00918	0,05317	0,016636	0,015385	-0,023364	0,082474	-0,004717	0,034483	0
			91	37	00:00:09	9	-0,006253	0,008813	0,009119	0,021602	-0,014652	-0,012005	-0,004673	-0,007326	-0,014493	0
			92	30	00:01:18	78	-0,009528	-0,001323	-0,00197	0,013825	-0,008811	-0,014981	0,005155	0,004717	0	0
			93	110	00:00:01	1	-0,08514	-0,03228	-0,018405	0,007394	-0,066667	-0,037383	-0,030928	0,004717	0,068966	0
Yes			94	44	00:01:57	117	-0,000558	-0,010213	0,005693	0,012111	-0,007576	-0,021661	0	0,007194	0	0
	Yes		95	44	00:01:20	80	-0,004251	-0,004118	0,003344	0,027754	-0,007605	-0,030266	0	0,007273	-0,0125	0
			96	32	00:00:16	16	0,008449	-0,037188	0,04724	0,032999	-0,036145	-0,067551	0	-0,014652	0,016667	-0,014085
			97	32	00:00:16	16	0,007148	0,016071	0,041548	0,037781	-0,032129	-0,020507	-0,004739	-0,010989	0	-0,014085
			98	32	00:00:15	15	0,022551	0,037991	0,034483	0,034137	-0,011765	0,003676	0,009479	0,003676	0	-0,014085
Yes			99	32	00:00:16	16	-0,002649	-0,03186	-0,009676	-0,02439	0,008032	-0,007238	0	-0,007326	0	-0,014085
			100	70	00:00:17	17	-0,002066	0,032987	0,021053	0,054299	-0,022642	-0,020214	-0,009302	-0,007168	-0,014085	0
	Yes		101	33	00:01:35	95	0,009117	-0,018251	0,008493	0,004388	0	-0,022919	0	-0,007299	0	0
			102	70	00:00:18	18	0,006885	0,050578	0,039834	0,040488	-0,030888	0,004988	-0,004717	0,00365	0,014085	0
			103	34	00:00:08	8	-0,009591	-0,032073	0	-0,011387	-0,011905	-0,020433	-0,004717	-0,00365	-0,020833	-0,015152
			104	34	00:00:07	7	-0,033796	-0,010268	0,026786	0,010155	-0,059524	-0,020433	-0,004695	-0,015038	0,022222	0
			105	34	00:00:10	10	0,015176	-0,002467	0,009056	0,011813	0,008065	-0,01444	-0,018779	-0,014706	0	-0,015152

Figure 9-10: Percentage difference of test data to average values of the same welding list and time. Including outlier labels. (4/4)

Outlier LOF	Outlier SVM	Outlier percent	Index	Schweissliste	Dauer	SZ	eval_pow_mean_percent	eval_pow_max_percent	eval_cur_mean_percent	eval_cur_max_percent	eval_vol_mean_percent	eval_vol_max_percent	eval_gas_mean_percent	eval_gas_max_percent	eval_wire_mean_percent	eval_wire_max_percent
			106	34	00:00:06	6	-0,011585	-0,006004	-0,02648	0,010695	0,015748	-0,016807	0	-0,003788	-0,02381	-0,015152
			107	50	00:00:16	16	0,0144	-0,025462	0,020896	-0,009505	-0,008	-0,015606	0	-0,021898	0	0
			108	45	00:00:27	27	-0,000599	-0,038897	0,004963	-0,014831	-0,004032	-0,02401	0	-0,003802	-0,016949	-0,015152
			109	45	00:00:35	35	0,004804	-0,024238	0,007194	0,021505	-0,004082	-0,044632	0	0	0,016667	-0,015152
Yes	Yes		110	80	00:00:17	17	0,029656	-0,002004	0,036458	0,025963	-0,007299	-0,027446	0	0,007273	0,011905	0
Yes	Yes		111	45	00:00:14	14	-0,017126	-0,054243	0,012658	0,036082	-0,029412	-0,087179	-0,018779	-0,014652	0	-0,015152
			112	45	00:00:15	15	-0,002299	-0,014647	-0,003155	0,002024	0	-0,017348	-0,009434	-0,01845	0	-0,015152
Yes		Yes	113	31	00:00:19	19	-0,004939	-0,16765	0,030268	0,032934	-0,035156	-0,194647	-0,009479	-0,021818	0	-0,054054
			114	31	00:00:16	16	0,005863	0,008096	0,008065	-0,008146	0	0,015404	-0,004717	-0,00369	0	-0,014085
Yes			115	31	00:00:18	18	0,008983	-0,018674	0,016166	0,017678	-0,007813	-0,035409	0,004739	-0,003802	0,016393	-0,014085
			116	31	00:00:05	5	-0,002231	-0,0197	0,060127	0,022727	-0,059524	-0,042169	0,00939	0,003676	0,02439	0
Yes	Yes		117	45	00:00:44	44	0,011172	-0,026904	0,012195	-0,009793	0	-0,016867	0,004762	0,003676	0	0
			118	45	00:00:06	6	-0,023113	-0,032817	-0,007421	-0,007332	-0,018248	-0,02515	-0,014151	-0,011029	0	-0,015152
			119	44	00:00:12	12	0,002915	-0,009723	0,02358	0,029787	-0,019305	-0,038694	-0,004717	-0,014545	0	0
			120	44	00:00:42	42	-0,007301	0,009319	-0,001947	0,025424	-0,003831	-0,015777	-0,004762	-0,014652	-0,012821	0
			121	44	00:00:08	8	0,031639	-0,01551	0,074523	0,020931	-0,038911	-0,036276	0,004695	0,003676	0,032787	0
			122	30	00:00:17	17	0,014902	-0,009866	0,044974	0,009464	-0,027237	-0,019208	0	0,00365	0,017241	0
			123	32	00:00:15	15	-0,004874	0,015381	-0,001189	0,019076	-0,003922	-0,003676	-0,014218	-0,011029	-0,016667	-0,014085
Yes			124	44	00:00:05	5	0,046677	-0,018308	0,050824	0,006178	-0,003831	-0,02401	-0,00939	0	0,02	0
			125	44	00:00:08	8	0,027373	-0,029362	-0,006355	-0,013242	0,035019	-0,016929	0	-0,011029	0	0
			126	44	00:00:08	8	-0,023196	-0,016399	0,005199	0,012388	-0,027237	-0,029021	-0,018779	-0,007353	0	0
			127	44	00:00:08	8	0,016931	0,01841	0,034084	0,050833	-0,015564	-0,031439	0	-0,036765	0,016393	0
			128	44	00:00:08	8	-0,021512	-0,001178	0,010976	0,01666	-0,031128	-0,018138	0	0,011029	-0,016393	0
		Yes	129	44	00:00:08	8	-0,004805	-0,207919	0,0803	0,089278	-0,077821	-0,273277	-0,018779	-0,029412	0,032787	-0,012048
			130	44	00:00:06	6	-0,006683	-0,012454	0,031421	0,017544	-0,038168	-0,028881	0	-0,026119	-0,02	0
			131	44	00:00:09	9	0,032666	0,039951	0,076747	0,056911	-0,042969	-0,015912	-0,018868	-0,070632	0,016129	0
			132	44	00:00:23	23	-0,007653	-0,012523	0,005558	0,013723	-0,011407	-0,02518	-0,014151	-0,00365	-0,013514	0
			133	35	00:00:10	10	0,013126	0,000513	0,044286	0,04149	-0,030418	-0,039568	-0,018957	0	0,014286	0
			134	35	00:00:10	10	0,004389	-0,033487	0,011482	-0,00508	-0,007605	-0,028777	-0,004739	-0,011152	-0,028571	0
			135	35	00:00:12	12	-0,00994	-0,011707	0,0138	0,012931	-0,022989	-0,023981	-0,018868	-0,025547	0,013699	0