CHALMERS
UNIVERSITY OF TECHNOLOGY

A data-driven approach to detect air leakage in a pneumatic system

Master's thesis in Production engineering

JOHAN LENÉ
MOHAN RAJASHEKARAPPA

Department of Industrial and Materials Science

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

A data-driven approach to detect air leakage in a pneumatic system

JOHAN LENÉ
MOHAN RAJASHEKARAPPA

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
Production Engineering
Chalmers University of Technology
Gothenburg, Sweden 2021

A data-driven approach to detect air leakage in a pneumatic system
JOHAN LENÉ, MOHAN RAJASHEKARAPPA

Supervisors: Ebru Turanoğlu Bekar, Chalmers University of Technology
Thomas Sundquist, SKF AB
Robert Andersson Jarl, SKF AB
Jonas Vallström, SKF AB
Examiner: Anders Skoogh, Chalmers University of Technology

Master's Thesis 2021
Department of Industrial and Materials Science
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Gothenburg, Sweden 2021

A data-driven approach to detect air leakage in a pneumatic system
JOHAN LENÉ
MOHAN RAJASHEKARAPPA
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

Maintenance practices in production systems have for the past decades followed a reactive approach. With the production domain currently shifting towards a new era, Industry 4.0, maintenance practises have experienced an incremental drift towards predictive approach. With key enabling technologies such as BigData, Industrial Digitalization, and Machine Learning that enables data-driven decision making, the field of production is now more ready than ever to establish new practises in maintenance. The fast-emerging area of data-driven decision making with key enabling technologies is where this thesis finds its roots.

This thesis work aims to develop a data driven approach based on machine learning to identify an early stage of a pneumatic leakage in a production process, before the leakage is so severe that it would impact the process adversely and cause machine breakdown. Through two experiments and extraction of historical data coming from sensors, a supervised machine learning model was built on the extracted significant statistical features. Adding on, an unsupervised model was developed to distinguish separate clusters representing the normal working state and leaking state of the machine for another process in the same production line.

The supervised machine learning model is successful in detecting the early stage of leakage with an accuracy of 98.2%, thus giving plenty of time to perform appropriate maintenance on the equipment. The unsupervised model with an accuracy of 99.3%, is successful in detecting the physically evident stage of leakage. The presented thesis makes salient recommendations for deployment of this model. Moreover, the provided valuable insights pertaining to features which were previously believed to be insignificant for modelling purposes, provide a standardized work methodology and a concrete platform for future study in the area.

Keywords: Machine learning, data driven decision making, predictive maintenance, pneumatic leakage, sensor data.

Acknowledgements

This master thesis has been written at the Production Engineering program at Chalmers University of Technology and in cooperation with project PACA (Predictive Maintenance using Advanced Cluster Analysis) which is an ongoing research project in the Department of Industrial and Materials Science at the Chalmers University of Technology. We would like to express our deep gratitude to our Academical Supervisor Dr. Ebru Turanoğlu Bekar who is a post-doctoral researcher at department of Industrial and Materials Science at Chalmers University of Technology for her patient guidance, enthusiastic encouragement and useful critiques of this thesis work.
We would also like to thank Alexander Karlsson who is a Senior Lecturer in the department of School of Informatics at University of Skövde for his advice and assistance in keeping our progress on schedule.
Our grateful thanks are also extended to Anders Skoogh who is a Professor of Production Maintenance and Director of Master's Programme in Production Engineering, division of Production systems, Department of Industrial and Materials Science for his valuable and constructive suggestions.
Last but not the least, we would also like to express very great appreciation to our Industrial Supervisor Thomas Sundqvist who is the Maintenance Engineer at SKF AB, Robert Andersson Jarl who is a Automation Engineer at SKF AB, and Jonas Vallström who is a Manufacturing Reliability Champion at SKF AB, for their professional guidance and valuable support enabling us to visit the company and observe the operations.

Johan Lené, Gothenburg, June 2021
Mohan Rajashekarappa, Gothenburg, June 2021

Contents

List of Figures

List of Tables

1
Introduction

In this chapter the background, aim, delimitations, and research questions will be presented, as well as a short description of the production line and the investigated processes within the scope of the thesis.

1.1 Background

Production systems have long followed a reactive approach for maintenance practices, but during the past decades they have shown an interesting development of more proactive approaches, mainly focussing on the relevance of failure avoidance [1]. Industry 4.0 and Industrial Digitalization is a key enabler for developing the field of maintenance, especially for developing predictive maintenance (PdM) in the manufacturing industry with Artificial Intelligence (AI) and Machine Learning (ML) solutions. Data-driven decision-making is an important precondition within industrial digitalization. The integration of IT and communication technology with the manufacturing process has led to the growth of smart manufacturing, which is reforming the traditional manufacturing industry. In consequence, tons of machine data are generated which can be used to make data-driven decision-making using computerized algorithms. It is evident that the enabling technologies of Industry 4.0 such as Internet of things (IoT), big data, cloud, etc. enables manufacturing companies to collect huge volumes of shop floor data as a part of manufacturing execution systems. Getting a good grasp of this new abundance of data while simultaneously having to search for meaningful knowledge, as well as developing insights with the help of informative analytical algorithms, is a recent challenge for these companies. This fast-emerging area of "Introduction of intelligence to control maintenance processes" is where this thesis finds its roots.
The project PACA (Predictive Maintenance using Advanced Cluster Analysis), which is an ongoing research project in the Department of Industrial and Materials Science at the Chalmers University of Technology, develops a framework for PdM that can exploit an algorithm in order to give a recommendation for the purpose of effective maintenance planning. AB SKF is one of the industrial partners of this project and is interested in analyzing and understanding data coming from multiple streams to investigate a specific packaging process if an AI/ML system can predict failures from the data. This process is pneumatically driven, and leakage in the tubing is quite common due to the high temperatures generated. These malfunctions are often detected too late, causing costly unexpected downtime, thus stressing on an urgent need for a system of early detection possibly in the form of smart alarms to inform

maintenance personnel for effective planning.

1.2 Aim

The objective of this master thesis is to find a data-driven detection method that can predict future failures on the pneumatic system by detecting an early stage of a leakage. Further to provide a standardised methodology for any kind of future work related to predictive maintenance on pneumatic leakages.

1.3 Research Questions

Knowing the aim of the thesis, the following research questions are specified for investigation:

- RQ1: How can data be prepared and labeled in case of a lack(or limited access) of historic evidence of fault?
- RQ2: What are the significant features that help to build a machine learning model to detect the pneumatic leakage in a system?
- RQ3: Can an unsupervised approach find distinct clusters for another machine with unlabeled data based on the features derived from RQ2?

1.4 Process and Problem Descriptions at Case Company

This section will give the reader an overview of the production line and its purpose, and a deeper insight of the processes further down. The wrapping and packaging machine are at the end of a highly automated flow production line in the D-plant at SKF Gothenburg. The production line assembles bearing rolls, inner bearing ring and outer bearing ring to complete the bearing as a whole. After the bearings are assembled they are greased with oil to withstand corrosion and to function as lubrication. Now the bearings are wrapped separately into a plastic bag, where the bag is sealed by making a seam with a high-temperature stamp and cut with a vertical knife. The bearings are transported on a conveyor to the packaging process, where they are packed in cardboard which is glued, and also labeled for deliveries.

Everytime the model of bearing is changed there is a new setup on the machines in the flow. For the two processes this thesis looks into, different setups do not change the data from the sensors, since the only change is for guides that control the movement of the bearing.

In this plant there are two separate production lines that have worked in the same way, and where the two processes wrapping and packaging schematically look ex-

actly the same. Though there are some minor differences when comparing extracted data from both of the machines, this is due to a higher constant leakage for one of the processes. When needed in this thesis, the processes will be separated as Pack1 and Pack2, respectively Wrapper1 and Wrapper2 to distinguish the separate lines.

As a security precaution, all the processes and machines in the production flow are protected with barriers and security doors, separating them from the operators and creating a safer work environment. For the processes this thesis works with, if a security door or security hatch is opened, the pneumatics are turned off and the whole process stops instantly. This is usually done during maintenance of the machines, and setups, or when there is some minor issue that can be fixed by the operators.

1.4.1 Wrapping Process

The bearings arrive at the wrapping process from a lubrication process by a conveyor one bearing at a time. They are transported onto a long plastic bag, which is folded and seamed by heat around the bearing. The bearings are dragged/transported to a cutter to close the "short ends" of the bag. When the bearing has arrived at its position the conveyor stops, one clamp fixates the bearing to its position, and another vertical clamp puts pressure on the bag on one side, heating the plastic to conceal the bearing. The bags are separated by a knife moving in the vertical direction inside the seaming clamp, as seen in figure 1.1. The knife and the clamps are pneumatically driven by compressed air. The most common error in this process that is investigated is to find leakages for the pneumatics to the knife, which is applied on both ends of the knife pushing it down. When leakage occurs the knife is only pushed downwards on one side, causing it to stick or make an unfinished cut of the plastic bag.

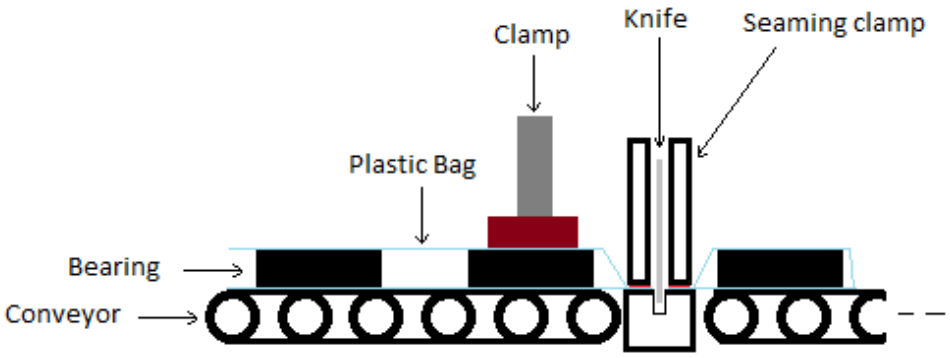

Figure 1.1: Conceptual model representing the wrapping process seen from the side

1.4.2 Packaging Process

This process has two separate inputs. One is the wrapped bearings from the previous process arriving from one conveyor, and the other is a flat cardboard that is refilled regularly. In this process one wrapped bearing is packed in each box. Figure 1.2 illustrates a conceptual model of the packaging process seen from above.

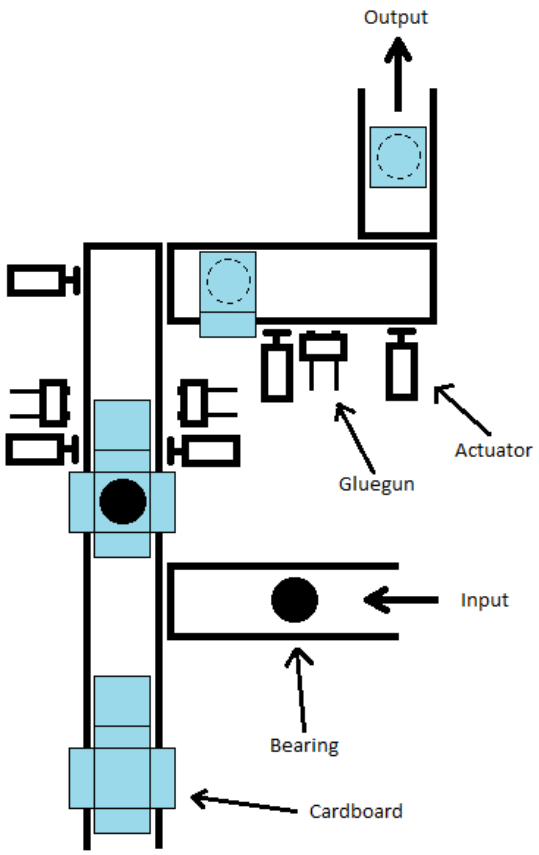

Figure 1.2: Conceptual model representing the packaging process seen from above

Initially the bearing is moved onto the middle of the cardboard by an actuator pushing it to the correct position. The two components are transferred forward by the conveyor, where actuators fold the sides and the top of the cardboard. The glue gun is activated when the box has arrived at a starting position spraying two strings of glue on both sides of the box whilst moving forward, covering the length of the side on the cardboard. Actuators then put pressure on the outside of the box so that the sides stick together. Later it's moved onto the next conveyor by an actuator, making the same process but folding and glueing the last side of the box to conceal the bearing completely. See figure 1.3 for a conceptual view of the glue process and folding.

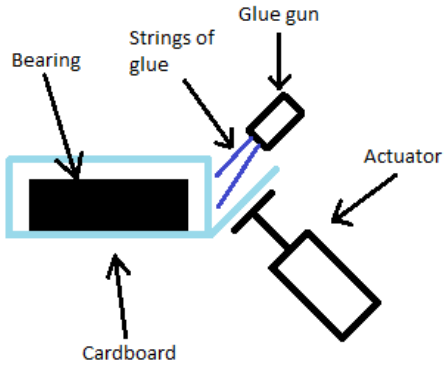

Figure 1.3: Conceptual view of the glue gun and folding process

Lastly in the process the box is labeled by a labeling machine and then transferred to a pallet for delivery. The three glue guns and the actuators are pneumatically driven, where compressed air activates the movement of the actuators and sprays a liquid string of glue from the glue guns.

1.4.3 Sensors

Two separate sensors record data for both of the processes described above. Both of them are of the brand IFM and model SD6500, and give a precise measurement of airflow, the temperature of the air, and pressure. They are both placed at the inlet of the compressed air to the whole pneumatic system, hence not detecting any pressure drops within the system. More details of the data collection and the sensors is described in chapter 4.1

1.5 Delimitations

This thesis work will only consider the specified pneumatic packaging process and wrapping process the development of machine learning algorithms. Possible deployment will not be considered by this thesis group, instead handled internally by SKF. Due to the current COVID situation, both parties need to be flexible and respect changes from the public health authority, Chalmers or AB SKF regarding company visits and transportation and the limitations that might follow.

2
Theory

This chapter is a short summary of the literature review that gives the reader a brief understanding about the terminologies and concepts necessary to fully understand this thesis work. Adding on, this chapter consists of a brief summary of the related work previously done in this field and gives some insights regarding gaps and challenges in this area.

2.1 Production Disturbances

Scheduling and controlling manufacturing systems is a challenge. The primary reason for this is its complexity. Machine breakdowns, operational delays and rush orders are few among plethora of reasons for the production disturbances [2]. Production Disturbance is defined as an unpredictable disruption(internal or external) to the production process[2]. Few examples with the categories of their origin are stated below [3]:

-Upstream disturbances: These disturbances correspond to the delays and disruptions related to the inputs or raw materials which fulfill the basic requirements of the beginning phase of the production process. Examples: Materials quality problems, supplier production problems, materials delivery delays, material property variations, incorrect deliveries[2], [3].

-Internal disturbances: Disturbances which originate internally because of the nature of the functionality of the institution are called internal disturbances. They can be categorized as follows:

- Information, control and decision-making: Examples: Control and communication system failures, operator errors and omissions, recording / communication errors, materials ordering errors, materials stock control problems.
- Production Equipment and Labour: Machine breakdowns, variability in machine performance (quality, cost, production rate), unavailability of labour
- Material Handling and Flow: Blockages, handling equipment failure[2], [3].

-Downstream Disturbances The disturbances caused to the processes involved in the conversion of upstream stage into finished product or during the sale of the finished goods are termed downstream disturbances. Other than the malfunctioning in the tail end of the manufacturing process, their origin can also be tracked to customers

involved in the sale as well. They can be categorized as follows:

- Make to order: Rush orders, changes to orders (quantity, due date), quantity and mix variations (e.g. due to demand variations in customer's business), customer production problems[2], [3].
- Make to stock: Demand variations (e.g. due to seasonality, marketing activity, competitor activity), forecasting errors, finished goods delivery delays, lost stock, poor stock monitoring[2], [3].

Irrespective of the kind of production disturbances, they always tend to pose a problem for the smooth functioning of the production system. Maintenance is an important function in a production system which can tackle production disturbances and provide a failure free performance under stated conditions.

2.2 Maintenance

According to the British Standard Glossary of terms[6], Maintenance is defined as:

"The combination of all technical and administrative actions, including supervision actions, intended to retain an item in, or restore it to, a state in which it can perform a required function. Maintenance is a set of organized activities that are carried out in order to keep an item in its best operational condition with minimum cost acquired."

The objectives of maintenance are subordinates to the production goals[7]. The sophistication or the degree of maintenance management is defined by the production goals. In simple words, a system with elite production goals requires its machines and equipment to be well conditioned and with good reliability[7]. Some of the basic maintenance objectives are:

- Increasing the production
- Efficient energy consumption
- Provide good and safe conditions for the labourers to work
- Reducing down times or emergency stoppages
- Provide good equipment conditioning thus improving equipment efficiency and reduce scrap rate

The dependability of a production system has three important pillars- Reliability, Maintainability and maintenance support [8].

2.2.1 History of Maintenance

2.2.1.1 Maintenance in the first generation

Before the 1950s, most of the machines used were simple and had low down times. This was the era of 'Reactive Maintenance'. The Focus was to complete the main-

tenance process with minimum time as possible with a 'Run to failure' approach. Maintenance was termed as a 'necessary evil'[8].

2.2.1.2 Maintenance in the second generation

During the 1950s, Industries became more dependent on equipment making them complex. There was an urgent need to stop the machine from failure as down times were not a thing to afford, thus giving birth to the idea of preventive maintenance. This was the time period when first "maintenance departments" were introduced in the industries[8]. During the 1960s Maintenance measurement terms were coined. MTBF(mean time between failure) and MTTR(Mean time to repair), maintainability(rate at which machine can be maintained), reliability (the rate at which equipment fails is low), cost of maintenance were some of the terms on which light was thrown for the first time. This generation saw the beginning of predictive maintenance; but as a mere terminology and still had a long way to be practically implemented on a large scale Soon the maintenance issue was more of a "technical matter" rather a "Necessary evil"[8].

2.2.1.3 Maintenance in the Third generation (1980-2000)

This was seen as the time of growth of computer technology and its application to the manufacturing sector. Equipment were smart machines with capabilities to self monitor, self calibrate, and self adjust. This could be precisely seen as the birth time of the digital age[8]. Computerized maintenance and management systems(CMMS) were just beginning to boom with tracking and recording of every maintenance based activity into databases. This generation witnessed the view of maintenance changing from ""Technical issue" to a "Profit contributor" [8].

2.2.1.4 Maintenance after 2000: Industry 4.0 and Digitalization

There is a steady growth in the complexity of the product as well as the equipment. Customization of products has led to frequent alteration or replacement of equipment, thus making the maintenance management domain an important research area[8]. Principles and technologies from the Internet of Things (IoT) applied to the manufacturing industry was the backbone of Industry 4.0[8]. The key features of Industry 4.0 belong to three main categories namely; Human, Technology and organization[1]. Stressing on the category of Technology, Automation and Big data are one of the main key enablers. Pertaining to the organization category, Interconnectedness, Internet Of Things, Cyber physical systems and Data Management are one among the many key features which play a vital role in Industry 4.0 and digitalization to be implemented successfully [4]. Automation and Big Data complement each other. Big Data is responsible to help and get useful information from the massive raw data in real time using appropriate tools and technologies, thus supporting automation. Automation refers to sensible implications of automated service providing, communication, production etc. commonly at production level which boosts the efficiency.[4]

Smart Maintenance is a concept for maintenance in digitalized manufacturing that is defined as "an organizational design for managing maintenance of manufacturing plants in environments with pervasive digital technologies"[9]. It is a research area intending to determine the current state of the organization and also specify what is the goal that is needed to be achieved in the domain of manufacturing in digitalized manufacturing systems. 'Data driven decision making', 'Human capital resource' 'integral integration' and' external integration' are the four important pillars of smart maintenance commonly defined as the four dimensions. These dimensions focuses on evaluation of smart maintenance within and across organizations [1]. Speaking of 'data driven decision making as a strong dimension to asses smart maintenance, data driven approach for predictive maintenance can be seen as an indispensable recent development playing a crucial role in making digitalized manufacturing systems smart in the domain of maintenance.

2.2.2 Data Driven Approach for Predictive Maintenance

Predictive maintenance comprises a set of maintenance activities which are initiated on the basis of changes sensed in the physical condition of the equipment thus, throttling the working life of the equipment and at the same time reducing the risk of failure [7]. The two popular kinds of predictive maintenance are:

- Condition based predictive maintenance:
  The condition of the equipment is monitored with the help of sensors continuously or periodically and threshold values and indications are defined. The condition of the equipment is the basis for the maintenance activities to be triggered if necessary.
- Statistical based predictive maintenance:
  Pre historic data recorded in the database which gives insight about the previous stoppages along with the patterns examined which was made available from the sensors is collected and examined to find trends, calculate the RUL(remaining useful life) and predict failure with the help of machine learning models[7].

Some of the major technological advances which helped to pave the way for the fourth industrial revolution include autonomous robots, simulation, horizontal and vertical system integration, the industrial Internet of Things, cybersecurity, the cloud, additive manufacturing, augmented reality, and big data and analytics[10]. Data-driven decision-making is an important precondition within industrial digitalization. The integration of IT and communication technology with the manufacturing process has led to the acclivity of smart manufacturing, which is reforming the traditional manufacturing industry. In consequence, tons of machine data are generated which can be used to make data-driven decision-making using computerized algorithms. The data can be studied and the information recovered from the data can be used to predict the degradation of components, the current state of the system or its remaining useful life.[11] According to one of the survey [11], data driven models can be classified into statistical models, stochastic models and machine learning models [11]. Statistical models aim to analyse the behaviour of the random variables based

on a recorded data. For the purpose of predictive maintenance, statistical models are very handy to determine the current rate of degradation and RUL(remaining useful life) accomplished by comparing the current behaviour of measured random variables against the known behaviours represented by a series of data [11]. Stochastic models aim to observe the behaviour of random variables over time. Stochastic processes are the fundamentals of the stochastic models. Though these models require high computational power requirements and advanced mathematical knowledge to be implemented they are quite popular because of their capabilities to handle regression suitable for degradation modelling and RUL estimation [11].
The techniques used in machine learning can be broadly classified as in 2.1

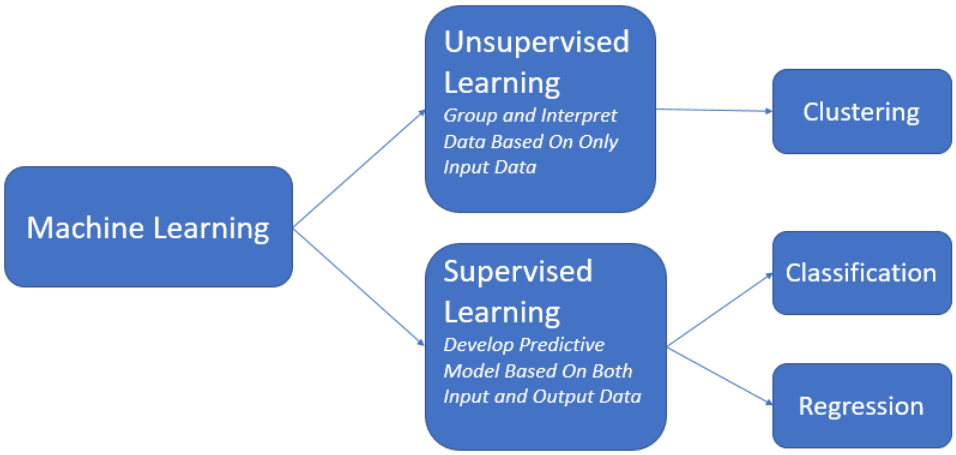

Figure 2.1: Different types of machine learning solutions

Machine learning is an important approach for predictive maintenance as it is very handy for anomaly detection. Machine learning models learn the normal behaviour of the system by monitoring the data generated by the IoT sensors over in real-time. Further, these models automatically detect the unusual activities, hidden patterns, and correlations which are useful to provide precautionary recommendations. Interesting thing about machine learning models is that they dynamically adjust to the new set of input data without the need for manual configuration or threshold settings that the traditional maintenance practices demand.

2.2.2.1 Machine learning Solutions

This section discusses the wide area in which the machine learning solutions used in this thesis belong to as well as giving a brief introduction of the same.

Two framework which captures the wide range of machine learning solutions are:

- Descriptive, diagnostic, predictive and prescriptive:
  It is a frameworks useful to begin the machine learning approach with a concrete understanding of the problem statement. Following this framework, a systematic approach towards understanding the pattern of historic failures, diagnose the root cause of the failures, predict the future occurrence of the

same and prescribe steps to tackle the problem in the future.

- Classification, Regression, Clustering and Association:
  Supervised learning: It is a type of machine learning where the expected outcomes are known and the model is trained on a labelled data. The output error is compared with the expected labels in the test dataset. The training of the data set is done with an iterative approach until the accepted level of error is reached [11]. Classification models: are one among the popular types of supervised modelling which try to draw some conclusions from the observed values. Based on the set of inputs these models try to predict the labels which can be binary classification or multi-class classification [12].

RUSboosted Model: The problem of class imbalance is quite common in the real world datasets resulting in the skewed training data posing a great challenge to construct good models. Data sampling and Boosting are two common techniques to tackle the above stated problem of class imbalance. Data sampling tries to balance the class distribution by either adding the minority class (oversampling) or by deleting some of the data entries belonging to the majority class(Undersampling).Boosting is a technique which improves the performance of a weak classifier. AdaBoost is the most common boosting algorithm which iteratively builds an ensemble of models. The weights of the classification examples which were wrongly classified in the current iteration are modified so that they are classified correctly in the next iteration. The RUSboost model is a hybrid combination of data sampling and boosting algorithm. It combines Random Undersampling and AdaBoost algorithm. This model is simple and has an advantage of less computing time and better classification performance compared to the hybrid of oversampling and AdaBoost algorithm called SMOTEBoost [13].

Regression analysis: Under the category of supervised learning, Regression analysis is an algorithm used to learn a mapping function from the input variables to the continuous output variables. The primary focus of such a learning is to approximate the learning function as accurately as possible such that a output variable for the data set can be predicted for every instance of a new input variable in the data set [14].

Unsupervised learning: This technique is used when the preliminary outcomes are not known (unlabelled data)[11]. The model is made to work on its own and discover patterns or information that was previously undetected. Clustering comes under unsupervised learning which is a simple task of grouping the population into a number of clusters where the data points in a particular cluster are similar to each other and dissimilar to the data points from the other clusters[15]. Association analysis is also an important type of unsupervised machine learning where the algorithm focuses on establishing relationships between different entities in an unlabelled dataset. These relationships are represented as association rules and play a vital role in extracting hidden information and are widely used in transactional data or relational databases.[16]

Gaussian Mixture Models (GMMs): Unlike K-means clustering models which are distance-based models and create clusters in circular shape, GMMs are distribution based models which assume that the data set has a certain number of Gaussian distributions (normal distributions: A bell shaped curve with data points distributed symmetrically around the mean) equivalent to the number of clusters specified. GMMs are probabilistic models that use a soft clustering approach for distributing the data points into different clusters [17]. They are of great use when the distribution of data points are not in a circular form.

2.3 Data Quality

The ongoing trend of the development of IoT and Big Data enhances the complexity of data. And with the large data architectures and their growth it can easily cause issues with data quality [18]. When modeling failure predictions, data quality problems are one of the major sources of poor accuracy for the predictions since the data might not represent the reality [19]. Another problem with data quality is as [20] mentions, that data is seen as of high quality only if it meets the needs of the user of the data, which implies that data quality is inherently subjective. This leads to a problem where different users can rate the quality of a data set differently depending on the needs. The Hierarchical Data Quality assessment framework [19] is a more objective framework to determine data quality which consists of widely accepted quality dimensions. These dimensions have several sub-elements followed by indicators. The dimensions of data quality are assessed and based on these arguments a score; either 1 or 0 is awarded. 1 representing the corresponding data quality element of the dataset is above satisfactory conditions and 0 representing below. See A.1 for the framework of the arguments.

2.4 Descriptive Statistical Features

For describing the characteristics of variables in a specific dataset statistical features are used that are based on the properties of the variables. These features provide information about the variables, and can also provide potential relationships between the variables [21].

2.5 Evaluation Metric for algorithm modeling

2.5.1 Confusion Matrix

When evaluating predictive models and the predicted classes a confusion matrix gives a good overview of the model performance whether it is predicting well or poorly. This matrix has two dimensions where one dimension is indexed by the actual classes, and the other is indexed by the class that the model predicts. [22] By labeling the known classes with e.g True and False or binary numbers, the confusion matrix will be a 2x2 matrix as shown below in figure 2.2.

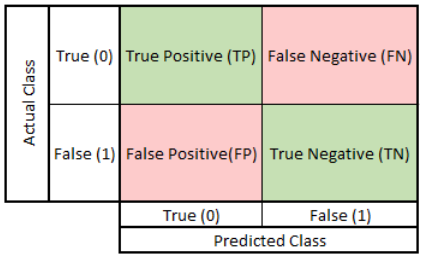

Figure 2.2: Simplified visualization of a confusion matrix

If the observed class is True and the model predicts it as True it will result as a true positive (TP) observation, if the model predicts it as false it will result as a false negative (FN). If the observed class is false and the model predicts it as false it results as a true negative (TN), and if it predicts it as true it will result in a false positive (FP). Based on the confusion matrix four different metrics can be calculated as shown by equations 2.1-2.4

$$Accuracy = \frac{TP + TN}{Total} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4}$$

Accuracy gives a performance measure that shows the ratio of correctly predicted observations against the total observations. The accuracy is a good measure if the classes are symmetrically divided in the data set. Precision gives the ratio of the correctly predicted positives against the total predictive positives. Recall shows the ability of the model to predict positive observations in the observed class. F1 Score gives the weighted mean of recall and precision, which then include both false positives and false negatives. The F1 score is useful when the distribution of the classes are uneven and if the impact of them are different.

2.5.2 Silhouette Score

For unsupervised clustering techniques each cluster can be represented with a silhouette which is based on how closely linked the clusters are. The silhouette coefficient will show how well a certain sample will belong to its assigned cluster compared to other clusters [23]. The coefficient ranges between -1 to +1, where a score off +1 means that the sample well belongs to its assigned cluster, 0 means the sample is in between two clusters or might belong to another cluster, and -1 that the sample is in the wrong cluster. Based on the silhouette coefficients for all the samples in a dataset a silhouette score can be determined by calculating the mean value of all the silhouette coefficients [24].

2.6 Double-acting Cylinders

Double-acting cylinders, or actuators, are commonly used in automated industries. The force from the pneumatic connection is performed in both extending and retracting direction, moving the piston dependent on which port that receives the compressed air. It can have two standard positions, either in an extended position when the pressure is applied on the positive chamber *Ps* or in a retracted position when the pressure is applied on the negative chamber *Pe* as shown in figure 2.3

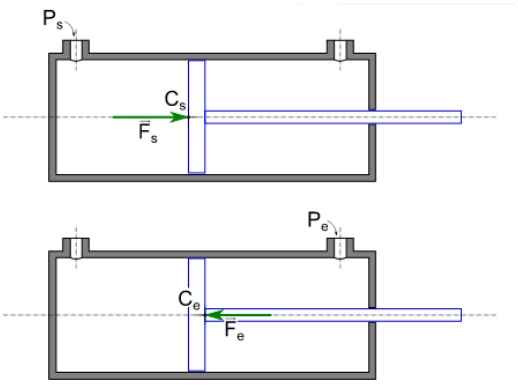

Figure 2.3: Double acting cylinder, [25], CC-BY

2.7 Related work

Previously, there were several studies [26][27][28][29] carried out pertaining to the data driven fault detection in the pneumatic systems. Fault detection which is data driven is a powerful approach to detect pneumatic leakage that is growing popular in predictive maintenance problems. Sticking to the basic definition of a fault or an anomaly; it is an event or an occurrence which deviates from the normal and expected trend. Studying the normal behaviour of pneumatic systems is of utmost priority, doing which there can be a platform to differentiate faulty behaviour from what is studied to be normal. In [26], although it does not actually connect the data driven-decision making with pneumatic leakage, it consists of a concise summary of literature analysis which highlights the recent increase in the emergence of data-driven decision making methods developed to exploit the abundance of sensor-generated data along with the emergence of cyber-physical systems and cloud technologies to process and store data, decision making for maintenance in the future will be more responsive and capable of providing accurate and proactive decisions. In [27] a signal based approach with wavelet method, an analytical faults detection and diagnosis for pneumatic systems is described. Multiresolution wavelet decomposition of sensor signals(pressure, flow rate etc.,) is used to determine leakage configuration. With the help of affine mapping, pattern recognition technique and analytical vectorized maps are developed to diagnose an unknown leakage based on the established faults detection and diagnosis information. The fault detection techniques quoted in this literature can be put to use extensively to reduce unneces-

sary downtimes thus increasing productivity. Literature [28] describes a data-driven predictive maintenance approach to detect anomalies on an Air Production Unit (APU). The key takeaway from this literature would be the concept analysing the sensor data. The paper considers frequency of peaks a reliable method for anomaly detection, which can be further used as anomaly indicators setting some rules about the same. The future scope of this literature where smart alarms will be used to notify the maintenance personnel in case of a fault detected is interesting. Literature paper [29] discusses application of unsupervised learning methods to detect developing air leaks in critical components (pipelines or air actuated components) of mobile mining machinery. Machine learning approach is used to associate patterns in pressure drop from the accumulator with the activation of each air-powered component. Wavelet transform is applied to the accumulator pressure trend and most informative wavelet scales are selected. Two anomaly detection methods(Local Outlier Factor and Autoencoder) are trialed and compared to find out the method having the best evaluation metrics. The major takeaway from this literature was the concept of comparing models with the evaluation metrics to reason the trade offs and select the most suitable one.
Using machine learning approaches for data driven predictive maintenance might sound the optimum solution with guaranteed results, but challenges and gaps in this area are quite considerable in number. Unsupervised learning approaches are common and it's a challenge to label the data especially if information regarding historic anomalies are limited. Moreover, data analytics being the core element of data driven decision making, it's always a challenge to translate the insights obtained from the data analytics into business actions for the companies and obtain competitive advantage from it, thus highlighting a sizable gap between the production and consumption of analytics[30]. Given the aforementioned literature, in most of the pneumatic leakage detection cases, abundance of sensor-generated data in the context of Industry 4.0. is a prerequisite. Sensor-generated data are examined to set rules that define the normality which inturn helps to define anomalies. Given the bulkiness of the sensor-generated data and their importance in the scenario of pneumatic leakage detection, it's a challenge to handle them appropriately to obtain fruitful results.
This section of related work provided a good understanding and helped to structure the approach of the presented thesis work intending to find a data-driven detection method that can predict future failures on the pneumatic system by detecting an early stage of a leakage and further to provide a standardised methodology for any kind of future work related to predictive maintenance on pneumatic leakages.

3
Methodology

In this chapter the methodology of this thesis work is presented. This chapter aims to give the reader an structured overview and understanding of the work procedure for this project, how the analyses and experiments were developed and how tools and processes were applied. This thesis and the aim for the project is heavily linked with data mining and data science, hence the use of the Cross-industry standard process for data mining (CRISP-DM) methodology which will be explained further in this chapter.

3.1 CRISP-DM

CRISP-DM is a standardized process for data mining and is a common method for machine learning projects [31] and was created since, at the time, there weren't any standardized framework for data mining projects [32]. The method is an iterative process of data mining and consists of six top level phases as seen in figure 3.1. Each of these phases has second level generic tasks, covering all the necessary data mining situations.

Figure 3.1: General methodology of CRISP-DM [31]

3.1.1 Business Understanding

In this initial step of the process the goal, scope and business objectives were discovered. This was made by discussions with the stakeholders of this project at the

start of this project. These stakeholders are:

- SKF Maintenance Engineer and domain expert
- SKF Reliability Manager
- PACA Project Leader and Supervisor

The discussion lead to the aim for this project and a mature problem definition for the goal of this thesis, as mentioned in chapter 1.2 Data mining objective: The process of data mining is tasked to find anomalies, patterns and correlations of the extracted data using various techniques that detect leakage.

3.1.2 Data Understanding

Data understanding is an important phase of this framework. Some kind of initial preliminary data understanding was necessary for the definition of the thesis problem. Later on, full fledged data understanding started with Data Collection followed by a set of steps and activities focused on:[32]

- Getting familiarity with the data.
- Identifying data quality problems if any.
- Discovering initial preliminary insights into the data.
- Detecting interesting subsets to form hypotheses for hidden information.

To further understand the data and the general behaviour of the processes two experiments were conducted. The opportunity for the first experiment was given during a planned maintenance stop in other processes in the production line. Based on the results from the first experiment, a second experiment was conducted as well. Chapter 3.1.2.1 and chapter 3.1.2.2 will explain the purpose and methodology of the experiments.

3.1.2.1 Experiment 1

During the thesis work, there was an opportunity to do an experiment in the cell of the packaging process during a planned maintenance on other equipment in the same production line. The maintenance was planned for eight hours during the day. First a small hole was drilled in an external plug with a 1 mm drill (see figure 3.2). The plug was then connected to a pneumatic tee tube-to-tube adapter (see figure 3.3), simulating a leakage of a tiny hole in the tubing.

Figure 3.2: Drilled hole of 1 mm inside the plug.

Figure 3.3: Tee tube-to-tube adapter with plug and a small drilled hole.

The installation and connections were all made by an experienced maintenance engineer from SKF. Due to the planned maintenance break the whole process of the machine could not be run, just one conveyor which was a part of the packaging process consisting a total of four cylinders was run in the process. A cylinder was chosen that had a retracted standard position, meaning that if the machine was in an idle state and the leakage occurs on the minus side, there was a constant leakage even though the machine was idle. On the contrary if the leakage was on the plus side of the cylinder, the leakage only occurred while extended.

During the experiment no cartons or bearings were loaded into the machine, and the glue guns were disconnected. The experiment consisted of four different phases. The first phase there was no hole in the tubing acting as the normal state of this specialized process. In the second phase, the adapter was connected to the plus side of the cylinder. In the third phase the adapter is connected to the minus side of the cylinder, making it leak constantly. The fourth phase, the plug is removed, making a large hole of 4mm in the tubing. See figure 3.4 for the different phases of the experiment visualized in Grafana.

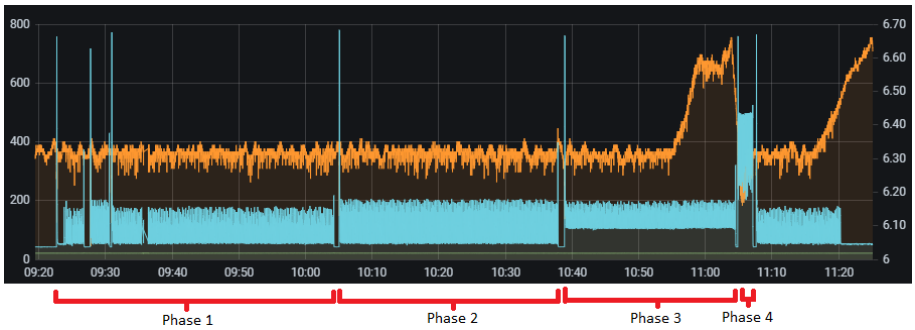

Figure 3.4: Visualisation of time series for packaging machine using Grafana

The start and end time of the experiment along with the comments was noted for each phase, See Appendix B.1 for the times and notes for each change. The purpose of this experiment was to induce artificial anomalies in the machine of the running cycle to observe and document the behaviour of the sensor values corresponding to the 'Normal machine Working' or 'Machine leakage' states of the machine. The purpose of this experiment is intended to head towards the direction of labeling the data.

3.1.2.2 Experiment 2

After reviewing the result from experiment 1 and how the machine behaved during this process, as can be read in chapter 4, a second experiment was conducted on the equipment. The purpose of this experiment was to record the data corresponding to 'Machine Leakage' status for a longer period of time, thus providing more data points for the 'Machine Leakage' status. Since experiment 1 was conducted during a maintenance break and just one conveyor was running, experiment 2 was conducted in the packaging process on a normal running cycle, on a normal production day with the packaging process happening normally.
Note:The actuator cylinder worked normally in both experiments though there was a leakage. Meaning, The leakage was so small that the normal working condition of the actuator cylinder was not hindered. The ambition here was to identify this minute leakage using data-driven machine learning models thus giving an opportunity to identify leakage at its earliest stage and perform the necessary steps in advance before the leakage turns into a severe one and stops the normal production process. The same type of plug and pneumatic tee tube-to-tube adapter as in experiment 1 was used. This was connected to the plus side of a cylinder with an extracted standard position, meaning there was a constant leakage even when the machine was idle. A small hole was drilled in the plug with a 0.8 mm drill in Chalmers Prototype Workshop, resembling an early stage of leaking in the tubing. The Maintenance Engineer Thomas Sundquist installed the connection the 27th of April at 08:30 in the machine, and removed the connection the same day at 15:26 from the process.

3.1.3 Data Preparation

It is very common that the process of cleaning and preparing the data has a lion's share of 80% of time and effort spent during data analysis[33]. The data preparation

phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are an iterative process. Figure 3.5 represents The workflow for data preparation that is followed.

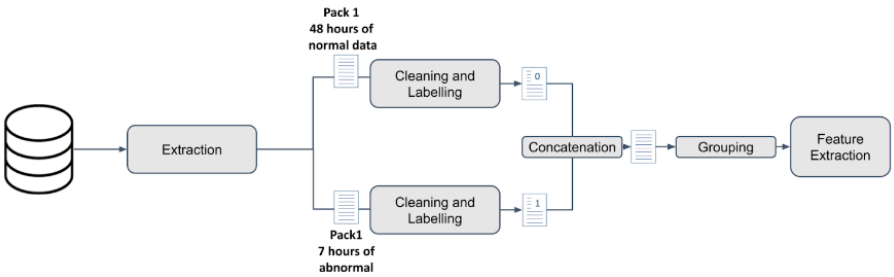

Figure 3.5: Workflow of data preparation

Tidy datasets are successful in providing a standardised way of linking the structure of the data set with its semantics[34]. A major share of statistical datasets are rectangular tables consisting of rows and columns as the basic elements of the data structure[34], A data set is a collection of either quantitative values(numbers) or qualitative values(strings)[34]. Each and every value in a dataset belongs to a variable and an observation[34]. A variable consists of all the values that measure the same attribute(pressure, temperature and airflow). An observation consists of all the values measured on the same unit(an observation is made every second in the data set that is used for this thesis work)[34]. An experiment design throws light on the structure of the observation made. In the given scenario, every combination of sensor values(pressure, temperature and airflow) along the dimension of time is measured[34]. Experimental design assists to make decisions on missing value treatment, whether or not not it could be dropped and what could be after effects of dropping the missing values[34]. In the dataset being used for the thesis work there isn't any missing values.
Tidy data is a standardized method to link the meaning of the dataset with the structure of the dataset [34]. Below are the five most common scenarios which represent a messy data:[34]

- Column headers are seen to be values instead of variable names
- A single column having multiple variables stored in it.
- Both the rows and columns contain stored variables.
- The same table contains multiple types of observational units.
- A single observational unit is stored in multiple tables.

The presented dataset in the thesis work follows Codd's 3rd normal form for relational databases [35]. According to which it follows three simple rules of data management, they are: [34]

- Each variable forms a column.
- Each row comprises an observation.

- Each type of observational unit forms a table.

3.1.4 Modeling

The 'Modelling' section is a systematic search for a model which meets the business objectives as defined earlier, which is efficient in predicting with minimum error and is most suitable to be used considering the properties and characteristics of the given data.
The final dataset with the significant features are partitioned into one set of training data and one set of validation data with the command 'cvpartition' in Matlab. This feature randomly partitions the observation into two separate sets based on the 'Label' class information. This will result that both the training and validation sets will have roughly the same proportion of Label class information. The size of the set is divided to 80% of training data and 20% of validation data. The train data is inserted to Matlab's Classification Learner app, with the response value set to the labels in the data set. Considering the characteristics of the dataset, it is very important that the right model is chosen. Imbalanced dataset is an important characteristic that is to be considered during the process of model selection. The group labels do not have equal data entries. 8.6% of the data after cleaning belongs to group '1' representing the " Air Leakage" state of the machine and 91.4% of the data set after cleaning belongs to group '0' representing "Normal" state of the machine. This is a strong argument which supports usage of 'accuracy' alone would not be sufficient to evaluate the performance of the model. Other evaluation metrics such as Precision, recall and F1 score are considered for the evaluation of performance.
The systematic search for the optimum model follows the work flow as mentioned in the figure 3.6.

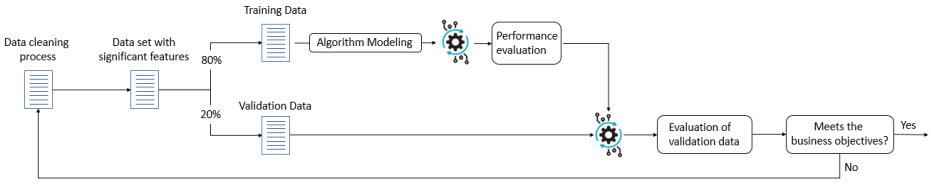

Figure 3.6: Workflow of Algorithm Modeling

The classification learner app provides an opportunity to train all the available models and check for the best performance. As mentioned before, the models' performance was evaluated using the confusion matrix which was available as an in-built feature in classification learner app.

3.1.5 Deployment

Building of the model is not the end of this framework. The model should be presented in an organized way to the company so that they make use of it[32].The

Deployment phase is out of the scope of this thesis mainly considering the limited time to work on the thesis. However, recommendations are made to SKF regarding deployment and possible best ways by which the model could be utilized. These recommendations are stated in Chapter 6: Conclusions and recommendations.

3.2 Unsupervised Modeling Approach

According to the Maintenance logging system an error was reported on 09/12/2020 for another machine called Wrapper Machine, where a leaking tube was the cause of the error. This error was fixed on 01/18/2021 16:00(see figure 3.7). An unsupervised approach was conducted to find clusters which could capture this scenario successfully. All the methodology including data extraction, data visualization, data understanding and data cleaning were the same as described in section 3.1. Although there were few changes: Data labelling was not necessary as it was unsupervised learning In the idle time removal stage, the same logic as for the idle time removal in packaging process was used except that the upper threshold for the sensitive feature(Airflow) was set to 45 l/m for 15 seconds time duration. (If airflow was less than 45 l/m for continuous 15 seconds, it was removed considering it to be idle time, this threshold was set by observing the variations in the features with respect to time in Grafana)

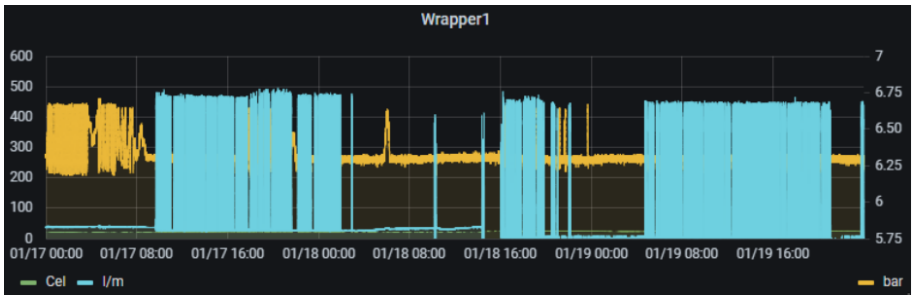

Figure 3.7: Visualization of features during the leakage and after it was fixed in Wrapper

Seven features which were defined as the significant features based on the Kruskal-Wallis ranking system (See Figure 4.12: Ranking of the features based on Kruskal-Wallis test) for the packaging machine were used as significant features for the wrapping machine dataset as well. The seven significant features were: Mean, RMS and peak value of temperature and mean, RMS, peak value and shape factor of airflow. For modeling, the Python module Scikit-learn was used. This is a module with a wide range of machine learning algorithms for both supervised and unsupervised problems [36]. Gaussian Mixture model is used as the classification model to form two distinct clusters for the machine 'Normal' and 'Leaking' state. GMMs is used as it has the ability to capture data points which are not spread in circular form. They are advantageous in the given scenario as they are distributional models which use a soft clustering approach to cluster the data points into their respective clusters.

4

Results

This chapter will shed some light on the information gained by following the CRISP-DM methodology and will present the results for each step accordingly. Adding on, the overall results from the experiments and validation will be presented for both an unsupervised and supervised machine learning model.

4.1 Data Understanding

Data has been collected from the two machines since July 2020 continuously with two separate sensors, from IFM model SD6500. The sensors are installed on the pneumatic pipes that execute the glueing process, measures the pressure (bar), airflow(l/m) and temperature of the compressed air (C°). Data is collected once every second, and sent as an analogue signal to an IO-Link master, converting the signal to a digital output. The master is an IFM I/O Link that is connected to the TCP/IP network and sends all the raw data to a database where the data is stored. The database collects sensor data from the whole production line, and contains information from a large number of sensors from different processes.

In order to get hands on the data from the SKF database, access to the EAA client platform was issued. One way to extract the data was with a scripted text file including various SQL commands that was handed by the SKF IT-department. A couple of functions could be edited to change the number of samples to extract, and to extract data from a different process. The resulting dataset using this method has five columns showing: Device name, machine name, Timestamp, unit of measurement, and lastly the measured value. The timestamp is of the format YYYY-MM-DD hh:mm:ss.SSSSSS+00, where S is equal to a fraction of a second. For this case, the first two columns will have the exact same content on every row. Each time stamp is saved on three consecutive rows where the measured pressure, airflow and temperature are saved separately as shown in figure 4.1.

	A	B	C	D	E
1	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:44.836993+00	bar	6.3
2	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:44.836993+00	Cel	21.8
3	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:44.836993+00	l/m	30.17
4	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:43.677469+00	bar	6.31
5	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:43.677469+00	Cel	21.7
6	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:43.677469+00	l/m	27.01
7	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:42.60766+00	bar	6.31
8	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:42.60766+00	Cel	21.8
9	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:42.60766+00	l/m	30.17
10	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:41.53778+00	bar	6.3
11	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:41.53778+00	Cel	21.8
12	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:41.53778+00	l/m	28.51
13	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:40.467754+00	bar	6.29
14	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:40.467754+00	Cel	21.8
15	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:40.467754+00	l/m	29.34
16	LCHO Pack 1 Airflow[SD6500-Port 1]	D3_LCHO_PACK1	2021-03-11 01:10:39.397004+00	bar	6.3

Figure 4.1: Extraction of data using SQL

The other approach was to extract the data from Grafana where the data ranges between specified time intervals for a single process. The resulting data set for this method has four columns: Time, temperature, pressure and airflow where the timestamp is of the format hh:mm:ss. Each row contains the three features for each sample, see figure 4.2.

	A	B	C	D
1	Time	Cel	bar	lm
2	2021-03-30 14:54:16	22	6.31	32.2
3	2021-03-30 14:54:17	23.2	6.28	393
4	2021-03-30 14:54:18	22.9	6.32	80.4
5	2021-03-30 14:54:19	22.6	6.32	31.2
6	2021-03-30 14:54:20	22.3	6.32	23.7
7	2021-03-30 14:54:21	22.1	6.32	22.3
8	2021-03-30 14:54:22	22.1	6.32	22.5
9	2021-03-30 14:54:23	22.1	6.32	22.7

Figure 4.2: Extraction of data using Grafana

For both of the approaches the extracted data resulted in a csv file. Data set extracted from Grafana differed from the SQL data extraction method in a few ways which is considered to be trivial. The data values for airflow (l/m) had two decimal values for the data set extracted via SQL and just one decimal value for the data extracted using Grafana. for the other two features, the data is exactly the same. Also, the time stamp had a different format. The time stamp when extracting with SQL commands results in the database time zone, GMT, whilst when extracting from Grafana results in the local computer's time zone. It was decided to extract the data with Grafana since xxxx

4.1.1 Data Description and Visualisation

This phase includes various activities to get a deeper understanding and familiarity with the data using visualization techniques. Figure 4.3 visualizes the behavior of the process Pack 1 over 45 minutes. The blue line shows the airflow (l/m) with values from the left y-axis, the green the temperature(C°) and the orange shows pressure (bar) with the values of the right y-axis. Observing the graph and the

behaviour of the airflow it is quite easy to distinguish the state of the machine as the pattern of l/m reading in Grafana shows a cyclical behavior when the machine is working, and a flat non cyclical behaviour when the machine is idle.

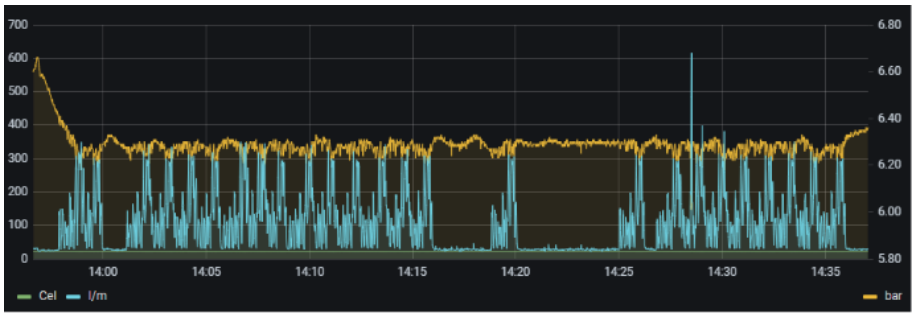

Figure 4.3: Visualized data of pack machine with grafana

As a first progressive step soon after the extraction of data from Grafana, it was visualized using Tableau Prep Builder (see figure 4.4). This gave an opportunity to obtain valuable insights of the data range in which the maximum data points for various features existed. It was also an initial step to know what the outliers could look like in the data set. Examining the visualizations, it was learned that 93% of celcius data entries were in the range 22-23 CC°. 60% of data entries corresponding to pressure spread in the range 6.2-6.3 bar. 99% of data entries corresponding to airflow spread within the range of 0-400 liters per minute.

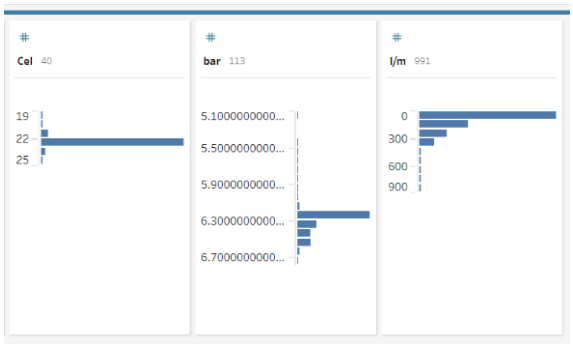

Figure 4.4: Distribution of raw data of the three variables

Using the function 'corrcoef' in matlab the correlation between the three variables was plotted, as seen in figure 4.5. There are both negative and positive correlation between the variables, but since the abs(correlation) < 0.8 it's not considered to be strong. [37]

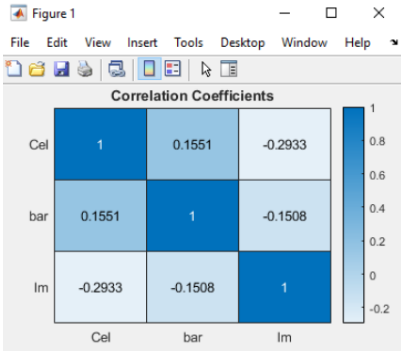

Figure 4.5: Correlation coefficient of the variables

4.1.2 Experiment 1

The prime takeaway from the experiment was that a 'sensitive feature' could be defined which visually helps to identify the different states of the machine. The details of the 'sensitive feature' is explained in section 3.1.3(data preparation).

4.1.3 Experiment 2

The outcome of experiment 2 gave an opportunity to label data representing induced anomalies during the normal working cycle in a normal production day.
The data gathered during the experiment 2 was visualized using Tableau prep builder as shown in the Figure 4.6 to know that it followed the similar distribution trend of the 'normal working state' data(Figure 4.4: Distribution of raw data of the three variables) thus stressing on the need for machine learning approaches to differentiate the two machine states.

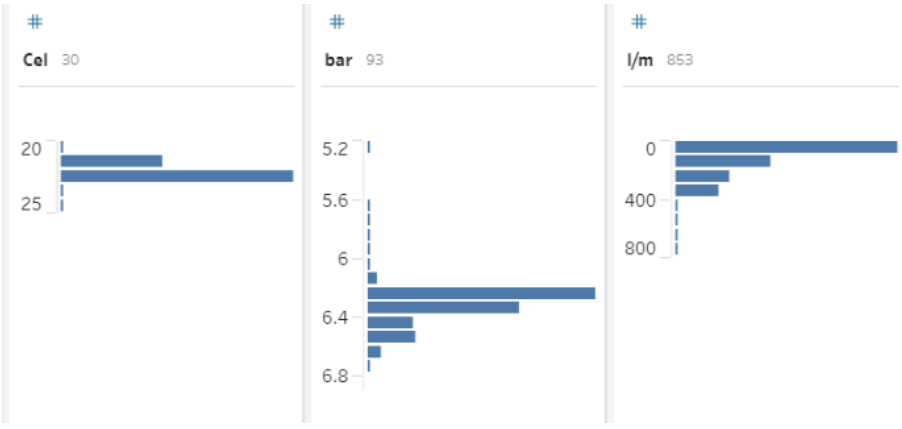

Figure 4.6: Visualization of data corresponding to leakage gathered during experiment 2

4.1.4 Data Quality

A data quality report based on the framework from chapter 2.3 has been made on the extracted datasets from Grafana, see Appendix C.1 for the result. Ten elements

belonging to five dimensions of data quality were assessed. All the elements are awarded 1 point indicating that they were above satisfactory level. The quality of data that is used in this thesis work is outstanding with a score of 10 points. The comment column gives the reasoning based on the data quality indicators for the allotted score.

4.2 Data Preparation

With help of knowledge and framework of Tidy data stated in section 3.1.3 , The dataset presented in the thesis work was tidied as previously shown in figure 4.2
Data cleaning: Data cleaning is an important step in data preparation. The data quality report gives a good insight regarding the anomalies of the current dataset. As stated before, the dataset does not have any missing values but contains outliers. The procedure of outlier detection and removal of the same is explained in the below section.
Sensitive feature: In the context of the given problem statement within the boundaries of the presented thesis work, The feature Airflow with the unit liters per minute is defined as a sensitive feature. During the factory visits for conduction of the experiments, with help of repeated observations in Grafana during the experiments, it was established that the feature airflow showed visual and clear variations which were sensitive to the different status(Idle, Normal, Leaking) of the machine. The other features(pressure and temperature) were not logically sensitive to different status of the machine. With repetitive observations of the sensor values in Grafana during numerous factory visits, the graphical trend of the sensor values of airflow (unit: liters per minute) corresponding to different status(Idle, Normal, Leaking) of the machine were learned. The details pertaining to the trend of this sensitive feature with respect to different status of the machine is discussed in the section 3.1.2.1 Experiment 1.
Outlier Detection and Removal: A Box plot is plotted on sensor values of the sensitive feature (airflow) with an intention to detect the outliers. During experiment 1, it was understood that the sensor values of the sensitive feature corresponding to the status of the machine (Idle, Normal and Leaking) lie in an approximate range of 0 to 400 liters per minute. This observation is statistically strengthened with the box plot presented in the figure 4.7. The outliers consisted of high airflow values corresponding to the security door openings. These security doors once opened, seizes the pneumatics and the process comes to halt. It is very important to know that all the airflow values corresponding to door openings are outliers, but not all the outliers are airflow values corresponding to door openings. Further on, in the modelling section, as classification approach is adopted, the presence of these logically unexplained values of the outliers, in both the labelled groups of 'Normal data' and 'leakage data' might lead to wrong classification of the data points during the prediction. They also pose a threat of adversely affecting the descriptive statistical measures. As shown in the figure 4.7, There are no outliers in the bottom and the outliers in the top comprises a small portion of 1.66% of the total data points. The whole dataset is of high resolution, that is, data points are recorded every second. Closely observing the arguments stated above regarding outliers' role in misclassi-

fication, given the fact of availability of abundance of data points and the outliers comprising a mere small portion of 1.66%, 'Removal of outliers' outliers treatment method was adopted.

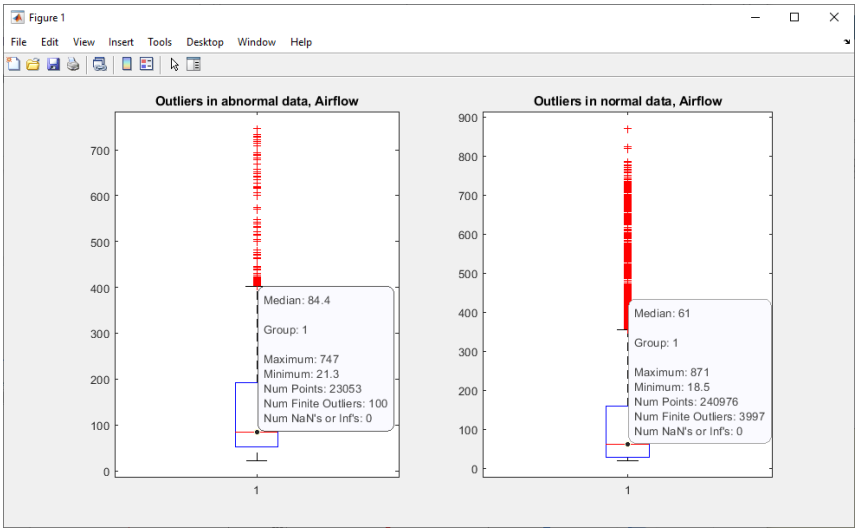

Figure 4.7: Box plot for outlier detection

Figure 4.8: Comparison of airflow before and after idle time was removed for normal data.

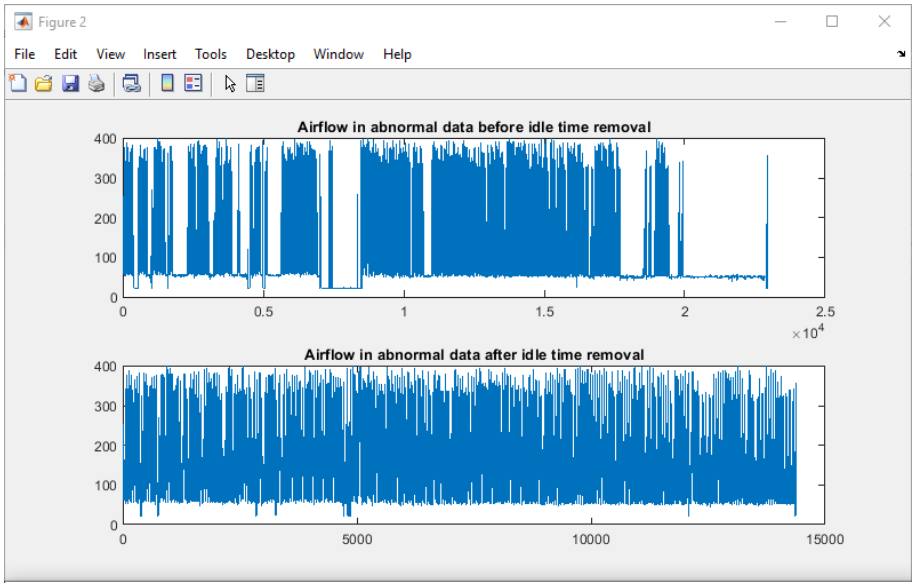

Figure 4.9: Comparison of airflow before and after idle time was removed for abnormal data.

Idle time hinders the trend of the data as it is completely dependent on the demand for production determining when the machine is kept idle or running. Removal of Idle time is necessary because the scope of this thesis work is to alarm pneumatic leakage during the working condition of the machine. The role of behaviour of the sensitive feature in establishing standard range for idle time is immense. During conducting experiments, the machine was kept at idle state to learn the behaviour of the sensitive feature. With repeated and careful observations in Grafana it was noted that all the airflow values when the machine was idle were below 45 liters per minute for 'normal' dataset. Again, it is important to note that, all the airflow values corresponding to idle time were below 45 liters per minute, but not all the airflow values below 45 liters per minute correspond to idle time. To tackle this scenario, an algorithm is built, which checks the values of airflow in a 20 seconds time window. Inside the window, every value is checked one after the other and if all the values are below 45 liters per minute for continuous 20 readings, then the whole window is removed considering it to be idle time. (Note: The time window of 20 seconds was chosen and the threshold of 45 l/m is set by careful and repeated observation of the behaviour of sensitive feature during machine idle time in Grafana during the onsite company visit.) The logic was same for the dataset corresponding to 'leakage' except that the threshold was set to 65 l/m (Note: The time window of 20 seconds was chosen and the threshold of 65 l/m is set by careful and repeated observation of the behaviour of sensitive feature during machine idle time in Grafana with an intention to capture all the idle time correctly.)

Grouping of Data: The count of data points per minute sums up to less than 60 after removal of idle time data points of the machine. As a result, regrouping of the data points is necessary. This is done on a newly defined time dimension called the "Processing time". Processing time is the time which represents only the working status of the machine. The grouping is done in such a way that every cell has 60

processing time seconds with a sliding window of 30 processing time seconds. In simple words, every cell has the last 30 seconds of the previous cell and the first 30 seconds of the next cell. The sliding window ensures that the hidden pattern or trend(in any) in the middle of the 60 seconds cell is not lost. The grouping of data is necessary to define the processing time dimension and to make the dataset concise with a better structural meaning. This structure of grouped data is very handy to extract the features using the Diagnostic Feature Designer app in MATLAB.

	1 Cel	2 bar	3 Im	4 Output
1	*60x1 table*	*60x1 table*	*60x1 table*	1
2	*60x1 table*	*60x1 table*	*60x1 table*	1
3	*60x1 table*	*60x1 table*	*60x1 table*	1

Figure 4.10: Grouped Data

Feature extraction Diagnostic Feature Designer app is used to extract the time domain features from the variables of the dataset (Pressure,Temperature and airflow). 39 time domain features were extracted into a feature table. These Time domain features are arranged in descending order with the feature of highest importance at the top. Meaning, The feature which has the highest potential to differentiate the groups (Normal machine working or Machine Leakage) was placed at the top of the feature table. These extracted 39 features are arranged in the feature table based on Kruskal Wallis feature ranking score. Among these 39 time domain features, 33 features were selected. This selection was based on the Kruskal wallis scoring of the time domain features. All the time domain extracted features having a Kruskal wallis score less than 33 are not considered for the modelling purpose to avoid overfitting and more importantly, they are considered to have less (or no) potential to differentiate the groups as the feature rank is lower. The Diagnostic Feature Designer app in matlab provides two ranking systems: 'One-way ANOVA' and 'Kruskal-Wallis'. Among the two, selection of a particular system is to be done considering the characteristics of the dataset. In the given scenario, 'Kruskal-Wallis' system is chosen as the dataset, especially the distribution of the sensitive feature is skewed and cannot be assumed to be normally distributed (see figure 4.11).

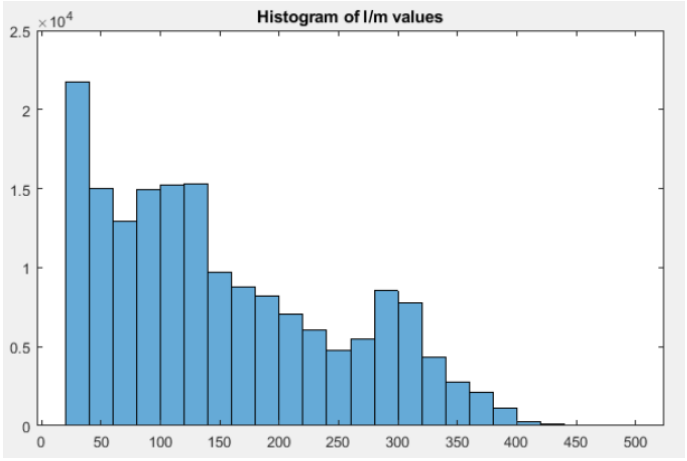

Figure 4.11: Histogram of l/m values

Thus a non parametric approach needs to be adopted which is fulfilled by using Kruskal-Wallis scoring which does not make any distribution assumptions. All features ranked in descending order is seen in figure 4.12

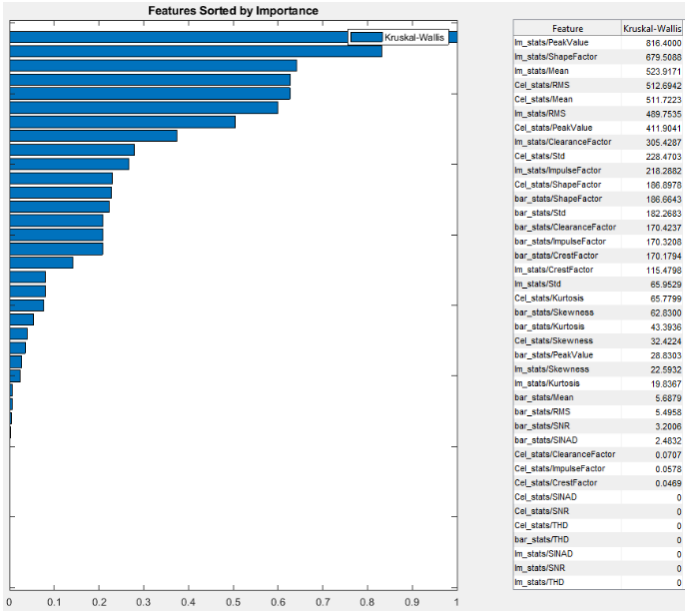

Figure 4.12: Ranking of the features based on Kruskal-wallis test

4.3 Algorithm Modeling and Evaluation

The top two models with good performance that were initially shortlisted were 'Logistic regression' and 'RUSBoosted bagged trees'. Table 4.1: Performance metric comparison shows the comparison of the best performing model with the second best performing model and how both of them are evaluated using the evaluation metrics. These models are trained with all the features having a Kruskal-Wallis score more than 0, 100, 200, and 400 (see figure 4.12 Ranking of the features based on Kruskal-wallis test) separately and the performance (accuracy, precision, recall

Table 4.1: Performance metric comparison

Model	Dataset	Number of Features	Kruskal-Wallis Score	Accuracy	Precision	Recall	F1Score
LR	Training Data(80%)	33	Above 0	0.9829	0.9887	0.9926	0.9906
RUS				0.9867	0.9911	0.9943	0.9927
LR		18	Above 100	0.9763	0.9812	0.9931	0.9871
RUS				0.9795	0.9889	0.9886	0.9887
LR		11	Above 200	0.9801	0.9843	0.994	0.9892
RUS				0.982	0.9894	0.9911	0.9902
LR		7	Above 400	0.9759	0.9782	0.9958	0.9869
RUS				0.9792	0.9872	0.9901	0.9886
RUS	Validation Data(20%)	33	Above 0	0.9873	0.994	0.9921	0.993

and F1 score) were analysed (see table 4.1: Performance metric comparison). The purpose of using a different number of time domain features extracted from the feature designer application was to check and select the optimum number of features which gives the best performing model without overfitting. The performance metric of both the models are compared for different numbers of features which is showcased in the Table 4.1: Performance metric comparison. With the help of this table, optimum number of features and the best model is selected based on the performance metrics. Once the optimum number of features and the model is decided, The test data is run on the trained model to validate the performance of the model with the selected number of features.

From the Table 4.1: Performance metric comparison, It can be concluded that all the 33 extracted features above Kruskal-Wallis Score of 0 should be used to train the RUSboosted bagged trees model. It has an accuracy of 98.672%, precision of 99.112%, a recall of 99.433% and F1 score of 99.2745% when trained with 33 extracted features having a Kriskal-Wallis score above 0. F1 score is an important attribute to decide how the model treats the minority class labels, and having a good F1 score proves that this model has handled the minority class labels well and given good importance to it. Figure 4.13 shows the confusion matrix for the RUSboosted model when 33 features were used. It can be seen that the false positives rate is 7.8% and false negative rate is 0.6%. In simple words, the rate of leakage going undetected and the rate of false alarms is low.

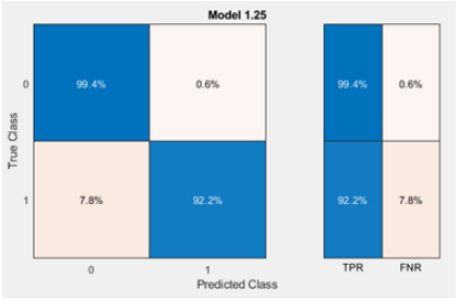

Figure 4.13: Confusion matrix for RUSboosted model with 33 features on training data.

The RUSboosted bagged trees model with the selected number of features (33 features) is trained and the test data set which was reserved initially (20% of the cleaned data set) for the purpose of validation is used on this model to perform predictions. The confusion matrix for the test data is shown in figure 4.14. The validation procedure resulted in 98.73% accuracy, 99.40% precision, recall of 99.21% and F1 score of 99.30%.

Figure 4.14: Confusion matrix for the validation data on RUSboosted model

4.4 Validation of the Model

In order to double check the working of the model and record its performance of leakage detection on another cylinder of the same packaging machine set up, the plug with 0.8 mm hole was inserted to another cylinder of the same setup. The plug was inserted at 08:30 on 25-05-2021 until 14:00 25-05-2021. Data extraction and Data preparation(idle time removal, outlier removal, grouping and feature extraction) was all done according to the previously stated methodology. The previously built classification model was used to predict the two states of the machine('normal working' and 'leaking'). Note: The data set on 25-05-2021 from 08:30 to 14:00 consists of only faulty data entries corresponding to leakage. The results of the prediction which is a good validation of the working of the model is summarised in the confusion matrix figure 4.15.

As it can be seen in the confusion matrix, the model was 95.23% accurate in detecting the faults with a false positive rate of just 4.76%.

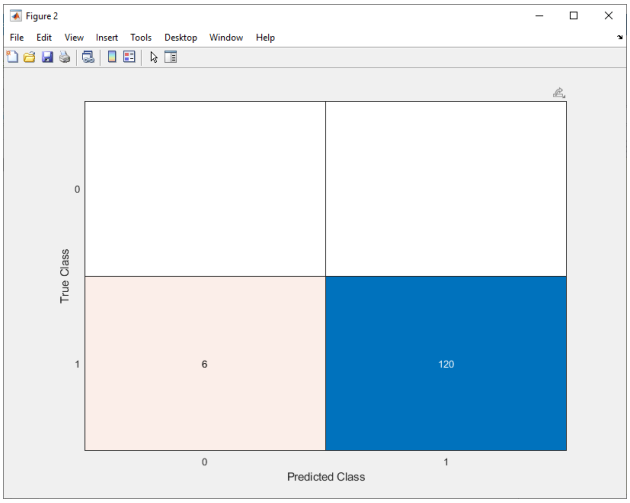

Figure 4.15: Confusion matrix of validation data set.

4.5 Unsupervised Modeling Approach

In the unsupervised approach, which GMM as a powerful clustering technique was implemented, the result shows that finding distinct groups for the machine state of 'Working normal' and 'Leakage' was successful. The GMM model was trained with the seven significant features It should be noted that the labels('Normal' or 'Leaking') of the data set is known as we are aware that a ticket for leakage in the wrapper machine was raised on 09/12/2020 and fixed on 18/01/2021. These labels were not introduced to the model thus making it unsupervised learning. Moreover, the known labels were only used to validate the performance of the model Figure 4.16 shows a seven by seven scatter plot of the different features in relation to each other, and is grouped by the cluster the model has predicted. It shows that when the temperature is on one of the axes it results in two very distinct clusters, but only looking at the airflow the two clusters are overlapping. The silhouette score for the two clusters was calculated to be 0.580, meaning that the majority of the samples are distinctively assigned to the correct cluster.

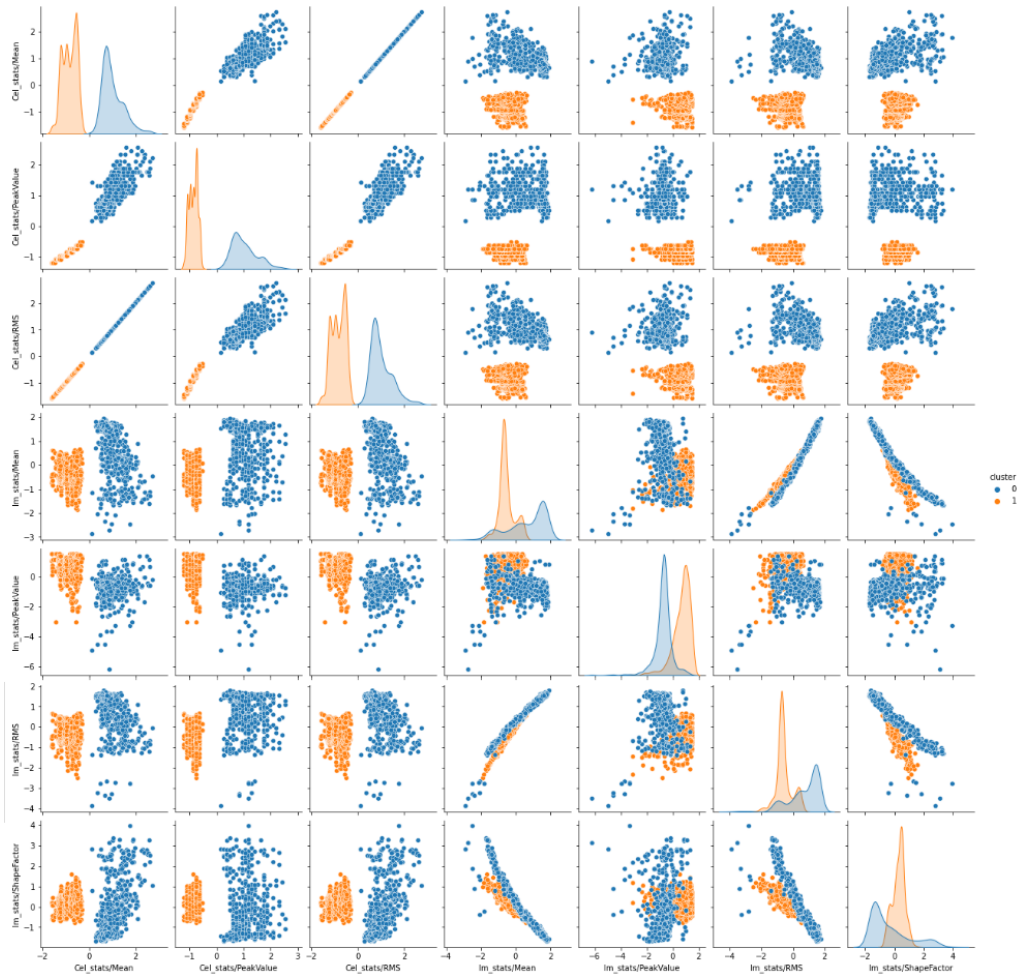

Figure 4.16: Pairplot of different features in relation to each other, and grouped by predicted cluster

The labeled dataset was compared with the predicted clusters the model generated (see figure 4.17) and the F1 score was calculated to be 0.994. As stated before, the labels(0=‘normal’ and 1=‘leakage’) of the machine which were labelled based on the fault reported in the system, were not introduced to the model. The known true labels are compared with the prediction made by the model as a validation for the performance of the model. This good F1 score indicates that the model prediction is outstanding.

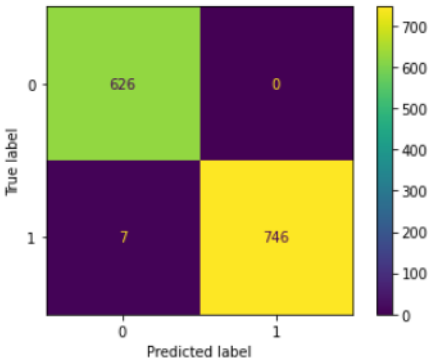

Figure 4.17: Confusion matrix of GMM-model

5
Discussion

It is a common challenge for companies to translate the insights obtained from the results into a competent business action. This chapter focuses on discussing the results and interpreting the outcomes thus facilitating the stakeholders to get necessary motivation for the future measures to be taken. Adding on, this chapter discusses the research questions that avail in deeper understanding of the results.

RUSboosted bagged trees model stands for random undersampling combined with standard boosting procedure AdaBoost. This is an algorithm built to handle imbalanced data with discrete class labels. The given dataset poses an imbalance in the population of data entries belonging to different groups and RUSboosted bagged trees model is the most suitable to handle it. This explains why the RUSboosted bagged trees model was the best performing model among the many models that were trained using the classification learner app. Considering the fact that there was limited evidence of historic anomalies of pneumatic leakage recorded in the SKF database, the presented approach of performing experiments to induce artificial anomalies and labelling the data points to conduct a supervised learning and build a classification model was successful. As an effort to answer the research question1(RQ1) of the presented thesis, labelling of data was carried out by performing two experiments in which pneumatic leakage was introduced to the system manually. The results of experiment 1 gave a valuable insight that the sensitive feature (Airflow) behaves differently with different size of the leaking hole and most importantly all of the behaviours were distinct. With this important insight experiment 2 was conducted on a normal production day to see distinct behaviour of the sensitive feature for pneumatic leakage produced by a 0.8 mm manually drilled hole and label the same. However there are some drawbacks with this approach. Starting with the whole approach being 'contextual anomaly detection', the model may fail to detect anomalies other than this particular type of pneumatic leakage. Although, with the limited time frame of this thesis work only two experiments could be conducted, it gave a solid foundation by providing a standard procedure to further work on more number of experiments with the possible following variations:

- Variation in the diameter of the hole resulting in leakage.
- Placement of the plug in different actuator cylinders.
- Conducting the experiments on different machines.
- Conducting the experiments with a focus of finding multiple states of machine fault like early stage of leakage, middle stage of leakage and severe stage of leakage.

These variations are essential to capture all the scenarios that might lead to leakage thus making the model more robust.
In order to validate the performance of the model (other than the validation performed on 20% of cleaned data calling it as test data) and to cover one of the variations discussed above, a validation dataset was extracted on 25-05-2021 from 08:30 to 14:00 consisting of only faulty data entries corresponding to leakage. The placement of the plug with 0.8 mm hole was carried out on another actuator cylinder. The good performance(95.23% accurate and 4.76% FPR) of the classification model on this variation of set up, gives some kind of confidence that this model can work right in capturing leakages due to other variations as well, given that more experiments with variations are performed to confirm the same.
This thesis work throws an ample amount of light on the significance of the features used for modelling which relates to the research question number 2 (RQ2). During the iterations for finding the optimum number of features with good evaluation metrics, it was established that training the model with 33 significant features yielded the best accuracy, precision, recall, and F1 score. However, examining the histograms of descriptive statistical features created in Diagnostic feature designer app, seven features(see figure 5.1) which also happens to be the top 7 ranked features in Kruskal-Wallis scoring system, with a score above 400, are the most significant features playing a major role in helping to differentiate the groups. (Note: The elements of evaluation metrics vary by a close margin irrespective of the number of features the model is trained on, see Table 4.1: Performance metric comparison)

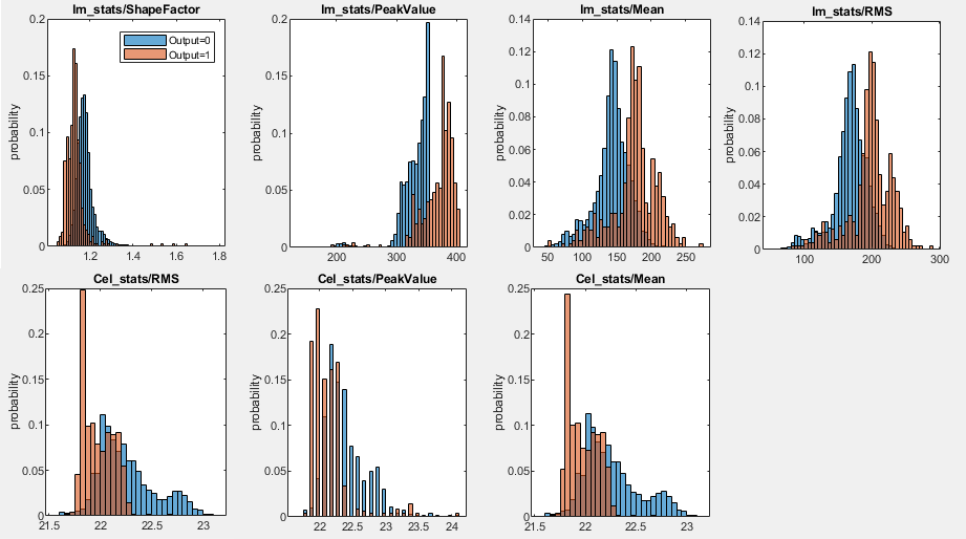

Figure 5.1: Eight top statistical features for experiment 2

Initial examination of raw data in Grafana made it seem that temperature feature had no significance in detecting the leakage as it showed almost no sensitivity to various states of the machine. Although in the feature extraction stage, it was learned that derived features such as peak value, mean, and RMS of the temperature had good potential to identify the distinct groups(see figure 4.16). Machine learning

is a tool tasked to identify the hidden patterns which are non identifiable by human interpretation. This is a major takeaway from the thesis.

The insights obtained pertaining to RQ2 was useful to conduct the unsupervised learning which is the foundation for answering RQ3. Based on the significant features, a clustering model was built which successfully classified the two states of the wrapper machine('Normal', and 'Leaking') into two distinct clusters. This model was built using the significant features that are mentioned in RQ2 for the purpose of the training data extracted from the wrapper machine.

Looking at the Figure 4.16: Pairplot of different features in relation to each other, and grouped by predicted cluster, It was clearly grasped that descriptive statistical features(mean, RMS, peak value, shape factor) pertaining to airflow(sensitive feature) alone could not provide distinct clusters (the overlap of the clusters can be seen in the pairplots) but, descriptive statistical features pertaining to temperature along with descriptive statistical features pertaining to airflow had coalesced successfully to provide distinct clusters. This also strengthens the importance of temperature, which was not believed to be significant in the early stages, as a feature when used with appropriate machine learning approaches, has the potential to provide a positive outcome. Stressing on the limitations of the presented unsupervised approach, the ticket regarding the malfunction of the wrapping machine due to leakage was raised at the extreme stage where the malfunction was visually evident and physically interrupting the production line. The primary focus of this thesis work is to detect the leakage at the early stage, at a stage where it is not physically hindering the normal functioning of the machine, and provide enough time for the personnel to take necessary proactive actions before the failure has any undesired consequences. Adding on, instead of labelling the data from the day ticket was raised till the day it was rectified as 'abnormal' and training the clustering model on it, the data points captured before the day ticket was raised could be labelled as 'abnormal' which represents the early stage of the leakage. The problem with this approach is that the exact timeline of the data points which represents the early stage of leakage could not be affirmatively established. It could have been leaking for days, weeks are even months before it becomes severe enough for a ticket to be raised. Stressing on the methodological discussions, particular phases in the methodology used in the presented thesis were iterative. Working on the framework of CRISP-DM, many different approaches were tried and the best approach which is logically explainable and with good results was chosen to progress further. The thesis could have further explored the power of deep learning and neural networks as well, but given the obtained accuracy of 99.4% using simple machine learning models, stepping into complex models would be unnecessary.

The whole approach of unsupervised modelling could have been eliminated if RUSBoosted classification model was used in the approach modelling the pneumatic leakage that occurred on December 18th 2020. This would have given a solid validation of the model which was built on artificial anomalies to identify true real world anomalies. However, an unsupervised approach was adopted to visually strengthen the importance of temperature features for modelling purposes. Using a whole new approach to validate temperature to be significant was the sole purpose of using an unsupervised approach.

Another possible approach would be to use the digital outputs and inputs for the cylinders, and with these define the start end for a complete work cycle. This would give another result for the final data set used for algorithm modeling, where the samples would not be grouped by an arbitrary 'processing time' of 60 seconds, but instead a proper unit for processing time. Here also the idle time removal would be automatically corrected with known start and end times. This would preferably be tested in future research for less data preparation tasks in general, and possibly could increase the accuracy even more.
As deliverables to SKF, a supervised model, an unsupervised model and insights about temperature features are crucial. The algorithms are delivered to IFM Electronics and possible deployment solutions are discussed. Most importantly, this was an initial foundation and a first approach of machine learning to detect pneumatic leakage. This presented thesis work acts as a helping guide for more future research by providing knowledge about previously unknown insights. Dashboards that display real time group labels and alarms the maintenance personnels in case of leakage detected from these models are some of the future ambitions and goals that can be achieved.
Considering the demerits and merits of the presented approach, it can be learned that machine learning opens up the dimension of problem solving with numerous approaches, choosing the appropriate one is a challenge.

6
Conclusion and Recommendation

The fact that AI and Machine learning has opened up a treasure of propitious possibilities is undeniable. However, it can be concurred with certainty that these possibilities have led to the emergence of 'AI solutionism', a philosophy referring to the belief that machine learning has a solution for all kinds of problems. As discussed before, the boom in the volume of data procured as a result of digitalization has led the companies to think and act in the newly opened dimension of Data-Driven Decision Making. It is a common challenge for the companies to translate the insights of these data analytics into business actions. Quoting the success of finding out the previously unknown hidden particulars of descriptive statistical features of temperature prominently influencing the modelling capabilities proves that this problem was an optimum one for the machine learning approach to be befitting. Speaking of the second part of the earlier discussion which raises concern for translating the insights into business actions; this thesis throws light on two different kinds machine learning approach for the same problem, hence emphasising on the actuality that there exists multiple approaches in finding a solution and machine learning is more fruitful with iterations and trial and error methodology. As an attempt to narrow the gap between data analytics insights and deployment, and moreover as deployment stays out of the scope of this thesis work, we are currently are in touch with, IFM electronic and delivered the algorithm with a recommendation to make use of these pneumatic detection techniques to raise smart alarms or even notify maintenance personnel with a automated mail. Most importantly, the work methodology and the derived useful previously unknown insights provide a concrete platform for further research in this area and facilitate a full fledged deployment of predictive maintenance in the near future.

Talking about the contributions of this thesis in a practical perspective, RUSBoosted Bagged Trees Classification model, GMM Clustering model, establishment of significant features that play a vital role in differentiating the different state of the machine, and the key understanding about the importance of temperature feature in modelling were the crucial findings. The supervised machine learning model was successful in detecting the early stage of leakage with an accuracy of 98.2%, thus giving plenty of time to perform appropriate maintenance on the equipment. The unsupervised model with an accuracy of 99.3%, was successful in detecting the physically evident stage of leakage.

When seen from an academic perspective, the contribution of this thesis is compelling. In the phase of literature review, it was observed that not a considerable

amount of work was available pertaining to data labelling by artificially induced anomalies using physical experiments. In a case, as ours, where the availability of data capturing the historic anomalies are limited, this approach proves to be useful. Moreover, for newly installed systems or a set up which has not seen any kind of malfunction yet, this combination of artificially induced anomalies by experiments and machine learning approach, helps the production system to proactively prepare for any kind of future hindrance for its smooth functioning.

This work gave an opportunity to identify the areas of production systems where data science tools and techniques could be applied for data-driven decision-making. Identifying leakage in early stages will help to minimize the wastage of power consumed to produce compressed air, thus one step taken towards achieving sustainable production systems.

Bibliography

[1] C. Lundgren, J. Bokrantz, and A. Skoogh, "A strategy development process for Smart Maintenance implementation," *Journal of Manufacturing Technology Management*, vol. 32 no. 9, pp. 142-166, Apr. 30, 2021 doi: https://doi.org/10.1108/JMTM-06-2020-0222

[2] G. Frizelle, D. McFarlane, and L. Bongaerts, "Disturbance Measurement in Manufacturing Production Systems", *In Proceedings of ASI '98*, Bremen, Germany. Available: https://www2.ifm.eng.cam.ac.uk/uploads/Research/DIAL/Resources/Papers/Disturbance_Measurement.pdf, Accesses on: May 26, 2021.

[3] D. Mcfarlane, J. Matson, "Assessing and Improving The Responsiveness of Manufacturing Production Systems", *IEE Seminar Customer Focused Manufacturing: Survival of the Fittest*, London, United Kingdom, 1999 doi: 10.1049/ic:19990802

[4] Lee, J., Ni, J., Singh, J., Jiang, B., Azamfar, M., and Feng, J. (August 18, 2020). "Intelligent Maintenance Systems and Predictive Manufacturing." ASME. *J. Manuf. Sci. Eng.* November 2020; 142(11): 110805. https://doi.org/10.1115/1.4047856

[5] Santos M.Y. et al. A Big Data Analytics Architecture for Industry 4.0. In: Rocha Á., Correia A., Adeli H., Reis L., Costanzo S. (eds) *Recent Advances in Information Systems and Technologies. WorldCIST 2017. Advances in Intelligent Systems and Computing*, vol 570. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-56538-5_19

[6] British Standards Institution, "Glossary of Maintenance Managament Terms in Terotechnology"m 1984.

[7] A. Schokry, "Introduction to Maintenance Second semester 2010/2011", Islamic university of Gaza, Palestine. Available: http://site.iugaza.edu.ps/aschokry/files/2011/02/Maintenance-II.pdf Accesses on: May 26, 2021.

[8] M. Gopalakrishnan. (2019). Maintenance & Reliability - Role of Maintenance in industries [Online].

[9] J. Bokrantz, A. Skoogh, C. Berlin, T. Wuest, and J. Stahre, "Smart Maintenance: an empirically grounded conceptualization," *International Journal of Production Economics*, vol. 223, May, 2020. doi: 10.1016/j.ijpe.2019.107534

[10] R. Tsvetkova, "WHAT DOES INDUSTRY 4.0 MEAN FOR SUSTAINABLE DEVELOPMENT?," *Industry 4.0*, vol. 2, no. 6, pp 294-297, 2017, Available:

https://stumejournals.com/journals/i4/2017/6/294 (accessed on: May 26, 2021).

[11] J. J. M, S. Schwartz, R. Vingerhoeds, B. Grabot, M. Salaün, "Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics," *Journal of Manufacturing Systems*, vol 56, pp. 539-557, July. 2020, doi: https://doi.org/10.1016/j.jmsy.2020.07.008

[12] K. Fuchs, "Machine Learning: Classification Models," *Medium*, [Online]. Mar. 28, 2017. Available: https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529 (accessed on: May 26, 2021).

[13] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1-4, doi: 10.1109/ICPR.2008.4761297.

[14] M. J. Garbade, "Regression Versus Classification Machine Learning: What's the Difference?," *Medium*, [Online]. Aug. 11, 2018. Available: https://bit.ly/34pQ8XU. (accessed on: May 26, 2021).

[15] A. Kassambara, "K-Means Clustering," *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA, 2017, ch. 5. Available: https://bit.ly/3vxRZp0

[16] M. Wagle, "Association Rules: Unsupervised Learning in Retail," *Medium*, [Online]. Mar 25, 2020. Available: https://medium.com/@manilwagle/association-rules-unsupervised-learning-in-retail-69791aef99a (accessed on: May 26, 2021).

[17] A. Singh, "Build Better and Accurate Clusters with Gaussian Mixture Models," *Analytics Vidhya*, [Online]. Oct. 31, 2019. Available: https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/ (accessed on: May 26, 2021).

[18] H. Xie, X. Hu, Z. Peng, J. Jiang, B. Wen and Q. Yang, "Energy System Time Series Data Quality Maintenance System Based on Data Mining Technology," *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, 2020, pp. 3096-3100, doi: 10.1109/EI250167.2020.9347363.

[19] L. Cai, and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal*, vol. 14, p. 2, May. 2015, doi: http://doi.org/10.5334/dsj-2015-002

[20] T. C. Redman, Data Quality Management Past, Present, and Future: Towards a Management System for Data. In: Sadiq S. (eds) Handbook of Data Quality. Springer, Berlin, Heidelberg. 2013. https://doi.org/10.1007/978-3-642-36257-6_2

[21] B. Esmael, A. Arnaout, R. K. Fruhwirth, G. Thonhauser, "A Statistical Feature-Based Approach for Operations Recognition in Drilling Time Series," *International Journal of Computer Information Systems and Industrial Management Applications.*, vol. 5, pp. 454-461, ch IV. Statistical Features Extraction. Available: https://pure.unileoben.ac.at/portal/files/1073786/A_Statistical_Feature_Based_Approach_for_Operations_Recognition_in_Drilling_Time_Series.pdf

[22] X. Deng, Q. Liu, Y. Deng, S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Information Sciences*, vol. 340-341, pp. 250-261, May, 1. 2016. doi: https://doi.org/10.1016/j.ins.2016.01.033

[23] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, Nov, 1987. Available: https://doi.org/10.1016/0377-0427(87)90125-7 (accessed on: May 26, 2021).

[24] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096

[25] Cdang. "Simplified cross section of a double acting cylinder, with the vector forces," 2011 [Electronic Image]. Available: https://commons.wikimedia.org/wiki/File:Coupe_verin_double_effet_exercice.svg.

[26] X. Li, I. Kao, "Analytical fault detection and diagnosis (FDD) for pneumatic systems in robotics and manufacturing automation," *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, AB, Canada, 2020, pp. 2517-2522. doi: 10.1109/IROS.2005.1545573.

[27] M. Barros, B. Veloso, P. M. Pereira, R. P. Ribeiro, J. Gama, "Failure Detection of an Air Production Unit in Operational Context. In: Gama J. et al. (eds) IoT Streams for Data-Driven Predictive Maintenance and IoT, Edge, and Mobile for Embedded Machine Learning. ITEM 2020, IoT Streams 2020", *Communications in Computer and Information Science*, vol 1325. Springer, Cham. Jan 2021. doi: https://doi.org/10.1007/978-3-030-66770-2_5

[28] A. Desmet, M. Delore, "Leak detection in compressed air systems using unsupervised anomaly detection techniques," *In Annual Conference of the Prognostics and Health Management Society.*

[29] A. Bousdekis, K. Lepenioti, D. Apostolou, G. Mentzas, "A Review of Data-Driven Decision-Making Methods for Industry 4.0 Maintenance Applications," *electronics*, vol. 10(7), 828, Mar. 31, 2021. doi: https://doi.org/10.3390/electronics10070828.

[30] S. Ransbotham, D. Kiron, P. Kirk Prentice, "Minding the Analytics Gap," *MIT SLOAN MANAGEMENT REVIEW*, Spring. 2015. Available: https://www.adaptiveplanning.com/sites/default/files/assets/Adaptive-Insights-Whitepaper-Minding-the-Analytics-Gap.pdf

[31] W. Vorhies, "CRISP-DM – a Standard Methodology to Ensure a Good Outcome," *Data Science Central*, [Online], Jul. 26, 2016. Available: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome (accessed on: May 26, 2021)

[32] R, Wirth., J, Hipp. CRISP-DM: Towards a Standard Process Model for Data Mining. Available: http://www.cs.unibo.it/ danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf, Accesses on: May 26, 2021.

[33] T. Dasu, T. Johnson, *Exploratory Data Mining and Data Cleaning*, Hoboken, NJ, USA: John Wiley & Sons, Inc, 2003.

[34] H. Wickham, "Tidy Data," *Journal of Statistical Software*, issue II. Available: https://vita.had.co.nz/papers/tidy-data.pdf (accessed on: May 26, 2021).

[35] E. F. Codd, *The RELATIONAL MODEL for DATABASE MANAGEMENT*, VERSION 2., USA: ADDISON-WESLEY PUBLISHING COMPANY, 1990. [Online]. Available: https://dl.acm.org/doi/pdf/10.5555/77708

[36] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. Available: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf (accessed on: May 26, 2021).

[37] J. Fernando, "Correlation Coefficient," *Investopedia*, [Online]. Feb. 23, 2021. Available: https://www.investopedia.com/terms/c/correlationcoefficient.asp (accessed on: May 26, 2021).

A

Appendix 1

Table A.1: Framework for data quality report

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	Whether a data access interface is provided
		Data can be easily made public or easy to purchase
	2) Timeliness	Within a given time, whether the data arrive on time
		Whether data are regularly updated
		Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	Data come from specialized organizations of a country, field, or industry
		Experts or specialists regularly audit and check the correctness of the data content
		Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	Data provided are accurate
		Data representation (or value) well reflects the true state of the source information
		Information (data) representation will not cause ambiguity
	2) Consistency	After data have been processed, their concepts, value domains, and formats still match as before processing
		During a certain time, data remain consistent and verifiable
		Data and the data from other data sources are consistent or verifiable
	3) Integrity	Data format is clear and meets the criteria
		Data are consistent with structural integrity
		Data are consistent with content integrity
	4) Completeness	Whether the deficiency of a component will impact use of the data for data with multi-components
		Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	The data collected do not completely match the theme, but they expound one aspect
		Most datasets retrieved are within the retrieval theme users need
		Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	Data (content, format, etc.) are clear and understandable
		It is easy to judge that the data provided meet needs
		Data description, classification, and coding content satisfy specification and are easy to understand
	2) Structure	Level of difficulty in transforming semi-structured or unstructured data to structured data

B

Appendix 2

Table B.1: Experiment 1 timeline

Timeline(GMT + 1)		Phase	Machine Status	Comments
Start	Stop			
09:36	10:04	Phase 1	Cycle Running	Normal cycle without leakage
10:04	10:06	-	Machine Down	Door is opened, plug with 1 mm hole is inserted
10:08	10:37	Phase 2	Cycle Running	Cycle with minor leakage, actuator still seem to work properly
10:37	10:38	-	Machine Down	Door is opened, adapter is switched to the other side of the cylinder
10:39	11:04	Phase 3	Cycle Running	Cycle with minor constant leakage, actuator still seem to work properly
11:04	11:04	-	Machine Down	Plug in adaptor is removed
11:04	11:07	Phase 4	Cycle Running	Cycle with major constant leakage, actuator is barely moving
10:07	-	-	Normal Cycle	After phase 4 the tubing was restored to original state

C
Appendix 3

Table C.1: Data Quality result

DIMENSION	ELEMENT	COMMENTS	SCORE
Availability	Accessability	The data was provided by SKF and was easily accessible by connecting to the database with the help of an open source management tool for Postgres called PgAdmin 4. The company also provided its database accessibility to Grafana which is an open-source platform for data visualization, monitoring and analysis. Making the data public was not allowed by the company.	1
	Timeliness	The data was updated in real time regularly and was available to be exported into the desired format of our choice for the further processing.	1
Usability	Credibility	The data obtained is credible as it is retrieved from one of the specialized organization; SKF, in the manufacturing sector of Sweden. Domain experts regularly keep a check on the data correctness. The SKF database is timely maintained and monitored by experts in this specialized field.	1
Realiability	Accuracy	Data provided is accurate to the milliseconds which reflects the true state of the source information without any ambiguity. The sensors that recorded the data were from the renowned company IFM who are a specialized organization in the sector of manufacturing sensors that read accurate values.	1
	Integrity	The data format is very clear and consistent with structural integrity and content integrity.	1
	Consistency	The fields in the data set contain values which adhere to the consistency of standard units (bar, celsius and liters per minute) and consistency of Time units(seconds) throughout the dataset.	1
	Completeness	The dataset does not contain any missing values or incomplete data. Though the data is recorded every second, it is complete and consistent throughout thus exhibiting row completeness and column completeness.	1
Relevance	Fitness	Not all the data collected in the database which could be accessed is used, the portion of data retrieved are within the retrieval theme users need and are highly relevant for the further processing according to the methodology.	1
Presentation Quality	Readability	The data in the CSV format is easy to read and understand. The content and format of the data correctly explains itself according to known terms, attributes, abbreviations and units.	1
	Structure	The data is extracted directly from Grafana in the form of a CSV file, there were no efforts to convert unstructured data into structured data as the extracted data itself was obtained in a highly structured manner.	1

Department of Industrial and Materials Science
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY