E. Ruijs

Automatic Failure Diagnosis for Flow Control Valves

Automatic Failure Diagnosis for Flow Control Valves

By

E. Ruijs

Master Thesis

in partial fulfilment of the requirements for the degree of

Master of Science
in Mechanical Engineering

at the Department Maritime and Transport Technology of Faculty Mechanical, Maritime and Materials Engineering of Delft University of Technology
to be defended publicly on April 15, 2020 at 02:00 PM

Student number:	4229665	
MSc track:	Transport Engineering and Logistics	
Report number:	2020.MME.8417	
Thesis committee:	Prof. R. Negenborn,	TU Delft committee Chair, 3mE
	Dr. X. Jiang	TU Delft, supervisor, 3mE
	Dr.. R. Ferrari,	TU Delft committee member, 3mE (DCSC)
	Dr. T. Park,	Supervisor, Royal Dutch Shell
Date:	April 6, 2020	

Preface

This Master thesis describes ’Failure diagnosis for flow control valves using supervised machine learning’. The graduation project is done as a final assignment within the track Transport engineering & Logistics of the MSc. Mechanical Engineering at the TU Delft. The research is executed in collaboration with the data science department at Royal Dutch Shell in Amsterdam. Predictive maintenance and data science were relatively new topics to me, which allowed me to learn many new skills and broaden my academic horizon. The development of new technologies and strategies in the field of maintenance in combination with digitalization will bring us tons of possibilities in the coming years, therefore, I am very proud to have been a part of this research.

First of all, I would like to thank Dr. Xiaoli Jiang for her supervision, and guidance during the research. Second of all, I want to thank the chair of my committee Prof Rudy Negenborn for challenging me during the project which definitely brought the result to a higher level.

Furthermore, I would like to thank all my colleagues at the digital centre of excellence of Shell in Amsterdam for always providing me with a fun experience in the office and helping me when needed for the past months. Especially, a big thanks to Dr. Timothy Park for helping me steer in the right direction throughout the whole process with critical questions, and teaching me with your passion for statistics. Finally, I would like to thank Dr. Maurice Hendrix from the refinery in Pernis for helping me out with the valve data, case study, and showing me around on the asset.

E.R. Ruijs
Delft, April 2020

Acronyms

Acronym	Description
AI	Artificial Intelligence
ATC	Air to Close
ATO	Air to Open
ANOVA	Analysis of Variance
ANN	Artificial Neural Network
OP	Controller Output
IP	Current to Pressure
DS	Drive Signal
FFT	Fast Fourier Transform
FDI	Fault Detection and Isolation
GNB	Gaussian Naïve Bayes
GBDT	Gradient Boosting Decision Tree
IG	Information Gain
IoT	Internet of Things
LR	Logistic Regression
ML	Machine Learning
MSE	Mean Squared Error
MTTF	Mean Time To Failure
MTTR	Mean Time To Repair
MLP	Multi-Layer Perceptron
MI	Mutual Information
NB	Naïve Bayes
PSD	Power Spectral Density
PV	Process Variable
RF	Random Forest
SP	Set Point
SGHP	Shell Gasification Hydrogen Plant
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TD	Travel Deviation
TA	Turn Around
CV	Valve travel setting signal

Executive summary

The increased implementation of digitalisation all over the world has led to an exponential growth of available data across various industries. Consequently, there is a large growth of advanced machine learning (ML) techniques applied to process data. Shell is an international energy company with expertise in the exploration, production, refining and marketing of oil and natural gas, and the manufacturing and marketing of chemicals. Maintenance for process equipment can be applied using three different strategies: preventive, predictive, and reactive maintenance. Traditionally, the two maintenance strategies used in refineries are preventive and reactive maintenance. With more data becoming available, new value can be created from existing businesses. Predictive maintenance is a digital strategy using condition-based monitoring techniques to track the performance of equipment to detect possible defects in advance. In oil refineries, various process equipment is used of which flow control valves are essential to regulate the throughput of heavy, possibly dangerous material. Control valves are fixed process equipment in oil refineries, where a network of closed control loops contributes to the generation of finished products. Component failure in flow control valves should be detected early in order to avoid unexpected breakdown, causing downtime. Therefore, the application of predictive maintenance strategies on flow control valves is researched. Currently, the Shell Pernis refinery uses ML to detect failures in control valves up to 75 days in advance. The problem, however, is that diagnosing failures by hand requires too much time, causing an overload of unresolved alarms. The goal of the research is to develop an effective automated failure diagnosis method for flow control valves in the refinery of Pernis when a failure is known to be present. Effective in terms of reduced time required for diagnosis and sufficient accuracy. Figure 1 shows the architecture of fault detection and diagnosis for flow control valves.

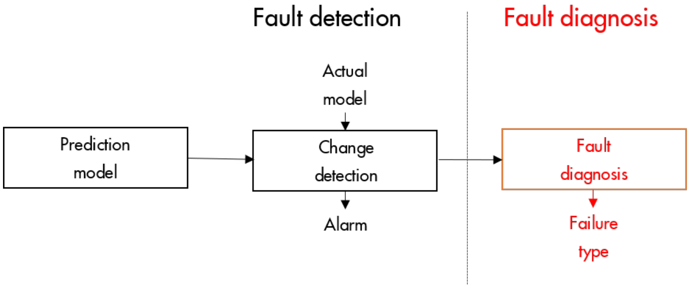


Figure 1: Failure detection and diagnosis in flow control valves. Black: fault detection model raising an alarm whenever a change is detected between the predicted and actual operating state of the control valve. Red: fault diagnosis model with a failure type as output when an alarm is raised by the fault detection tool.

Automatic failure diagnosis in process equipment and control valves has widely been discussed in the literature. The applied methods are ranging from trend analysis, clustering techniques, to deep-learning. Four criteria are used to choose the appropriate procedure for this research: interpretability for the user, ability to use the model generically for the whole asset, robustness to work with small sample size, and the risk of the model overfitting. Three methods applicable to this problem are scored on the criteria by multiple domain experts. Based on these results, the method ’feature engineering and supervised machine learning’ has been chosen, due to its high interpretability, ability to be used plant-wide, and robustness with small datasets.

In flow control valves, the Set-Point (SP), Process Variable (PV), and Controller output (OP) are recorded from the feedback loop. The control valve aims to minimise the difference between the PV [tonnes per day] and the SP [tonnes per day] by adjusting the OP [%]. The data of these three signals is assumed to contain the failure mode in the flow control valve and therefore are used as time series input for the model. Five failure mode categories are identified in Shell’s data: Fail to close, Fail to open, Hunting, Hysteresis, and Other. The ’Fail to close’ category contains failure cases where the valve has trouble closing, which results in the valve steering the OP to close but failing and therefore causing a significant difference between the SP and the PV. The characteristics of the ’Fail to open’ category are precisely the opposite. The ’Hunting’ behaviour shows a

constant overshoot of the positioner, resulting in an anti-phase between the signals OP and PV. ’Hysteresis’ is the difference between the valve position on the upstroke and its position on the downstroke at any given input signal, resulting in different values of the PV for the same value of the OP. The ’Other’ category show failures where no explicit failure behaviour can be identified in the input data. The training and test set for the machine learning classification model have been prepared using the input and output data. The training data, built from failure cases extracted from Shell’s maintenance notification database, covers five plants with different locations and applications. The case study includes data from the SGHP unit, where the gasification process takes place on refining residue materials. The dataset of this new unit is kept separately and has a different application from the plants involved in the training set.

Given the available data, the methodology for the extraction and selection of features and supervised ML algorithms are determined. The complete framework chosen for failure diagnosis in control valves for the Shell Pernis refinery consists of the three steps, shown in Figure 2. The data input consists of the time series of the control valves showing failure behaviour. In the first step, 13 features are extracted from the time series. Feature extraction refers to the transformation of data into formats that are suitable for an ML model. The set of features are built in order to distinguish the characteristics of the various failure behaviour categories, using statistical methods on the signals, such as the mean, variance, correlations, and ratios. Furthermore, Power Spectral Densities (PSD) of the signals are applied to distinguish between frequencies and therewith create features. The second step contains two selection methods used to choose the features that are significant predictors. The ANOVA (F-test) measures the degree of linear dependency between two random variables by calculating an F-value that compares the variability between and within the groups. The Mutual information (MI) test, based on Shannon’s entropy, is used to determine the non-linear mutual dependence between two random variables by quantifying the ’amount of information’ obtained about one random variable as a result of observing the other random variable. In the third step, three ML classification models are used to predict failure behaviour on new incident data. Logistic regression (LR) is a widely used ML classifier that determines the probability that a specific value of the predictor belongs to a particular class or category using a logistic function. Naive Bayes (NB) is a probabilistic classifier, based on Bayes theorem. The algorithm calculates the posterior probability, which is the chance of a specific class given the observation. A Random Forest (RF) is a classifier consisting of a collection of tree-structured classifiers where each decision tree casts a unit vote for the most popular class. This results in 6 different prediction outcomes which are assessed using k-fold cross-validation on the training set with the performance metrics: accuracy and log loss. The method with optimal performance is used to predict the outcomes of new incident data in the case study. Furthermore, the ML algorithm parameters are tuned and the input length of the time series is varied using $t = 200$, $t = 1000$, and $t = 5000$ [minutes]. Cross-validation is used on the training set to prevent overfitting, and due to the large variety of plants present in the samples.

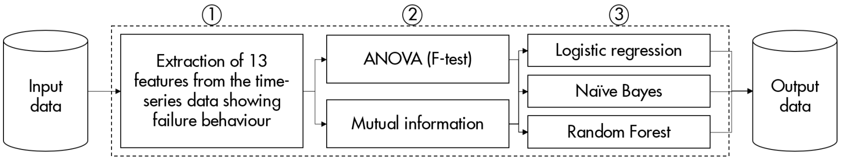


Figure 2: Final framework used for failure diagnosis in control valves for Shell’s refinery in Pernis.

The cross-validation result of the RF model, using a maximum tree depth of 10, shows a significantly higher accuracy on the training set compared to the other models when using the ANOVA selection method with a confidence interval of 95% and $t = 200[m]$. The accuracy achieved is 0.88 ± 0.06, and a log-loss value of 0.49 is obtained. Next to the cross-validation results, the use of the short time-window decreases the chance of noise, in this case, normal operating behaviour, being present in the time series.

The case study, performed on the gasification plant of Shell’s refinery using the RF classifier, consists of unseen plant data not represented in the training set. The input data consists of 6 features, tested as significant with a confidence interval of 95% using ANOVA, results in matrix x_{ij} for $(i = 1,..,18)$ and $(j = 1,..,6)$. After training the model, the most important feature, measured in percentage of the total features, is the mean OP with 30%. The least important feature is the absolute difference between OP and PV.

	Accuracy	Log-loss
RF ($t = 200[m]$)	0.81	0.46

Table 1: Case study results of the RF classifier on accuracy and log-loss using 6 features selected using ANOVA with $t = 200[m]$.

The results from Table 1 show, using the proposed novel framework for automatic failure diagnosis in control valves, incident behaviour in flow control valves can be classified and predicted with an accuracy of 81% using supervised ML classification. Another finding shows the prediction probability of the actual label is higher than 20% at all times, which opens up opportunities for diagnosis strategies where critical failure types should be prevented at all times for a specific valve. Also, due to the recall values of 1.0 for the classes 'Fail to close' and 'Fail to open', such a prevention strategy can be realised. The implementation of the automatic fault diagnosis, therefore, can have a considerable influence on the current state of the failure diagnosis process for engineers. Figure 3 shows the confusion matrix of the case study using the RF classifier.

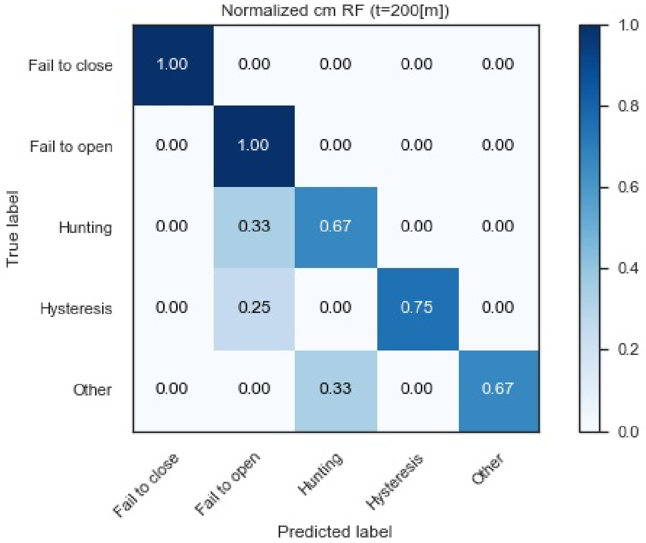


Figure 3: Confusion matrix of the case study using the RF classifier with 6 features selected by the ANOVA test with $t = 200[m]$.

The issue that Shell faces is the lack of information in the output of the available fault detection model, resulting in an overload of unresolved alarms. Therefore, the goal of this research was to develop an effective automated failure diagnosis method for flow control valves in the refinery of Pernis. Effective in terms of reduced time required for diagnosis and sufficient accuracy. On unseen data, the automatic failure diagnosis method can reduce approximately 7.5 minutes per fault, and obtain an accuracy of 81%. This accuracy is sufficient when compared to the identified minimum by the Shell engineers (73%), and when benchmarked with the current state of the literature (19% increase). Therefore, the goal of this research has been achieved.

The implementation of the proposed method will have an impact on the current failure diagnosis process. First of all, due to the ability to classify and predict within several seconds, the diagnosis time will reduce considerably. Currently, step two of the failure diagnosis process takes about 7.5 minutes on average for an engineer. The whole process will shorten to less than a second using the automatic failure diagnosis. Second of all, engineers can use the model to classify alarms, which will help them in actively prioritising and searching for issues that are critical and urgent.

However, there are also limitations to the model. First of all, the failure diagnosis model is entirely reliant on the predictive fault detection method, since failure behaviour should be present to classify the failure. In order to determine whether the proposed method can also be used for failure detection, the possibility of classifying normal behaviour should be explored. Second of all, the division of failure categories has been done from an engineering perspective. Nevertheless, the categories show overlap and therefore are sometimes hard to distinguish. The results when using a data-driven approach, with an unsupervised clustering method such as SVM, could possibly lead to new decision boundaries and distributions over the classes. The accuracy of the classifier was somewhat lower in the case study in comparison with the cross-validation results of the training set. Therefore, it can be assumed that the gasification data set is not completely represented by the five plants present in the training set. Thus, when new plants are analysed multiple supervised ML models should be considered.

More research should be performed on the creation of new features, in order to test the possibilities of making a better division between the three failure categories 'Hunting', 'Hysteresis' and 'Other'. Due to the implemented selection algorithm, just the significant predictors will be adopted in the model without fundamentally changing the architecture. Furthermore, more training data of new plants should be added, since this research shows it increases the accuracy of predictions. Finally, the implementation of new hardware for control valves, such as smart valve positioners, would provide many opportunities in the detection and diagnosis of control valves. The additions of smart positioner bring several advantages such as automatic calibration and configuration of the positioner, real-time diagnostics, and improved process control. When the actual position of the valve opening is known, many new potential failures could be predicted beforehand. Moreover, smart control valves often have built-in diagnostic capabilities in order to continually be up to date on the current state of the control valve.

List of Figures

List of Tables

Contents

1 Introduction

1.1. Digitalisation

Increasing trends of digitalisation all over the world has led to an exponential growth of available data across various industries. Consequently, there is a large growth of advanced machine learning (ML) techniques applied to process data. According to the discussion paper on artificial intelligence (AI) by McKinsey Global Institute Institute (2017), predictive maintenance has an extremely high potential in the energy industry using AI and other novel technologies. Predictive maintenance provides an integrated solution for monitoring the health status of equipment in the present and future. The goal is to reduce downtime and cost of maintenance by assessing the working condition of equipment and predicting failures. The rapid growth of new technologies such as the Internet of Things (IoT) allows communication between sensors and together with the use of AI algorithms, predictive maintenance is becoming smarter every day (Adhikari et al., 2018).

1.2. Royal Dutch Shell

Shell is an international energy company with expertise in the exploration, production, refining and marketing of oil and natural gas, and the manufacturing and marketing of chemicals. Oil companies are striving for increased efficiency of already existing utilities to remain competitive in a market where the oil prices have dropped. With data becoming exponentially more available new value can be created from existing businesses. Furthermore, Shell is committed to delivering energy responsibly and safely, preventing harm to employees, contractors, local communities and the environment. Therefore, the facilities are safely operated and properly maintained to prevent leaks of hazardous materials. The petroleum industry consists of three sectors: upstream, midstream, and downstream. Shell is an integrated oil and gas company, including all three sectors. The upstream sector includes the search for possible crude oil and natural gas fields, exploratory drilling in wells, and operating wells to bring the raw products to the surface. The midstream sector includes the transportation of raw products to refineries and the transportation of refined products to downstream distributors. The downstream sector includes the refining of crude oil and processing of natural gas to finished products such as gasoline, kerosene, lubricants, and liquefied petroleum gas. Oil refineries are large, complex industrial plants that contain many different process units and enormous amounts of pipelines to transport the goods. This research focuses on Shell's refinery in Pernis, which, with a capacity of 404,000 barrels of crude oil per day, is the largest integrated refinery and chemical manufacturing plant of Europe.

1.3. Predictive maintenance

Maintenance for process equipment can be applied using three different strategies: preventive, predictive, and reactive maintenance. Traditionally, the two maintenance strategies used in refineries are preventive and reactive maintenance. Preventive maintenance is largely done during a turn-around which is a scheduled event where an entire processing plant is taken off-stream to inspect components and execute the necessary repairs. Turn-arounds (TAs) can take up to weeks and are therefore extremely costly. The equipment is supposed to endure until the following TA. However, when working with raw products and high temperatures, this can never be assured. Therefore, when a component fails before its expected lifetime, this can cause highly expensive downtime and reactive maintenance costs. The use of predictive maintenance, using sensors and microprocessor technologies, leads to a decrease of downtime and safety risks (Samdani, 1992). Figure 1.2 shows total costs, consisting of breakdown and repair costs, can be minimised using predictive maintenance compared to other strategies. Due to the large amounts and variety of equipment, about several thousands of control valves in the refinery of Pernis, it is difficult to find an optimal solution. Furthermore, the instrumentation used in the refinery can be outdated, which limits the predictive possibilities available. Currently, Shell has produced a machine learning tool that can detect valve failures in the petrochemical refinery in Pernis.

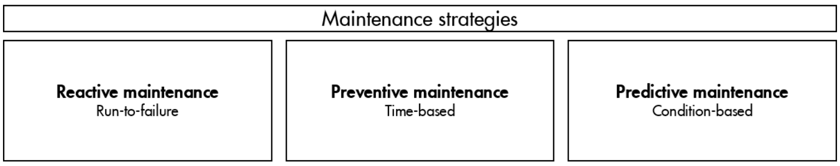


Figure 1.1: Overview of maintenance strategies used in Shell's refinery.

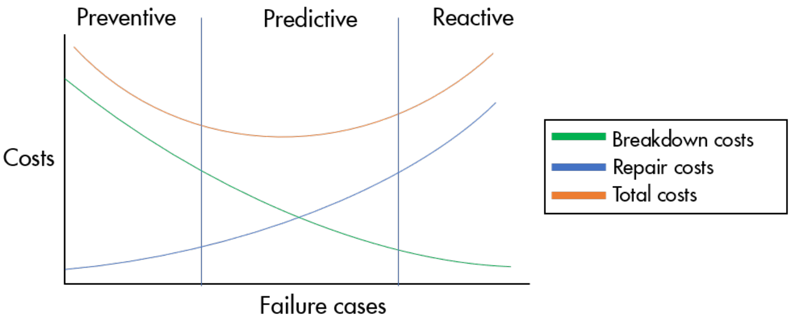


Figure 1.2: Three maintenance strategies applied on process equipment.

1.3.1. Failure detection

Currently, there is a fault detection mechanism live, using ML, that detects control valve failures up to 75 days in advance. The fault detection model uses ML to raise the alarm whenever there is a failure present in the control valve. The alarm specifies the functional location of the valve and a timestamp of when the issue occurred. The analytics engineer receives the alarm and, after diagnosing the failure by hand, reports to the maintenance engineer on-site.

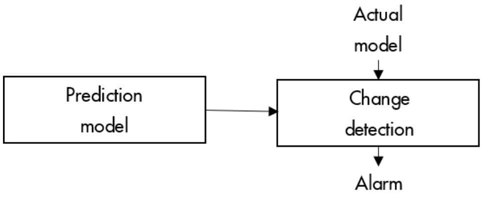


Figure 1.3: Control valve fault detection model.

1.3.2. Failure diagnosis

When the fault detection model raises the alarm, an engineer has to go through several process steps to diagnose the failure and report to the maintenance team on the site. First of all, information regarding the valve location of the valve has to be gathered from technical drawings from the plant. Furthermore, the latest repair has to be known to determine the state of the valve. Second of all, the data around the failure is analysed to decide what failure behaviour the valve is showing, which caused the fault detection model to raise the alarm. Third of all, the data of the valve since the latest repair until the current behaviour is analysed to determine when the issue started. Finally, conclusions on the valve problem are reported and sent to the maintenance engineer on-site. Figure 1.4 shows an overview of the process performed by an engineer.

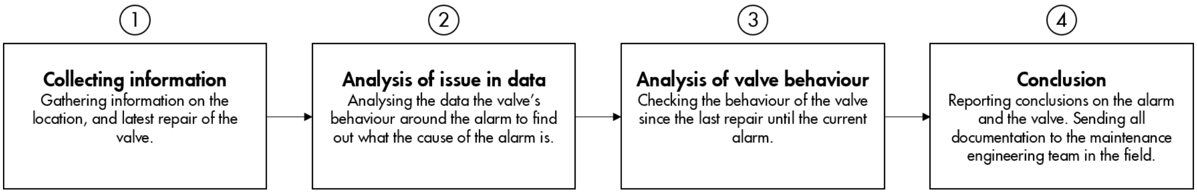

1
Collecting information
Gathering information on the location, and latest repair of the valve.
2
Analysis of issue in data
Analysing the data the valve's behaviour around the alarm to find out what the cause of the alarm is.
3
Analysis of valve behaviour
Checking the behaviour of the valve since the last repair until the current alarm.
4
Conclusion
Reporting conclusions on the alarm and the valve. Sending all documentation to the maintenance engineering team in the field.

Figure 1.4: Process diagram of failure diagnosis for control valves in Shell's refinery of Pernis.

1.4. Problem statement

The main problem Shell faces the lack of information regarding the specific fault given by the fault detection model resulting in an overload of unresolved alarms. The engineers cannot deal with the number of alarms, and this could result in missing a critical failure. Two issues cause a large number of alarms. First of all, failure diagnosis by hand requires too much time, shown in step 2 of Figure 1.4. Secondly of all, there is no possibility for classification or prioritisation of the alarms. This gives the engineers a hard time to work through the alarms given by the fault detection model. Classifying failures can help engineers to prioritise incidents correctly and mitigate the most severe incidents sooner, thus reducing downtime.

1.5. Research goals

The goal of the research is to develop an effective automated failure diagnosis method for flow control valves in the refinery of Pernis when a failure is known to be present. Effective in terms of reduced time required for diagnosis and sufficient accuracy. The automatic failure diagnosis for control valves will shorten step 2 in the failure diagnosis process. Furthermore, automatic failure diagnosis will allow the alarms to be classified and prioritised. Therefore, the result of implementing an automatic failure diagnosis model will reduce the surplus of alarms. The automatic failure diagnosis will work on top of the current fault detection model by triggering whenever an alarm is raised. An overview of the method architecture is shown in Figure 1.5.

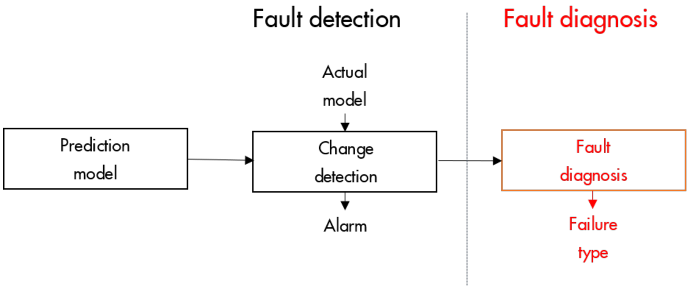


Figure 1.5: Predictive maintenance for control valves. In black: Shell's current fault detection model. In red: the additional automatic failure diagnosis model that will give a failure type as an output when an alarm has been raised by the fault detection model.

Previous research on automatic failure diagnosis in process equipment and control valves has mostly focussed on two methods: model and data-driven based failure diagnosis. Various papers by Bacci Di Capaci et al. (2013) and Scali et al. (2011) describe methods of failure diagnosis for control valves using trend analyses. However, due to the large number and variety of valves, creating a diagnostic model for all control valves in the refinery is difficult. Another research by Prabakaran et al. (2013) successfully applies an artificial neural network to 'DAMADICS', which is a benchmark dataset for control valves. Nonetheless, the interpretability of a deep learning model is low, which raises an issue for engineers willing to work with the model. A recent study on predictive maintenance for aircraft by Adhikari et al. (2018) describes a machine learning-based data-driven diagnostics framework. To the best of our knowledge machine, learning-based classification has not yet been used in control valve failure diagnostics. This study aims to propose a method using supervised machine learning classification to diagnose failures in flow control valves. The training set will consist of failure cases in historical data covering various failure types. This set will be tested using a case study containing new incident data to test the performance of the chosen supervised machine learning algorithms. The objectives of the research can be formulated into a main and several sub research questions:

Main research question

> *What performance can be achieved in the classification and prediction of failure types in flow control valves using supervised machine learning classification, and what is the impact on the failure diagnosis process?*

Sub-research questions

- What are the available failure types in Shell's data and can these be distinguished successfully in flow control valves using the available input of data?
- What features can be extracted from the model input to distinguish failures, and which of these are significant predictors for the output?
- Which supervised machine learning algorithm performs best on the classification and prediction of incident data using performance metrics?

1.6. KPIs

The goal of the research is to reduce the time required for failure diagnosis. Therefore, the time required for the classification and prediction is the first performance measure. Furthermore, the accuracy of the classification is also governed to determine the quality of the automatic failure diagnosis method.

Since multiple methods will be benchmarked, the performance of the classification will be tested using two performance indicators and cross-validation of the data. Afterwards, the optimal model and parameters for the data will be selected and tested on the case study introduced in Chapter 3.4. The two KPIs used are:

- Accuracy; the number of correctly classified failures to the total number of classifications.
- Log-loss; a measure of classification performance which takes the uncertainty of the prediction into account.

Furthermore, the results of the case study will be evaluated using the above KPIs and precision, which is the fraction of predictions that are relevant, recall, which is the fraction of relevant predictions correctly classified by the model.

1.7. Research scope

The scope of this research is flow control valves used as process equipment in Shell's refinery of Pernis. Valves that are not in operation, or are operating in manual mode are not taken into account. This research focusses on the diagnosis of failures which are known to suffer from failure behaviour. Therefore, normal behaviour is not applied as a failure type by the classification model. Furthermore, failure cases that happened before 2014 or after January 2020 are not adopted in the research.

Three assumptions have been made in this research to use the method plant-wide for all flow control valves. First of all, the specific valve type does not influence valve behaviour. There are variations in the valve type. However, we assume that these variations show similar failure behaviour. Second of all, the amount and type of material do not influence valve behaviour. The material or amount of material can vary per valve. However, we assume that this does not influence the failure behaviour. Finally, there are three different control modes: manual, automatic, and cascade. The behaviour where the valve is in manual mode is removed from the data. We assume that the other methods, automatic or cascade, do not influence the failure behaviour. More detailed information on the assumptions made can be found in Chapter 3.1.

1.8. Research plan

In this section, we describe the plan to answer the research questions successfully.
What are the available failure types in Shell's data and can these be distinguished successfully in flow control valves using the available input of data?
First of all, the control valve mechanisms used in the refinery of Pernis should be known in order to get a better overview of the input data that can be used to diagnose failures. A literature review has to be done

on common failures in control valves, and Shell's historical failure data should be analysed to determine the overlap. Furthermore, the current state of automatic failure diagnosis in literature needs to be known to choose a methodology suitable for the classification and prediction of failures in control valves.
What features can be extracted from the model input to distinguish failures, and which of these are significant predictors for the output?
Shell's control valve signal data and historical failure base need to be used to get a clear overview of the failures that can be distinguished using the data. Furthermore, literature research needs to be done on the extraction and selection methods that can be used for continuous time-series input data and categorical output data. Finally, the features need to be extracted and selected using the suitable methods to determine the feature overview that can be used as input for the supervised ML classification model. The set of features gathered from several plants will become the training set, and the set of features from a separate plant will become the case study or test set.
What supervised machine learning algorithm performs best on the classification and prediction of incident data using performance metrics?
Three ML classification algorithms have to be scored on the training set using the determined KPIs and a method such as cross-validation in order to define the optimal method for the failure classification and prediction given the available data set.
Afterwards, the optimal method will be tested on the unseen data of the case study plant to determine the quality of the classification and predictions of the model and the impact of the results on the current state of the failure diagnosis process.

1.9. Thesis outline

Chapter 2 of this report contains background information regarding control valves, failures in control valves, and the current state of failure diagnosis for process equipment in literature. Chapter 3 gives an overview of the data input and output of the model used for failure diagnosis in Shell's refinery in Pernis. Also, it gives an outline of the training set and case study. Chapter 4 describes the methodology used to classify and predict failures in flow control valves. Chapter 5 contains the experimental set-up used in the research, where more information concerning choosing the optimal model and parameter tuning of the training set is given using cross-validation. The results and discussion of the case study are also discussed in Chapter 5. Chapter 6 gives a conclusion on the research question and delivers the recommendations for further research.

2 Background information

Chapter 2 contains background information on various topics present in this research. Chapter 2.1 and 2.2 will provide more knowledge on the refining of oil and control valves, which are essential process equipment in refineries. Furthermore, Chapter 2.3 gives an overview of the existing literature on failure diagnosis for process equipment and control valves. Furthermore, our chosen method is proposed. Chapter 2.4 - 2.9 show more detail regarding the literature on our chosen method.

2.1. Refining of crude oil

Petroleum, a hydrocarbon mixture, appears in the Earth in the liquid, gaseous, and solid-state. However, when crude oil is extracted from the subsurface, it has no value to consumers, until, it is transformed into products. This process is called refining. The world's first refinery opened in Romania in 1856. Over decades the process has developed, primarily due to the rise internal combustion engine, to a mature industry. The second world war and the aviation industry required more variations in products allowing changes and growth to the refining of crude oil. This process requires several methods to convert crude oil into petroleum products. The first step of refining consists of boiling the crude oil to distillate into separate fractions. After this separation, the products are further modified by changing the size and structure of the material using cracking, reforming, and other conversion processes (Gary et al., 1988). Each refinery is unique depending on its location, and required input and output of products.

Shell's refinery in Pernis, which with a capacity of 404,000 barrels of crude oil per day is the largest integrated refinery and chemical manufacturing plant in Europe. The surface of the refinery, with 60 units processing crude oil, stretches out over 550 hectares. Due to the strategic location with a terminal on the harbour of Rotterdam products can be shipped overseas or transported inland with pipelines.

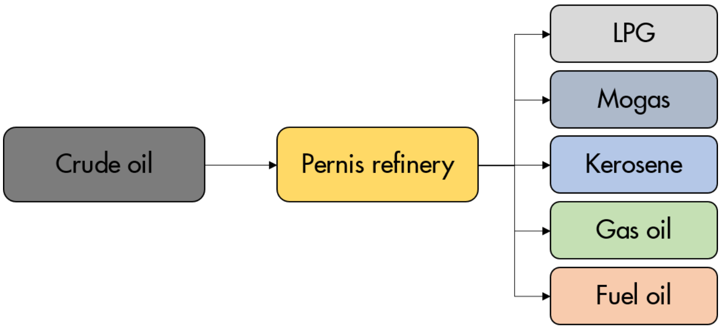


Figure 2.1: Overview of the refining process in Shell's refinery of Pernis. Crude oil enters the refinery, and after processing multiple products leave.

Refining crude oil into various products requires resilient machinery able to cope with rough materials. Equipment used in refining is often referred to as process equipment. Process equipment is often designed to perform a singular task, such as controlling flow, storage, or containing chemical reactions. Two categories can be distinguished in the machinery: fixed and rotating equipment. Fixed equipment consists of types of machinery that do not move, such as pipelines, furnaces, valves, storage tanks, and heat exchangers. Rotating equipment refers to machinery that rotates or is in motion, such as turbines, gearboxes, compressors, and pumps (Jones, 2006). This research focusses on the control valve. Spread out over the refinery of Pernis, approximately 5,000 control valves are currently in operation.

2.2. Control valve

The petrochemical refinery of Shell in Pernis uses a network of closed control loops to generate a finished product. The most common control element in these closed loops is the control valve. Control valves can modify the desired opening position by varying the controller output (OP) and thereby manipulate a measured process variable (PV), such as flow, as close as possible to the desired set point of the controller (SP). The process variables are measured by sensors in the system. Interactions with other control loops in the network cause disturbances influencing the process variable. Control valves can regulate the flow and are therefore capable of preserving the process variable at the desired set point even if the flow is affected by disturbances (Emerson Automation Solutions, 2017). Figure 2.2 gives an overview of the feedback loop used for control valves.

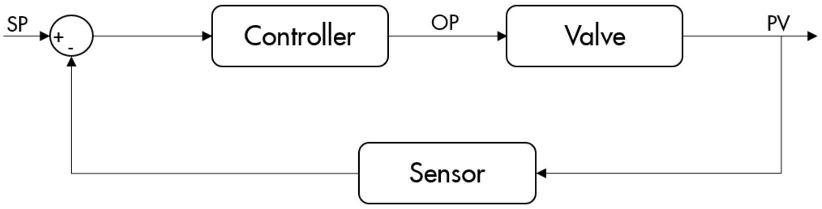


Figure 2.2: Control valve feedback loop.

The system controls various process variables such as flow, temperature, pressure, or in rare cases, other variables. The controller receives the data on the process variables from the sensor through the transmitter and calculates the error between the desired and actual state. This error is converted into a stroke value that will drive the current state of the process towards the set point. Control valves mostly have three basic components (Nwaoha et al., 2012): actuator, valve body sub-assembly, and accessories. The actuators in control valves are regularly pneumatic. With the help of a diaphragm and instrument air, the throttling element is positioned in the device. The valve body sub-assembly consists of the valve plug, valve seats, and the valve casing. The geometry of the plugin combination with the body determines the flow properties. The accessories of the control valves contain positioners, I/P transducers transforming current into pressure, and position sensors. Figure 2.3 shows a cross-sectional diagram of a pneumatic sliding stem control valve. New valve types have electronic positioners that also measure the actual opening position of the valve.

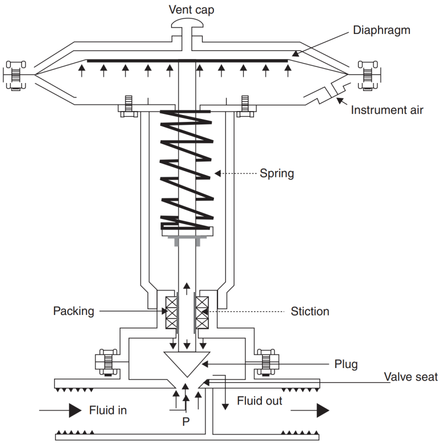


Figure 2.3: Schematic overview pneumatic control valve.

The flow characteristic is an essential property of valves, describing the relationship between the amount of liquid flowing through the valve and the opening of the valve. These characteristics depend on the shape of the valve seat. From literature, three essential characteristics can be defined: quick opening, linear, and equal percentage. Quick opening valves provide a substantial change in flow for minimal changes in the valve stroke when moving from a closed situation, which makes it favourable for batch or semi-continuous processes. In the equal percentage characteristic, the flow increases exponentially when increasing the valve stroke. Linear characteristic valve has a linear increase in flow when opening the valve.

Two types of pneumatic control valves are available: Air to Open (ATO) and Air to Close (ATC). Air to open control valves are held closed by the spring and air pressure is required by a control signal to open them. Applying more air pressure opens the control valve progressively. Air to close valves are held by the spring in the valve and require air pressure to move to a closed position. In the petrochemical refinery of Pernis, two types of control valves are implemented. Sliding stem valves and rotary valves depending on their location and application in the refining process.

2.2.1. Failure types

In literature, several failure types of control valves are discussed, ranging from mechanical issues on the valve to tuning issues of the controller. Table 2.1 gives an overview of failure types widely considered in literature with a short description (Bacci Di Capaci et al., 2013, Choudhury et al., 2004, Emerson Automation Solutions, 2017).

Failure type	Definition
Hysteresis	The difference between the valve position on the upstroke and its position on the downstroke at any given input signal.
Deadband	The range a measured signal can vary without initiating a response from the actuator, causing a backlash.
Stiction	Varying input preceded by a sudden abrupt jump, called the 'slip-jump', due to the static friction exceeding the dynamic friction.
Stuck	controller tries to open or close the valve, however, the outflow stays constant.
Packing leakage	leakage of process fluids caused by a loose or worn stem packing, for example, due to packing damage by cavitation.
Hunting	valve overshoot the target position after a change in the controller output.
Diaphragm rupture	leakage from the valve membrane contributing to a loss of efficiency of the valve action.
Blockage	change in internal dynamics of the valve due to variations of the forces opposed to the motion of the valve plug.

Table 2.1: Overview of failure types in control valves discussed in the literature.

This overview shows general malfunction types in control valves. However, this list can differ for refineries. Depending on the location and application of the valve, different failure modes can require an utterly distinct maintenance strategy. Some defect behaviours do not affect the performance significantly. Therefore, no immediate repair action is needed. However, other failure types could completely stop large parts of a plant possibly leading to extended downtime. Thus the early diagnosis of malfunctions can be a very effective method to classify failure types and thereby prevent critical issues.

2.3. Failure diagnosis in literature

The research area on fault detection and diagnosis contributes to the automation of discovering faults and diagnosing their causes in physical systems from raw data. Faults can be defined as an anomalous deviation from the normal behaviour in a system. Failures can frequently occur, due to deteriorating forces, after the machinery has reached a certain age. However, failures can also occur infrequently. In this case, the faults are unexpected, not trivial to detect, and hard to diagnose. Faults can cause major issues in operation plants and therefore should be dealt with carefully. The detection of failures relates to the indication that a fault is present in the system. Improving just the detection of failures can be done by fault classification or diagnosis. Fault classification is defined as the categorization of a fault into a popular category. After the fault is diagnosed, prior-knowledge on the failure type can be applied to deal with the failure (Pareti, 2010). Figure 2.4 shows a categorization of fault diagnosis methods. According to Wang et al. (2009), the application of failure diagnosis can be divided into two approaches: model-based and data-driven. The model-based method uses mathematical models to estimate the state and parameters of the system. The data-driven method focuses on high dimensional data and is applied to highlight important information from the available data. Within the data-driven fault diagnosis method is the knowledge-based approach, which applies a knowledge base composition and a set of qualitative models. The knowledge base composition includes process input and output variables, fault characteristics, and operational constraints and assessment criteria. Techniques used are cause-effect analysis, fault tree analysis, and rule and case-based reasoning.

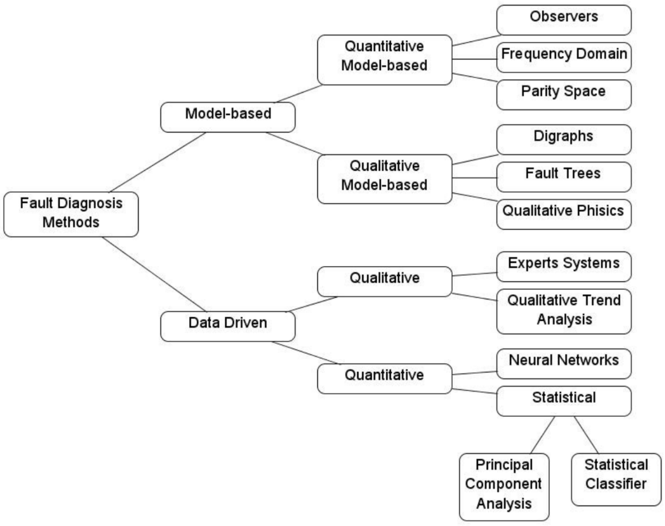


Figure 2.4: Overview of fault diagnosis methods split up into model-based and data-driven (Pareti, 2010).

2.3.1. Failure classification of process equipment

Failure diagnostics is one of the main challenges in predictive maintenance. Seo and Jun (2019) proposes a failure diagnosing method to estimate the state of equipment using clustering methods to extract abnormal patterns. The approach consists of a learning and a predicting stage. In the model, features are statistic summary values in a specific interval, such as average, standard deviation, maximum value, minimum value. The set of features over a specific interval is used in a feature vector. Clustering methods are used to find the patterns in the feature vectors. Based on these patterns, a value based on the Term Frequency-Inverse Document Frequency (TF-IDF) is gathered, that denotes the degree of occurrence of a pattern in a failure type set.

Xie et al. (2015) describes a method to detect failure events using sensor data. This is executed in three stages: data cleansing, feature extraction, and the application of machine learning tree classifiers. The features are extracted by obtaining the average, standard deviation, maximum, and minimum for a particular period. Finally, two supervised machine learning tree classifiers are used: Gradient Boosting Decision Tree (GBDT) and Random Forest (RF). The models obtain high accuracies. However, it is just used to detect failures not to classify them into failure types.

Predictive maintenance with an ML diagnostics framework for the aircraft industry is discussed by Adhikari et al. (2018). The diagnostics model consists of 5 stages: feature engineering, anomaly detection, enhanced fault isolation, fault identification, and prognostics. Firstly, the features are extracted, selected, and reduced using several methods. Finally, the anomalies are detected and classified using ML algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), and RF. The research of Santi et al. (2019) proposes a fault diagnosis classification model. The model consists of unsupervised clustering to group the failure types. Afterwards, an RF supervised machine learning classifier is used to improve the quality of the diagnosis.

2.3.2. Failure classification in control valves

Huang and Yu (2008) have presented a simple method that detects stick-slip failures in industrial control valves. Stick-slip failures are common in control valves and arise from changes of friction in the valve stem due to wear and tear. This method requires the actual valve travel signal and the valve travel setting signal assuming these are fault free. Under normal operating conditions, the valve stem moves smoothly corresponding the normal response. Due to malfunctions, caused by the stick-slippage, the travel of the stem varies in steps, where the speed of the stem is divided into moving and stationary states. When comparing the speed frequency at different speeds, a relationship between the mean and root mean square of the stem valve speed can be found. This indicates whether stick-slip occurs.

Huang et al. (2010) shows a trend analysis method to detect failures of actuators with digital positioner. The malfunctions analyzed are fully failure, offset and bias, change of gain, serious hysteresis, and stick-slip fault. In the analysis, three signals are used: valve travel setting signal (CV), actual valve travel (X), and flow rate through valve body (Q). Stick-slip can be detected using the method described by Huang and Yu (2008). The signals of the CV is compared to the spool opening and the relative flow rate in order to detect the failures.

Scali et al. (2011) has developed an integrated system that can give early diagnose of several anomalies such as stiction, change in valve dynamics, air leakage, and periodic disturbances. These failures cannot be distinguished by referring to classical loop variables. KPI's, such as Travel Deviation (TD) and Drive Signal (DS) are made available by the valve positioner and can identify the failures. Stiction can be detected using just the loop variables OP, SP, and PV. Oscillations will show up in loop variables PV and OP in the case of stiction. In the case of air leakage, a persistent difference is measured between the OP and the actual valve position (MV). Change in valve dynamics can be measured by OP by responses on step SP change. In the situation of changing internal valve dynamics, the time interval in which the OP is deviated from the MV, after a step SP change, increases when going into a steady state.

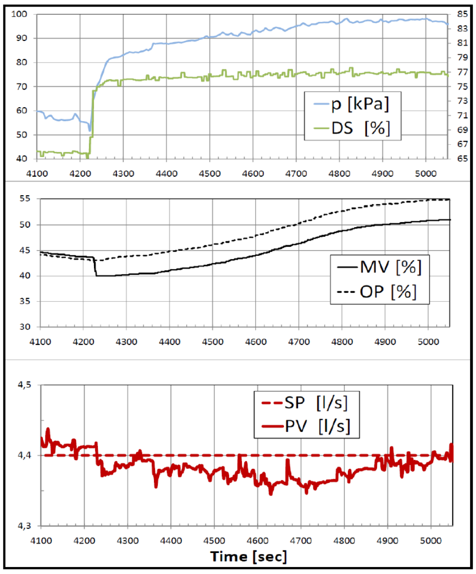


Figure 2.5: Example of an air leakage failure case using trend analysis in control valves in the research of Scali et al. (2011).

Bacci Di Capaci et al. (2013) introduces a performance monitoring system that based on trend analysis can detect different failures in control valves such as stick-slip, constant bias, change of valve gain, hysteresis, and stuck conditions. The monitoring systems use the OP, the PV, and the SP to extract the condition of the control valve. Suitable KPI's with threshold values have been defined and together with a developed logic, the valve status can be evaluated allowing the model to classify failures. The position of the valve stem MV can be determined with the drive signal and the pressure signal acting on the valve membrane. With MV and OP, the difference between the OP and the MV can be determined, further referenced to as the Travel Deviation (TD). The six KPI's used are all based on simple metrics of the TD. In the nominal case, the TD has a mean value close to zero with only small peaks in correspondence to changes in the SP. In the case of dynamic friction or jamming, similar results are shown, making it very hard to detect. Air leakage shows a clear decrease in the mean value of the TD. Clogging of I/P converter shows similar behaviour the air leakage conditions. Stiction shows persistent oscillations regardless of the constant value of the SP. Oscillations of the TD can also be caused by the presence of periodic disturbances. However, in this case, the amplitude peaks are small compared to stiction. Therefore, the two failures can be distinguished from one another.

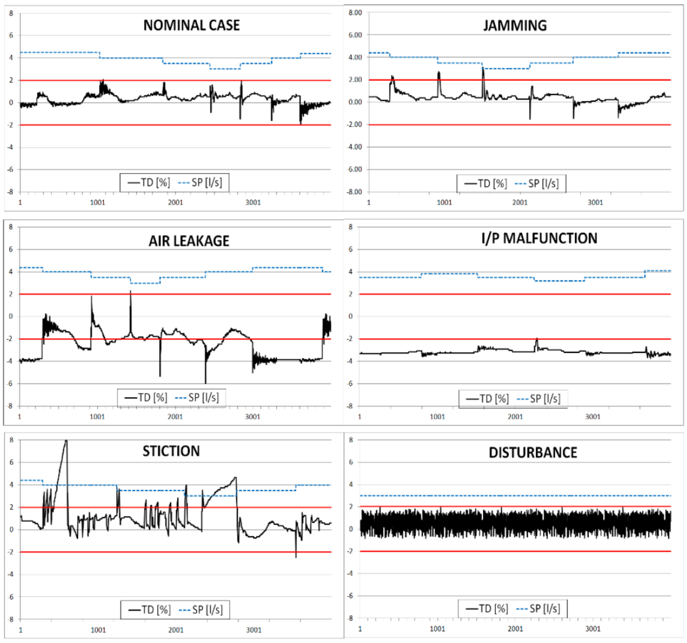


Figure 2.6: Example of failure diagnosis in control valves using trend analysis in the research of Bacci Di Capaci et al. (2013).

Mathur et al. (2019) demonstrates how Fault Tree Analysis (FTA) can be applied to process plants using a statistical software package in R. The research contains a monitoring system that predicts the likelihood of events in terms of quantitative methods. Four stages are used in the process. In stage 1, the hazards or failures are classified. Stage 2 identifies the factors causing the failure or hazard. Stage 3 localizes the factor places within the plant. Stage 4 connects all the events and creates the tree specifying the hazard or failure. The fault tree, in figure 5, shows a visualization of failure events or hazards of the control valve. Using tree metrics such as the Mean-Time-To-Failure (MTTF) and the Mean-Time-To-Repair (MTTR) can be estimated.

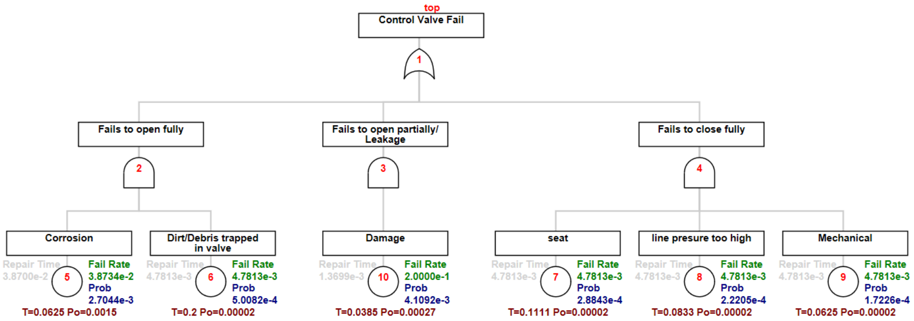


Figure 2.7: Diagnosis using a FTA package to distinguish failure behaviour in the research of Mathur et al. (2019).

Syfert et al. (2003) describes the actuator benchmark for fault diagnosis studies using the framework 'DAMADICS'. The purpose is to be able to benchmark a wide range of fault detection and isolation methods (FDI). Benchmarks are real or virtual sets of standard measures that allow the evaluation of the research approaches, algorithms and methods. The primary goal is to examine the KPI's of the algorithms prior to industrial application. For DAMADICS, two benchmarks have been created: model-based benchmark, and data-based benchmark. The model-based benchmark focusses on the simulation of the actuator's behaviour in normal and failure states. The data-based benchmark delivers process data from the evaporation station and steam boiler of a sugar factory. Artificial failure sets are introduced in the process. Based on the DAMADICS benchmark, several failure detection methods have been proposed, such as Calado et al. (2006). Marciniak et al. (2003) describes a pattern recognition approach to diagnose faults using the DAMADICS benchmark. In this method of automatic fault diagnosis based on pattern recognition, two steps can be defined: symptom extraction and the actual diagnostic task. In the research, after the extraction of a set of features, a decision space is defined within the state of the system. The states would contain normal behaviour, fault 1, fault 2 etc. The classification method can be defined as a mapping from the process measurements space to the decision space, which can be approximated with a neural network for feature extraction and a fuzzy classifier for decision making. The paper distinguishes the following faults: restriction of rod movement, leakage in the bypass valve, and faulty sensor measurements. The accuracy that is achieved, using the leave one out cross-validation, is 89%. Research by Prabakaran et al. (2013) is another paper on the fault diagnosis in process control valves using artificial neural networks.

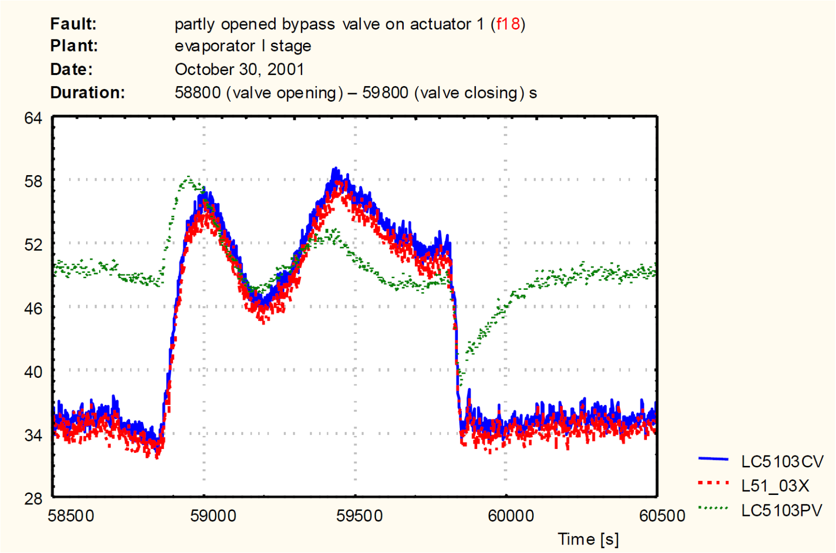


Figure 2.8: Example of the simulation of a 'party opened bypass valve' failure using the DAMADICS benchmark (Syfert et al., 2003).

Trunzer et al. (2018) contributes with a classification table in which expert knowledge on failure modes, underlying parameters and detection features are summarized to reduce significant losses of production due to unplanned downtime's. The research shows the usefulness of combining expert knowledge and a data-driven analysis on pressure regulated control valves. Four requirements have been set-up to reduce the complexity of the input variables. Requirement 1: failure modes must be identified and described by experts. The defined failure modes must be distinguishable and describe a specific effect. Requirement 2: The causal mechanisms for each failure mode should be identified by experts. However, in contrast to the failure modes, the causal mechanisms do not have to be distinct. Requirement 3: all relevant influencing parameters must be defined by experts. Requirement 4: detection features and fault detection models are developed based on the influencing parameters.

2.3.3. Conclusion

Based on the findings in the literature, a suitable method can be determined for this research. The failure diagnosis method is chosen based on the novelty of the method in literature when applied to control valves and on four criteria determined with Shell engineers. The four criteria are:

1. Interpretability of the model by the user.
2. Generic model for the whole plant.
3. Robustness when working with small sample size.
4. Risk of overfitting.

In order to make a proper decision on the optimal method for failure diagnosis knowledge of domain experts is gathered. Interviews are done in order to assign scores, ranging from 1 (bad) to 5 (good), for every criterion to the corresponding method. The full results of the domain expert survey can be found in Appendix B.1. However, a summary for every method in literature is shown in Table 2.2. The methods in the literature that cannot be applied to the available data obtain a score of 1 on all four criteria.

Title	Author	Method	Control valves	Interpretability	Generic model	Robust with small set	Risk of overfitting	Applicable to the data
DAMADICS failure diagnosis	Parbakaran et al. (2013)	Neural network	Yes	1.2	4.0	1.2	2.2	Yes
Stick-slip detection	Huang and Yu (2008)	Trend analysis	Yes	4.2	1.6	3.2	3.4	Yes
Failure detection method	Huang et al. (2010)	Trend analysis	Yes	4.2	1.6	3.2	3.4	Yes
Characterization and diagnosis of failures	Scali et al. (2011)	Trend analysis	Yes	4.2	1.6	3.2	3.4	No
Performance monitoring system to detect failure types	Bacci di Capaci et al. (2013)	Trend analysis	Yes	4.2	1.6	3.2	3.4	No
Fault tree analysis applied to process plants	Mathur et al. (2019)	Fault tree analysis	Yes	1.0	1.0	1.0	1.0	No
Failure diagnosis approach using the DAMADICS benchmark	Marciniak et al. (2003)	Neural network	Yes	1.2	4	1.2	2.2	No
Classification table using expert kwowledge for fault diagnosis	Trunzer et al. (2018)	Trend analysis	Yes	4.2	1.6	3.2	3.4	Yes
Failure diagnostics in condition based monitoring	Seo et al. (2019)	Feature engineering and supervised machine learning	No	3.8	4.4	3.6	3.6	Yes
Failure event detection using sensor data	Xie et al. (2015)	Feature engineering and supervised machine learning	No	3.8	4.4	3.6	3.6	Yes
Fault diagnosis classification	Santi et al. (2019)	Unsupervised clustering	No	1.0	1.0	1.0	1.0	No
ML diagnostistics framework for the aircraft industry	Adhikari et al. (2018)	Feature engineering and supervised machine learning	No	3.8	4.4	3.6	3.6	Yes

Table 2.2: Literature overview on failure diagnosis methods for process equipment.

The total scores of the methods in the literature that apply to the available data from Shell's refinery in Pernis, resulting from the domain experts scores from Appendix B.1. can be found in table 2.3.

Method	Score
Trend analysis	12.4
Neural network	8.6
Feature engineering & ML classification	15.4

Table 2.3: Total scores of methods for failure diagnosis that are applicable to the available data from Shell's refinery in Pernis.

The method scoring best is 'feature engineering and supervised machine learning classification', and therefore this method has been chosen for the automatic failure diagnosis model in control valves. This approach has not yet been used for control valves, and therefore will be novel to the literature on automatic failure diagnosis. Figure 2.9 is a diagram showing the stepwise approach of the methodology from input to output data.

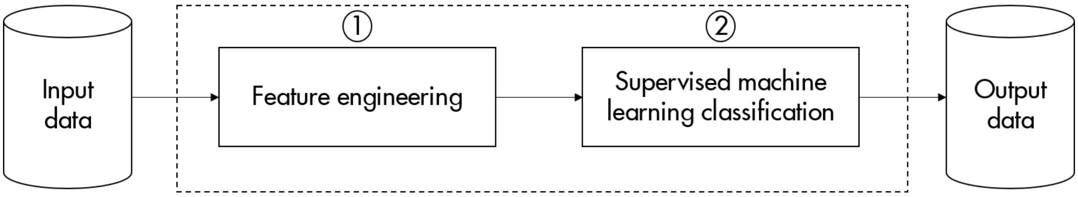


Figure 2.9: Stepwise diagram of the feature engineering and supervised machine learning classification process for failure diagnosis in control valves.

2.4. Feature engineering

Machine learning models predict and derive insights by fitting mathematical models to data using features as input. Raw data can be numerically represented by a feature, which functions as the link between data and model. Feature engineering is the refining of raw data by removing unnecessary information and transforming data into formats that are suitable for a machine learning model. Using the correct features can reduce the complexity of the data, and thereby improve the quality of the output. Figure 2.10 shows the workflow of machine learning, where feature engineering builds and cleans features that represent the raw data (Zheng and Casari, 2018).

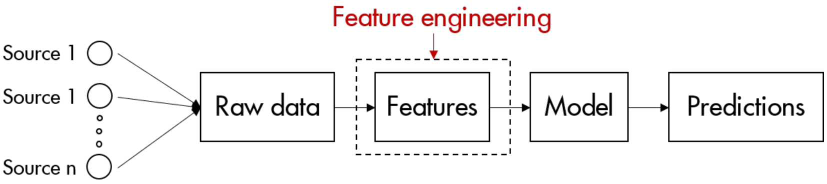


Figure 2.10: Machine learning workflow. Feature engineering transforms the raw data into features that can be used in the machine learning model.

The number of features used in the model has a strong influence on the performance of the model. If too few informative features are built, the model will most likely not represent the raw data well resulting in bad performance. This is known as underfitting. However, too many can make the model more problematic to train and complex, ultimately resulting in lower performance. This is known as overfitting. Therefore, feature engineering can be separated into two stages: feature extraction and feature selection.

2.5. Feature extraction

The extraction of features refers to transforming data into formats that are suitable for a machine learning model. In order to obtain features that can distinguish failure behaviour, quantitative and qualitative research on the data should be done. Qualitative reasoning is used to design features that can be used to separate failure behaviour. Quantitative reasoning functions as a filter on the features that determine which are good predictors this is called feature selection. Figure 2.11 gives examples of features which can be created from time-series data.

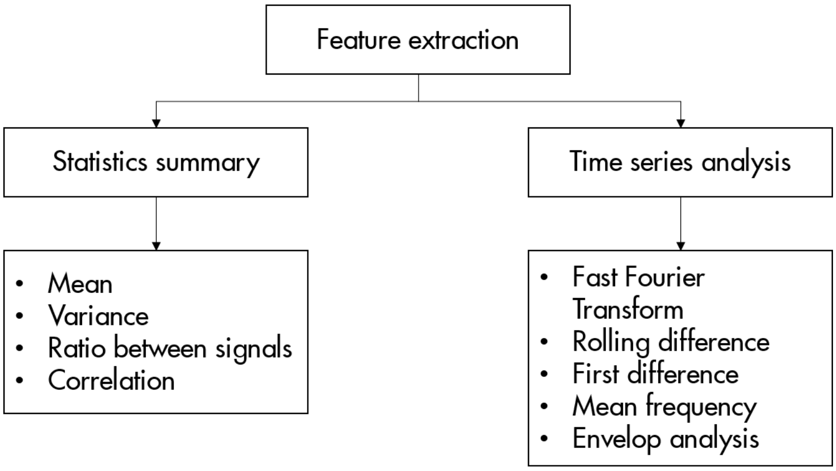


Figure 2.11: The extraction of features that represent the raw time-series data. Left: features that summarize the statistics of the signal. Right: features based on time-series analysis methods.

The category on the left contains a summary statistics of the signal, for example, the mean of the signal over a certain period can be used as a feature. If, for one of the failure behaviours, the controller output shows an extremely high mean (around 100%), while for the other failure behaviour categories it shows an intermediate mean (around 50%), the mean of the controller output signal can be an essential feature. The category on the right contains time-series analysis methods widely discussed for feature extraction. For example, Power

Spectral densities resulting from Fourier transforms can be used to separate signals built up from different frequencies.

2.6. Feature selection

The selection of features refers to removing unnecessary features before running the model to improve its quality in terms of run-time and performance. Statistical tests can be applied to test whether features are significant predictors for the output. Several statistical tests are available depending on the data types of the input and the output. A hypothesis test can be used to determine if the features are significant predictors. The null hypothesis in general claims that the feature is not a significant predictor for the output of the model. If the p-value of the test is below a threshold, often 0.05, we reject the null and conclude that the feature is significant. Figure 2.12 shows various tests from the literature that have different requirements on the types of data and distributions from the input and output data.

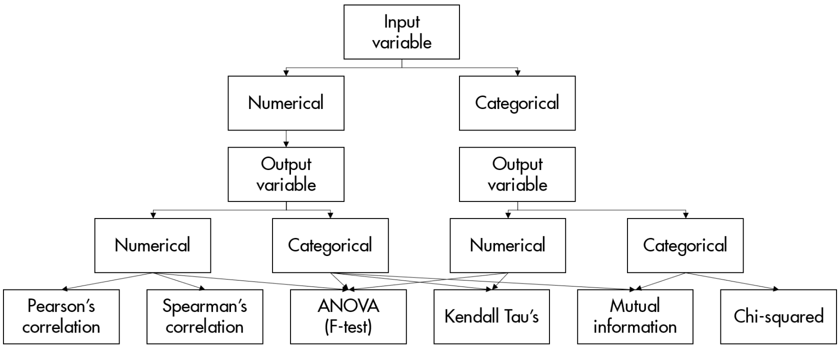


Figure 2.12: Methods used for feature selection depending on the input and output data type.

Feature selection methods are of often classified into three categories (Vergara and Estévez, 2014): wrapper, embedded, and filter methods. Wrapper methods apply predictive models to score feature subsets. Wrapper methods are computationally very heavy since it requires training a model for each subset of features. Embedded methods are performing feature selection in the model construction process, such as a lasso regression algorithm. Filter methods, use metrics to determine the predictive power of each feature subset for the output. Filter methods are robust against over-fitting; however, do not choose the type of predictive model and therefore, might result in lower model performance. This research considers filter methods since the model performance should not be influenced by feature selection when using different machine learning algorithms.

2.6.1. Pearson's correlation

Pearson's correlation measures the strength and direction of linear relationships between pairs of continuous variables by generating a coefficient r. This correlation measurement is often used to determine correlations among pairs of variables, or to determine correlations within and between sets of variables (Minckler, 1995). The correlation coefficient is calculated using Equation 2.1.

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)} \cdot \sqrt{\text{var}(y)}} \tag{2.1}$$

where r_{xy} is the correlation coefficient between x and y, $\text{cov}(x,y)$ is the covariance between x and y, $\text{var}(x)$ is the variance of x, and $\text{var}(y)$ is the variance of y. In the case of feature selection x is the input feature or variable, and y the output variable. The correlation coefficient has a value between [-1,1]. Where -1 is a negative linear relationship or negative correlation, 0 is no relationship or no correlation, and 1 is a positive linear relationship or positive correlation. The data has to meet several requirements to be able to use Pearson's correlation (Kent State University, 2017):

- input and output should be continuous variables.
- Linear relationship between the input and output.
- Independence between observations.

- Data should be normally distributed, therefore, the Pearson's correlation does not perform well on data containing outliers.

2.6.2. Spearman's correlation

Spearman's correlation or rank-order correlation coefficient is a non-parametric measure of the strength and direction of the relationship of two variables on an ordinal scale. Data on an ordinal scale means that it is ranked, for example, when the data varies between low and high. This test determines the statistical dependence between the two ranked variables by using a possibly non-linear monotonic function. A monotonic relationship works as follows: if one variable increases, the other also increases, or if one variable increases the other decreases. The correlation coefficient ρ_{xy} gets a value between [-1,1], where for -1, the data follows a decreasing monotonic relationship and for +1 an increasing monotonic relationship. Spearman's correlation is often used when the requirements for determining Pearson's correlation are not met. The requirements for the data are (Aerd Statistics, 2018):

- Data should be on an ordinal, interval or ratio scale.

2.6.3. ANOVA (F-test)

The ANOVA (F-test) measures the degree of linear dependency between two random variables and can be used to test whether a feature is a significant predictor to the output. The test calculates an F-value by comparing the variability between and within the groups. Large F-value, emerging from a large distance between the means of the variables, corresponds to a good predictor for the output. Since for the ANOVA test, the variance within and between the groups is required the Total Sum of Squares is applied between the feature and predictor, and within the feature and predictor. Furthermore, degrees of freedom are used, which refers to the maximum number of logically independent values that have the freedom to vary. The data requirements for the test are:

- Categorical data should be independent.
- Independence between observations.

2.6.4. Kendall Tau

Kendall Tau is a non-parametric measurement of the relation between two ranked variables. Ranked means that the data is on an ordinal scale. The Tau coefficient returns a value between [-1,1], where 0 is no relationship, and 1 is a perfect relationship. The Kendall Tau correlation coefficient can be calculated using the following formula 2.2:

$$\tau = \frac{C-D}{C+D} \tag{2.2}$$

Where C and D refer the concordant and discordant pairs to describe the relations between pairs of observations. If the direction of two pairs of observation is the same, for example, both positive or negative, the pairs are concordant. Discordant pairs of observations have exactly the opposite characteristics. The requirements of the data using Kendall Tau's test are (Statistics Solutions, 2019):

- Data input can not be negative.
- Data must be ranked.

2.6.5. Mutual information

Mutual information, widely used measure for non-linear dependency between variables, quantifies the amount of information that one variable obtains from the other. The Mutual Information value can vary between [0,1], where 0 is completely unrelated or independent, and 1 is completely related or dependent. Mutual information is calculated using formula 2.3:

$$I(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2.3}$$

Where p(x,y) is the joint probability density function of the input feature and output, and p(x) and p(y) are the marginal density functions of the input feature and output. The required values can be estimated using the k-nearest neighbours method, according to work by Ross (2014).

2.6.6. Chi-squared

The Chi-squared test is applied in feature selection to test the features which are highly dependent on the output. Chi-squared compares the observed output to the expected output, using Equation 2.4 (Van Hulse et al., 2009):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{2.4}$$

Where c is the number of classes in the output, r is the number of values of the input feature, $O_{i,j}$ are the observed values, and $E_{i,j}$ are the expected values. Independent features, thus bad predictors, obtain a small Chi-squared value. Features that are good predictors gain a high Chi-squared value. Using Chi-squared as a feature selection filter method requires:

- Data input and output should be categorical.

2.6.7. Conclusion

Several feature selection filter methods are available depending on the data type used as input and output for the model. Table 2.4 shows the discussed statistical tests, and their most important requirements for the data input and output.

Method	Important requirements of the data
Pearson's correlation	Input: continuous, output: continuous, linear, independent
Spearman's correlation	Input: continuous, output: continuous, non-linear (monotonic), ranked (ordinal)
ANOVA (F-test)	Input: continuous, output: categorical, linear, independent
Kendall Tau	Input: continuous, output: categorical, non-linear, ranked (ordinal), non-negative
Mutual information	Input: continuous, output: categorical, non-linear
Chi-squared	Input: categorical, output: categorical, non-linear

Table 2.4: Summary table of the statistical tests available for feature selection in literature and their most important requirements.

2.7. Machine learning

Humans are prone to making mistakes between various features during analyses or cannot always find relationships required to solve problems, especially when the dimensionality increases. It is therefore beneficial for computers to learn from their experiences and apply their knowledge to improve the efficiency and performance of systems. The term machine learning (ML) refers to the automated detection of meaningful patterns in data. Nowadays, machine learning is a critical aspect in various industries and is widely applied in research. This chapter will give a clear definition of ML, discuss its current strategies applied, and more in-depth information on ML classification algorithms.

2.8. Machine learning strategies

Machine Learning is the science of allowing computers to learn from data. The first definition of ML can be derived from IBM research by Samuel (1969):

> "Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed."

Arthur Samuel applied ML on a computer that learned how to play a game of checkers. Due to the lack of computing power, the computer was not able to outplay a checker master. However, its playing ability drastically improved. Over time the applications using ML greatly increased and therefore a more recent definition by Mitchel (1997):

> "A computer program is said to learn from experience E concerning some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

ML can be used in a large range of applications, and especially it works well for problems in which: solutions require much tuning by hand, complex problems beyond human capabilities, and fluctuating environments (Geron, 2017). Several types of ML algorithms can be applied in different situations. These types can be divided into three categories:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

In Supervised learning, the instances in the data include the desired outputs or known labels. Therefore if the model is trained, it can make predictions of instances using new data points. Two tasks in supervised ML are classification and regression. Classification is the process of predicting the class of given data points. Classes are referred to as targets, labels or categories. Regression is predicting a numerical value, such as the price of a product, based on available features. There are models which can be used for regression and classification. Examples of supervised ML algorithms are Linear regression, Logistic regression, Support Vector Machines (SVMs), Naive Bayes (NB), Decision trees, Random Forests (RFs).

In unsupervised learning, the instances in the data do not include the desired outputs or known labels. Unsupervised ML algorithms can be applied to discover classes of items that can be distinguished based on the features. Essential types of unsupervised learning algorithms are Clustering algorithms, visualization and dimensionality reduction algorithms, and anomaly detection algorithms.

In reinforcement, learning is very different from supervised and unsupervised learning consisting of: the learning system called an agent, and the external trainer called the environment (Kotsiantis, 2007). The agent is not told which actions it should take. However, it should discover it on its own by getting rewards or penalties in return. Therefore the agent will learn by itself what the best strategy called a policy is. Examples of reinforcement learning projects are robots learning how to walk by themselves.

The application of ML is rather challenging due to various risks that can lead to entirely wrong conclusions based on the results of the model. The most common problems are around bad data and models. Problems with data are mostly around the insufficient quantity, non-representative training data, poor-quality data, and irrelevant features. Model issues are often related to under or overfitting the model. When the model is performing very well on the training data but not so well on the test data, it does not generalize well, which means the user has been overfitting the model. Underfitting is the opposite of overfitting. Two definitions related to overfitting are (James et al., 2013): bias and variance. Bias refers to the error that is introduced by approximating a very complicated problem with a straightforward model. High bias models are easy to train but less flexible, e.g. as linear models. Variance is the amount by the estimate of which the target function would change if another training set were used. High variance models are often strongly influenced by specifics of the training data, such as decision trees. In ML, there is always the trade-off between bias and variance, since increasing one will reduce the other, and both of them influence the error of the model.

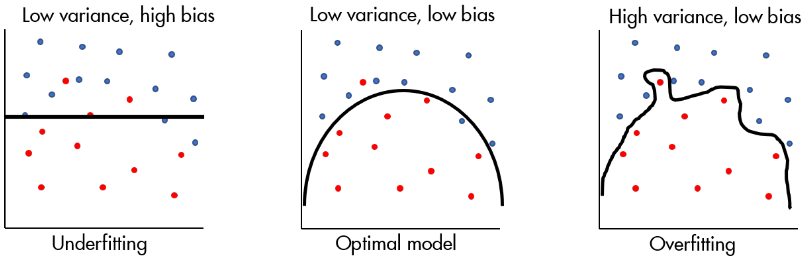


Figure 2.13: Bias, variance trade-off in ML.

When quantifying the error of the predictions made by the regression model in ML often the mean squared error (MSE) is used. This MSE is created using the variance and the bias of the predictions, with the following Equation 2.5:

$$\mathrm{MSE} = \mathrm{Variance} + (\mathrm{Bias})^2 \tag{2.5}$$

With a simple model, the variance is low and but bias high. In this stage, the model is underfitting. When the bias of a model is reduced, which increases the complexity of the model, the error reduces. However, at the same time, the variance increases. After improving the complexity, and the variance up to a certain level, the error on the test set starts to increase again. The error on the training set becomes low at this stage, however, when predicting new data, the error will be substantial. In this stage, the model is overfitting.

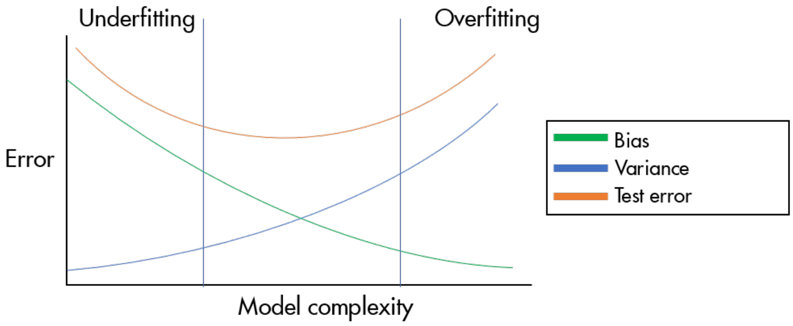


Figure 2.14: Error on the test set, when the complexity of the model increases.

2.9. Supervised Machine Learning classification

The prediction of a qualitative response or categorical variable for observation can be referred to as classification. The goal is to build a model of the distribution of class labels in terms of predictor features. The result is a classifier being able to predict unknown class labels of test instances using the values of the predictor features (Kotsiantis, 2007). Classification requires a set of training data $(x_1, y_1), ..., (x_n, y_n)$ which is used to train the model. The classification model should not only perform well on the training data, but also on the test data that has not been used to train the model (James et al., 2013). An example of the supervised ML classification process, where a classifier is created to learn from a set of rules in the training set and is used to generalize from new instances, is shown in figure 2.15. First of all, data sets have to be collected for the problem. Expert knowledge if available, can be used to determine which data sets are relevant and informative. If not available methods such as ’brute-force’ can be used, these consider everything and attempt to isolate the informative features. Second of all, the features should be prepared and selected to determine the training set. The preparation of data often consists of handling missing data and assuring the accuracy of data. In feature selection, irrelevant and redundant features are detached to reduce the dimensionality and size of the data. Feature selection increases the speed and quality of the model (Yu and Liu, 2004). After the training set has been defined a machine learning algorithms should be chosen. In this critical step, algorithms should be evaluated before they can be implemented as a well-performing model. The evaluation of the model and comparison with other algorithms can be based on a different statistical method using various techniques. A common metric to evaluate and compare models is the accuracy, which is the percentage of correct predictions divided by the total number of predictions. There are various techniques used to calculate the evaluation metrics. Preparing the training set to optimal performance on the data can be done using cross-validation. In cross-validation, the training set is split into mutually exclusive and equal-sized sets, and for each set, the model is trained. The average error is the estimate of how well the model is performing on the training set. The k-fold cross-validation is explained in more detail in Chapter 4.5. Various factors can have a large impact on the quality of the classification, such as not enough data available, too high dimensionality, inappropriate algorithm chosen, parameter tuning is needed, or an imbalanced data set. More on model evaluation in chapter 4.6. Depending on the algorithm chosen parameters can be tuned to increase the model's performance.

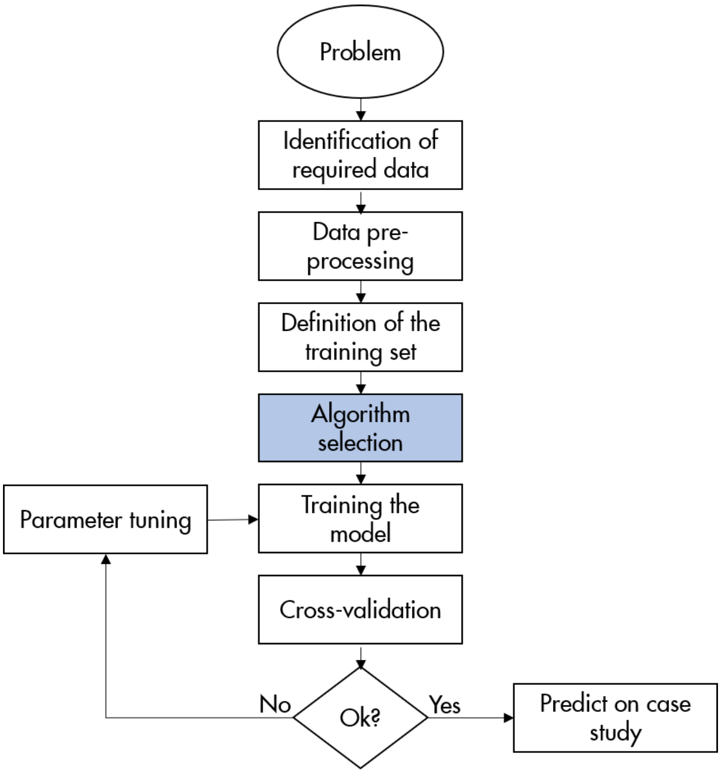


Figure 2.15: The process of applying supervised ML classification to a problem

2.9.1. Linear regression

Linear regression models the relationships between predictors using a linear approach. Linear regression consists of two steps. In the first step an assumption of a linear model is made with coefficients β_0 and β_1 in the following form (James et al., 2013):

$$f(X) = \beta_0 + \beta_1 X_1..\beta_n X_n \tag{2.6}$$

Where n is the number of observations of the predictor. In the second step, the model is trained using the ordinary least squares method to redefine the function of Equation 2.6. The function plot of linear regression, for input X and output y, is shown in figure 2.16.

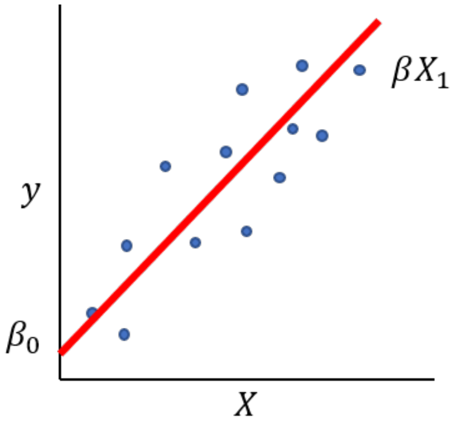


Figure 2.16: Example of linear regression plot for data input X and output y.

Linear regression is very useful for ML regression problems, however, for classification it is not very suitable due to the fact that the predicted value is continuous, not a probability.

2.9.2. Logistic regression

Logistic regression is a supervised ML algorithm that can be used for classification. The model calculates the probability between 0 and 1 that a specific value of the predictor belongs to a particular class or category using a logistic function. The model determines the intercept and regression coefficients using a method called the maximum likelihood. Since logistic regression uses a logistic function to calculate the probabilities resulting in an S-curve, which can be found for binary classification in Figure 2.17. Logistic regression can be used for binary classification, or multi-class classification using, for example, one-versus-rest or entropy lossJames et al. (2013).

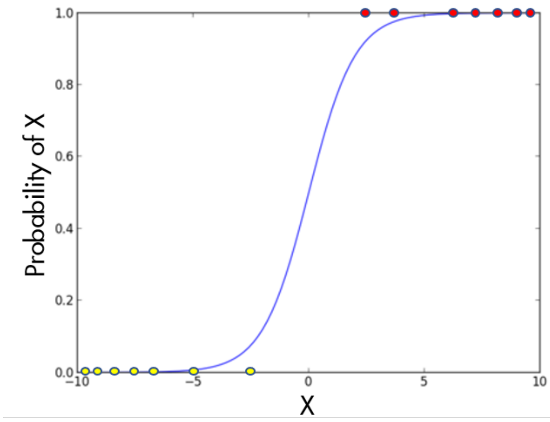


Figure 2.17: Example of a logistic regression probability plot.

Logistic regression is known to be a fairly simple algorithm. Therefore, it is highly interpretable, requires a low number of computational resources, and is easy to tune. The disadvantage of the model is that the decision surface can not be non-linear.

2.9.3. Support-Vector Machines

Support-Vector Machines (SVMs) are supervised machine learning models that can be used in regression and classification problems. SVMs construct a set of hyperplanes to separate two classes. The hyperplane is chosen so that it maximizes the distance from the plane to the nearest data point on each side. The hyperplane can have a different dimension depending on the kernel function chosen (Cortes and Vapnik, 1994). SVMs can use hard or soft margins, depending on the application of the model. Figure 2.18 shows an SVM classification, with input X and output y. The hyperplane, using a linear kernel function, is the dark line in the figure and maximizes the margin between the nearest data points on each side.

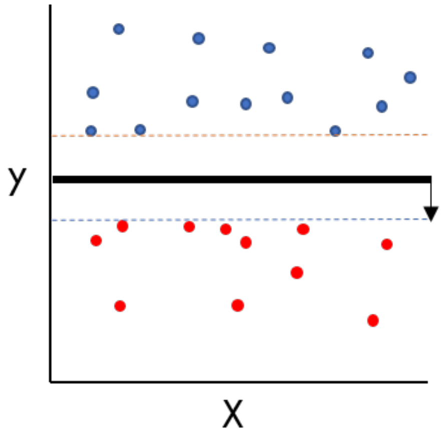


Figure 2.18: Example of a SVM classifier.

SVM's are very effective when classes are seperable, and work well in higher dimensions. However, SVM's often require a long time to process and have problems with cases where there is overlap in the data. Moreover, it can be very hard to determine the appropriate kernel function.

2.9.4. Naive Bayes

Naive Bayes (NB) classification models in ML are probabilistic classifiers that apply Bayes theorem. Bayes theorem calculates the posterior probability, using the likelihood, class prior probability, and the predictor prior probability. This means that the probability of the observation being in a specific class given the attribute value x (Zhang, 2004). In formula form Bayes theorem is given in Equation 2.7:

$$p(\text{class}|x) = \frac{p(x|\text{class})\,p(\text{class})}{p(x)} \tag{2.7}$$

NB classifiers can apply different distributions of p(x|class), such as the Bernoulli or Gaussian distribution. NB classification models are known to be highly interpretable, work fast, and perform well with a small dataset. A strong requirement for the NB classifier, however, is that the data should be independent.

2.9.5. Decision tree

Decision tree's (DTs) in ML are widely used for regression and classification problems to predict based on observations. The DT consists of decision nodes, leaf nodes, and branches. In classification, DTs are built using binary recursive partitioning, which is a process that splits data into partitions. These partitions are later on split up into branches. This process repeats until the leaf nodes are reached, and classification has been made for the observations. Several metrics can be used to construct the decision trees, and examples are Gini impurity, Information gain, and Variance reduction. Figure 2.19 shows an example of a decision tree with decision nodes, leaf nodes, and branches to connect them. The advantages of using decision trees are: highly interpretable, able to use categorical and numerical input, white-box model, and no feature selection is required. However, DTs have some limitations, such as: over complex tree's that do not generalize well and a small change in data can produce a completely different tree.

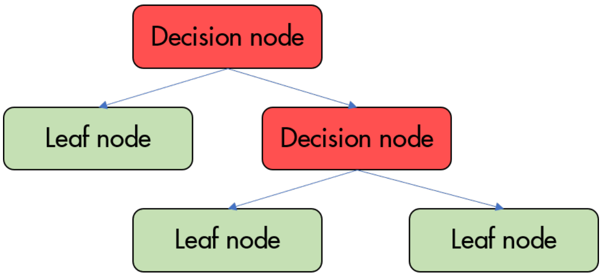


Figure 2.19: Example of a DT classifier.

2.9.6. Random Forest

The Random Forest (RF) classifier in ML uses a set of tree predictors. First of all, random samples from the data are drawn. Second of all, a DT is created for every sample to obtain a predicted classification result from each tree. Finally, the mode of the classification results, which correspond to the classification result that has occurred the most, is chosen as final prediction output (Breiman, 2001). The RF algorithm has several advantages: highly accurate due to the number of decision trees participating, and it cancels out overfitting due to the mode of classifications. However, RF models can be harder to interpret than DTs due to the lack of visualization, and it can be relatively slow compared to other models.

2.9.7. Artificial Neural Network

Artificial Neural Networks (ANNs) are widely used algorithms in the area of supervised ML classification. An ANN is a layered network consisting of an input layer, intermediate or hidden layer(s), and an output layer. Weights are associated with the connections between the layers, which are iteratively updated in the learning process increases the performance of the model. There are two categories of ANNs (Bala and Kumar, 2017): feed-forward and recurrent networks. In feed-forward models, the networks do not form a cycle, while in recurrent networks they do. Figure 2.20 shows an example of the architecture of a feed-forward ANN.

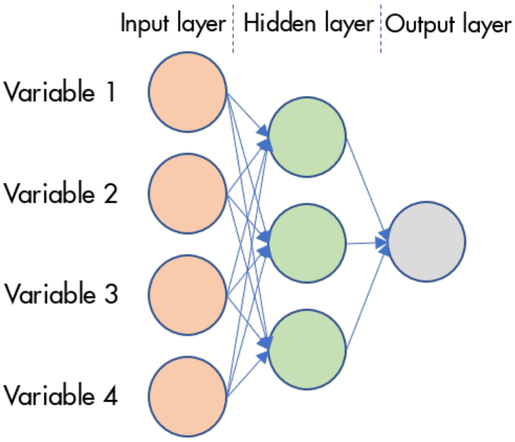


Figure 2.20: Example of a feed-forward ANN.

Using ANNs has several advantages for supervised ML classification: able to handle large amounts of data, able to detect non-linear relationships, and it can solve very complex models. However, there are also disadvantages: training of the model is time-consuming, and it is difficult for humans to interpret the meaning behind weights and hidden units.

2.9.8. Conclusion

Different algorithms are available, which all have different advantages and disadvantages depending on the data and application of the model. Table 2.5 shows an overview of the discussed algorithms and their advantages and disadvantages for classification.

ML algorithm	Advantages for classification	Disadvantages for classification
Linear regression	Easy to interpret	Not suitable for classification
Logistic regression	Simple, easy to interpret, low computational resources	Only linear decision boundary
SVM	Work well with higher dimensions	No overlap in data, hard to choose appropriate kernel function
Naive Bayes	Highly interpretable, fat, performs well on small dataset	Data has to be independent
Decision Tree	Highly interpretable, no feature selection needed	Over complex tree's, small change in data can completely change the tree
Random Forest	High accuracy, cancels out overfitting	Harder to interpret than DT, slow
ANN	Large amounts of data, non-linear data, able to solve very complex model	Very hard to understand, training is time consuming

Table 2.5: Summary table of supervised ML classifiers in literature and their advantages and disadvantages.

2.10. Conclusion

Typically, oil refineries are large-scaled assets processing extremely rough materials, which requires high-end equipment. Control valves have an essential role in refining, due to their ability to regulate processes based on measured target variables, such as flow, pressure, or temperature. However, due to the harsh material flowing through the valves, failures are not exceptional.

The classification and prediction of failure types in control valves are required for Shell's refinery in Pernis. Automatic failure diagnosis in process equipment and control valves has widely been discussed in the literature. However, many of these techniques are applied to simulation data with perfect conditions, not on technical data. The applied methods are ranging from trend analysis, deep-learning, clustering techniques, to supervised machine learning classification. In order to determine the optimal method for this problem, several criteria on the method have been surveyed by domain experts. The appropriate method is chosen based on these survey results and the novelty of the approach in literature. Due to the high interpretability, ability to use the model generically for the whole asset, and robustness to work with a small sample size 'feature engineering and supervised machine learning' has been chosen. This method allows us to create features based on qualitative reasoning on failure cases. Therefore, the knowledge of failures from experienced engineers can be used to perform automatic failure classification and prediction properly.

Feature engineering consists of two steps: the extraction of features that describe the signature of the raw data, and the cleaning of features using selection methods to optimize the performance of the model. In the extraction step, several approaches have been discussed around using statistical summaries or time-series analysis on the raw data. In the selection step, numerous statistical tests are available to determine the optimal predictors for the ML classifier. Examples of tests are correlations, ANOVA, Mutual information, and Chi-squared. Depending on the type of data input and output, the optimal tests can be chosen for failure diagnosis in control valves.

There is a very diverse offer of supervised ML classification algorithms available, which all work optimally for different data types and applications. This research describes the algorithms: linear regression, logistic regression, SVM, NB, tree-based methods, and ANN. Several criteria and the type of data input and output of the control valves are used to determine the optimal ML algorithms for failure diagnosis in control valves. In chapter 4, the methods will be chosen and highlighted for feature selection and ML classification.

3 Data

Chapter 3 describes the flow of data used in the automatic failure diagnosis model for control valves. Chapter 3.1 describes the input of the model consisting of the signals containing information on the valve behaviour. Automatic failure diagnosis requires the output of the model to be around the type of failure occurring in the valve, which is further elaborated in Chapter 3.2. Chapters 3.3 and 3.4 are introducing the training set and case study extracted from Shell's failure notification database.

3.1. Model input

Maintenance performed by Shell's engineers on process equipment is recorded using a notification database. This database contains information on the functional location of the control valve, priority of the incident, description of the failure, the action is taken to repair the issue and the notification date of the failure. The failure cases used in this research are all based on reactive maintenance, where equipment is repaired or replaced after it has run into failure (Swanson, 2001).

The input data of the model consists of time series data recorded in the feedback loop used for the control valves mentioned in chapter 2. The feedback loop aims to minimise the difference between the process variable (PV) measured by the sensor and the set-point of the controller (SP) by adjusting the controller output (OP). All these time series related to the control valves are linked to a functional location describing its operating unit, process and signal type. The operating unit corresponds to the plant in which the control valve is located. The process corresponds to the type of control valve and to the mechanism where the valve is operating in. Four main variants of control valves are active in the refinery of Pernis: Flow, pressure, temperature, and level control valves. These types are indicated with the acronyms: FC, PC, TC, and LC. The signal type is one of the three time series resulting from the controller or sensor: measured process variable, set-point, or controller output. The functional location of the control valve can be generalised as follows: [unit]:[process description].[signal type]. The controller output is a percentage varying between -5 and 105, where -5 corresponds to the controller steering to close fully and 105 to the controller steering to open fully. The PV and SP are the measured flow by the sensor and the set-point flow of the controller in tonnes per day. The amount of flow through the control valve depends on the location and application of the valve. The signals used in this research all have a frequency of 1 update per minute. Table 3.1 shows an overview of the signals used as an input of the model.

Functional location	Unit	Description
ABC:***FC***.OP	[%]	Output of the controller, value between (-5,105)
ABC:***FC***.PV	[tonnes/day]	Flow, measured by the sensor
ABC:***FC***.SP	[tonnes/day]	Flow, set point of the controller

Table 3.1: Name description of control valves in Shell's refinery

The valve can operate in various modes with different control characteristics: cascade, automatic, and manual. The cascade control mode uses the output of one control loop as a target for another control loop. In

Shell's refinery, it is most common for valves to operate in cascade mode to use the difference between the process variable and set-point as a target for the set-point of the control loop. In the automatic mode, the set-point of the control loop is manually set to a value that corresponds to the desired throughput of the control valve. The manual mode requires the operator to set the output of the controller to a specific value.

The time-window of three signals used can vary depending on the requirements of the operation. In chapter 5, cross-validation will be used to test the optimal time-window of the training set. The upper graph of Figure 3.1 contains 300 minutes of the controller output (OP) signal. The middle graph of Figure 3.1 shows 300 minutes of the process variable (PV) and the set-point (SP) signals. The lower graph of Figure 3.1 contains the same data as the previous two. However, it plots the OP versus the PV.

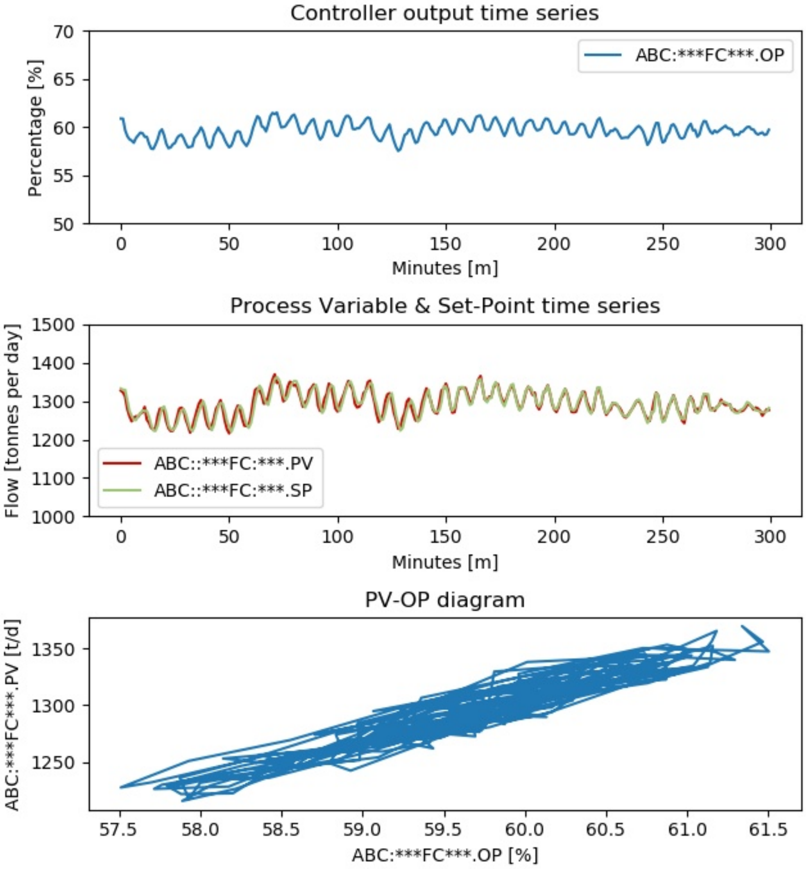


Figure 3.1: Model input time-series example. Upper: controller output time-series. Middle: measured flow and set-point of the controller time-series. Lower: plot of the controller output versus the measured flow

Several assumptions have been made on the input data in order to be able to diagnose failures in flow control valves successfully. First of all, several control valve types are present in Shell's refinery in Pernis, such as the sliding stem and rotating valve. Due to a large number of valves, this research assumes that the failures have similar effects on the signals of different valves. Second of all, the type of material and throughput of the valve can be very different depending on the location and application of the control valve. However, since the goal is to perform a large scale failure diagnosis, the second assumption on the data input is that this does not influence the failure behaviour of the valve. Finally, the last assumption presumes that the control mode of the valve, except for the manual mode, does not influence the significance of the failure behaviour. This results in removing the data where a valve is in the manual control mode from the dataset. Therefore, the only control modes that are considered are automatic and cascade. Table 3.2 contains an overview of the assumptions made on the input data and their results on the model.

Number	Assumption	Result
1	Valve type does not influence failure behaviour	Model that can be used plant-wide for control valves
2	Amount and type of material does not influence failure behaviour	Model that can be used plant-wide for control valves
3	Control mode, automatic or cascade, does not influence failure behaviour	Model that can be used plant-wide for control valves and the removal of manual mode data

Table 3.2: Overview of assumptions made on model input data

3.2. Model output

The output of the model should diagnose failures in control valves. However, the type and amount of failure categories used in fault diagnosis for control valves are very diverse in literature. Some papers, such as Syfert et al. (2003), Trunzer et al. (2018), divide the failures over different components of a control valve. Components that are distinguished are the valve body, plug or seat, air-driven part, and the positioner. In other research, the failure types are divided into various failure groups or behaviour. Bacci Di Capaci et al. (2013) divides the failures into six groups: nominal, disturbance, jamming, stiction, leakage and i/p converter malfunction. Ling et al. (2007) describes four failure categories: leakage, blocking, deadband, and backlash. Mathur et al. (2019) uses three categories, where the valve fails to open fully, open partially, or close fully. Zhang et al. (2012) distinguishes deadband, stiction, packing leakage, and valve saturation. Scali et al. (2011) divide the failure cases into five categories: stiction, jamming, air leakage, periodic disturbance, change in internal valve dynamics. The grouping of failures cases in categories is highly reliant on valve characteristics, transport materials, and failure types that occur in the valves. Therefore failure cases in Shell's historical data should first be thoroughly analysed to divide the failures into categories properly.

During the extraction of the failure cases using the notification database, described in Chapter 3.1, the descriptions created by the engineers were analysed. Seven failure types having various causes were identified in Shell's historical flow control valve data. These failure types are shown on the left side of the diagram shown in Figure 3.2. When just observing and analysing the data, these failure types are hard to distinguish; however, the control valves can show similar failure behaviour for different failure types. Therefore, based on the divisions made in literature and analysis of Shell's data, failure behaviour in flow control valves is divided into five groups that can be classified and predicted using new incident data. The failure behaviour categories are shown on the right in the diagram of figure 3.2.

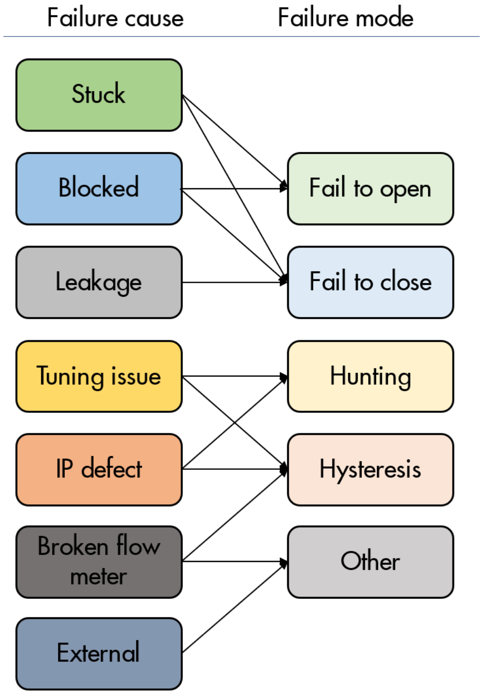
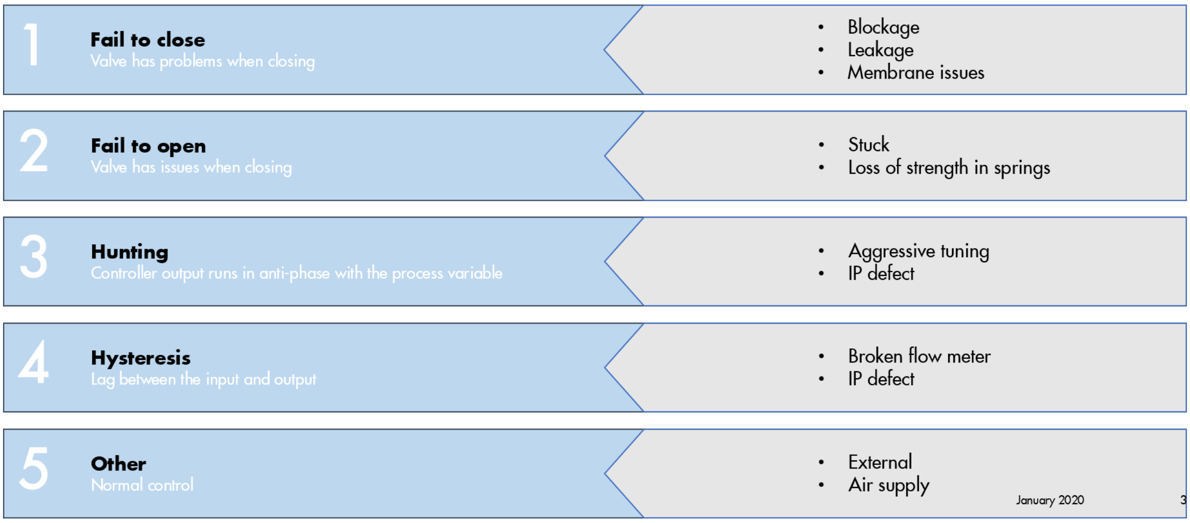


Figure 3.2: Left: seven failure types identified in Shell's data. Right: five failure behaviour categories that can be distinguished using Shell's data.

The 'fail to open' category contains failures where the valve has issues opening, for example, when the positioner is stuck, or the valve is blocked. 'Fail to close' is failure behaviour where the valve has problems closing, because of leakage, blockage in the valve opening, or stuck positioner. During 'hunting' behaviour, the controller output and measured process variable run in an anti-phase, which causes control problems for the valve. 'Hysteresis' behaviour, is another control issue, where the measured process variable has a significant lag in comparison with the output of the controller. There are several sources for the behaviour of these two control issue, hunting and hysteresis, such as a current to pressure (I/P) converter defect. The final behaviour category 'other' includes failure cases where the valve still controls well. However, the maintenance did report a failure in the notification database. Figure 3.3 contains an overview of the five failure categories identified in Shell's historical data of flow control valves.



1 **Fail to close**
Valve has problems when closing

- Blockage
- Leakage
- Membrane issues

2 **Fail to open**
Valve has issues when closing

- Stuck
- Loss of strength in springs

3 **Hunting**
Controller output runs in anti-phase with the process variable

- Aggressive tuning
- IP defect

4 **Hysteresis**
Lag between the input and output

- Broken flow meter
- IP defect

5 **Other**
Normal control

- External
- Air supply

January 2020 3

Figure 3.3: Overview of the five identified failure categories. Left: failure categories and short explanation. Right: possible causes of the failure behaviour

The five failure categories will serve as an output of the automatic failure diagnosis model. More in-depth analysis on the five failure categories can be found in the Chapters 3.2.1 - 3.2.5.

3.2.1. Fail to open

The category 'fail to open' represents failure behaviour where the valve has problems opening. The descriptions in the notifications of this category range from a stuck valve to a diaphragm leakage. Figure 3.4 shows historical data of a control valve that is not opening correctly. In the upper graph, the controller attempts to open the valve fully; however, when comparing this to the measured flow, the valve is not actually opening. The lower graph shows that the value of the set-point is higher than the measured flow most of the time. Therefore the controller tries to open the valve by increasing the controller output fully.

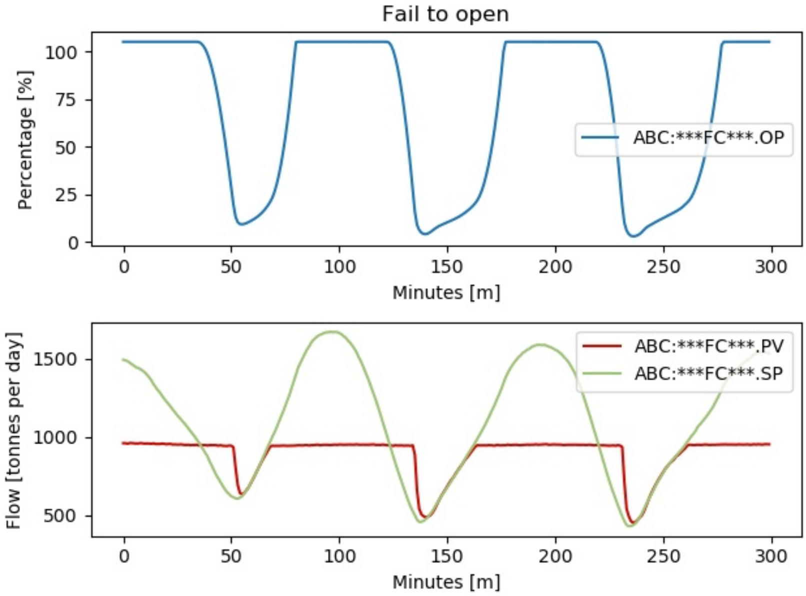


Figure 3.4: Data representing Fail to open behaviour in a flow control valve. Upper: 300 minute time-series data of the controller output. Lower: 300 minute time-series data of the process variable and set-point.

3.2.2. Fail to close

The category 'fail to close' contains cases where the valve has problems when closing. In the notifications of the failures, descriptions are ranging from flow through the valve while it should be fully closed to a valve stuck at 25%. Figure 3.5 contains 300 minutes of the control valve signals during the Fail to close behaviour, where material flows through the valve when it should not. The upper graph shows the controller output trying to close the control valve fully. The lower graph shows the measured flow above the set-point of the controller most of the time. Due to this difference, the controller tries to close the valve by steering the valve closed fully; however, it fails.

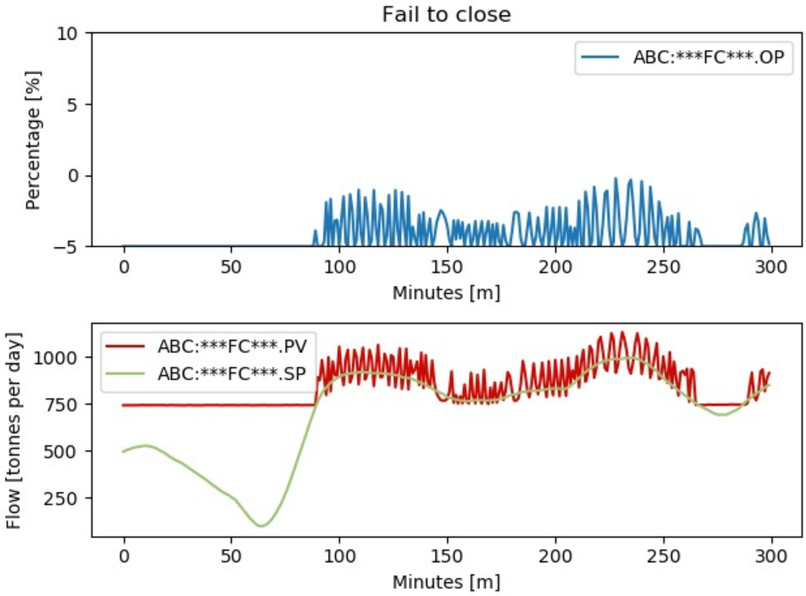


Figure 3.5: Data representing Fail to close behaviour in a flow control valve. Upper: 300 minute time-series data of the controller output. Lower: 300 minute time-series data of the process variable and set-point.

3.2.3. Hunting

In literature, a ’hunting’ valve is operating with the on and off-cycle close together, resulting in a constant overshoot of the positioner (Choudhury et al., 2004). In the five failure behaviour categories, ’hunting’ corresponds to a control issue where the signal of the controller output and measured process variable operate in anti-phase, which corresponds to a delay of 180 degrees ($-\pi$). When the valve is hunting, it does not control well, which could cause degradation of the valve material and heavy oscillations. The descriptions in the notifications of historical failure data range from the valve oscillating heavily to tuning issues of the controller. Figure 3.6 shows Hunting behaviour in a flow control valve using a time-series of 120 minutes. The upper graph compares the time-series of the controller output and the measured process variable. Since the unit of the controller output is a percentage and the unit of the flow is tonnes per day, standardised data is used to be able to compare the data properly. Standardisation of a dataset is done by removing the mean and scaling to the unit variance. The y-axis shows of the number of standard deviations the signal is from the mean. The lower graph, using the raw data of the upper graph, plots the controller output versus the measured process variable. When comparing the lower graph to the lower graph of Figure 3.1, the plot has rotated about 90 degrees clockwise. Depending on the underlying issue, hunting can be fixed using control tuning.

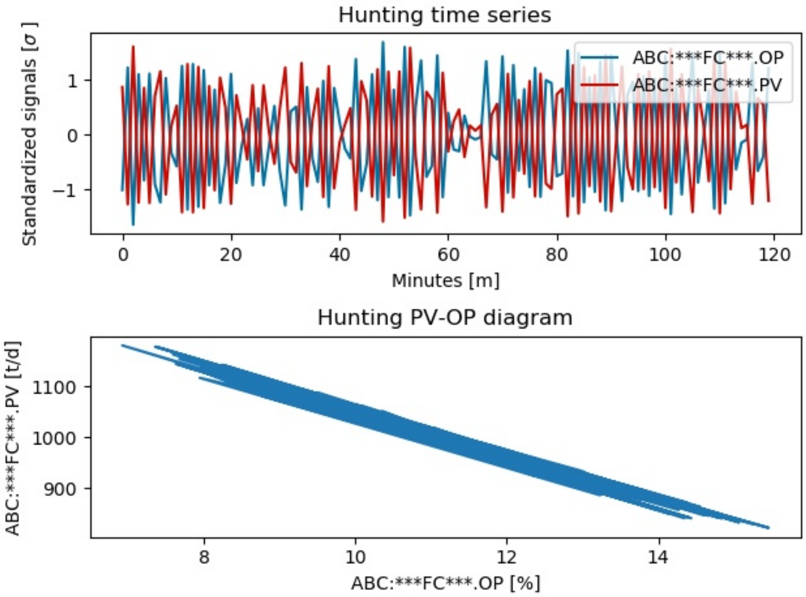


Figure 3.6: Data representing Hunting behaviour in a flow control valve. Upper: showing normalized data ($\frac{X-\mu}{\sigma}$) of the 120 minute time series OP and PV. Lower: graph showing the raw OP and PV data of the same 120 minutes dataset.

3.2.4. Hysteresis

In literature, the phenomenon ’hysteresis’ is defined as: "the difference between the valve position on the upstroke and its position on the downstroke at any given input signal’ (Emerson Automation Solutions, 2017). So for similar values of the controller output, the measured flow will have completely different values, which in the time-series causes a lag between the OP and PV. The upper graph of Figure 3.7 shows the time-series, where the lag is visualised. In the lower graph of Figure 3.7 it clearly shows that for the same value of the OP, a flow difference exists of about 60 tonnes per day. When comparing the lower graph of Figure 3.7 to the lower graph of Figure 3.1, a clockwise rotation of about 45 degrees is found. The notifications in of the failure cases contain descriptions such as flow meter is broken, or hysteresis.

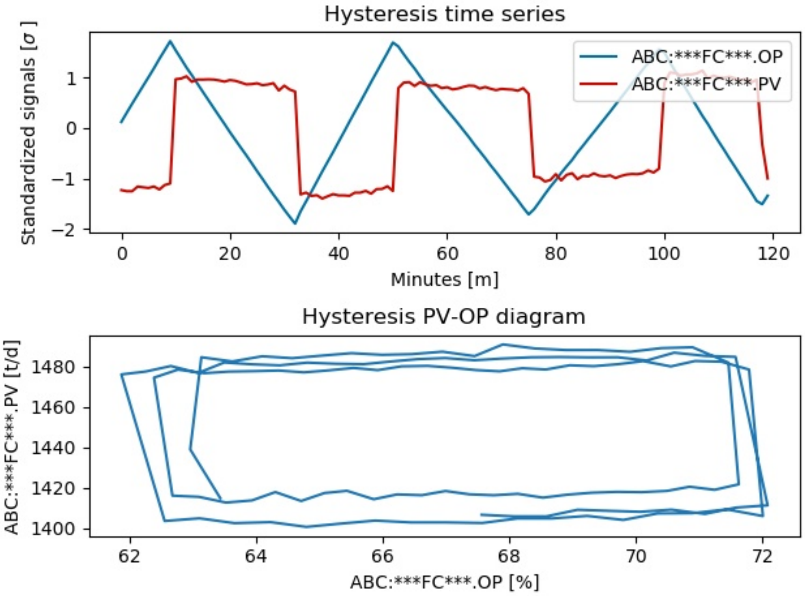


Figure 3.7: Data representing Hysteresis behaviour in a flow control valve. Upper: showing normalized data ($\frac{X-\mu}{\sigma}$) of the 120 minute time series OP and PV. Lower: graph showing the raw OP and PV data of the same 120 minutes dataset.

3.2.5. Other

In the failure notification database, there are also failures, where when analysing the data, no clear failure behaviour can be spotted. However, a report has been created by an engineer. An example of a failure in the 'Other' mode would be gland leakage, where there is a leakage which is not visible in the data. In the upper graph of Figure 3.8, the standardised time-series of the OP and PV are showing that the valve controls well. The lower graph of Figure 3.8 also shows a PV-OP diagram with a positive correlation. The Other category currently functions as a bucket for failures cases where the valve still controls well; however, there is an issue raised by the engineer.

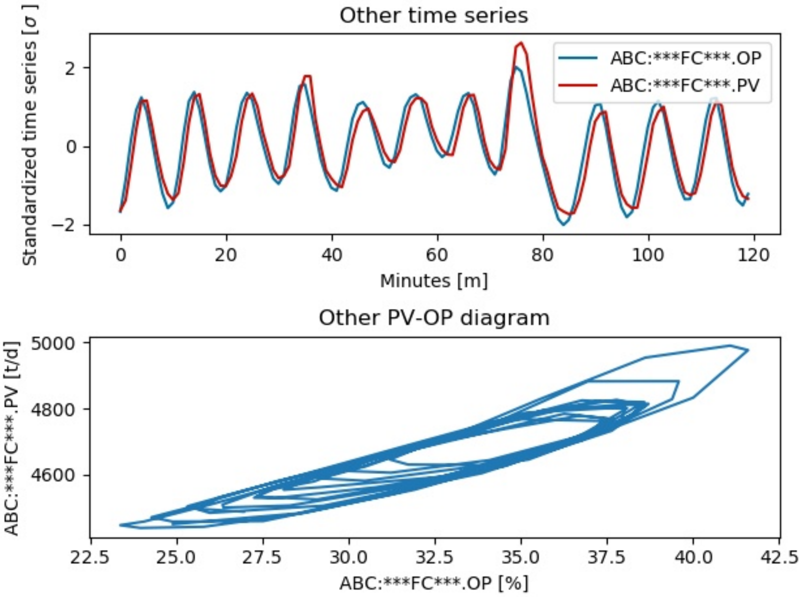


Figure 3.8: Data representing Other behaviour in a flow control valve. Upper: showing normalized data ($\frac{X-\mu}{\sigma}$) of the 120 minute time series OP and PV. Lower: graph showing the raw OP and PV data of the same 120 minutes dataset.

3.3. Training dataset

This section will give a clear overview of what the training set of the failure diagnosis model looks like, and how it is established. The failure cases training set is built from historical cases retrieved from Shell's failure notification database. In order to create a model that can be used plant-wide, meaning that it will have high performance when monitoring different plants within the refinery, we selected five plants with different locations.

The engineers creating notifications have the possibility to assign a priority to the valve incident. This priority level ranges from 1 to 5, where a priority 1 is the most critical. This research considers failure cases with priority (1 & 2). The notifications used in this research range from 2014 to the beginning of 2020. The following four steps were taken in order to create the extract and label the training set successfully:

1. Manually extract control valve signals (OP, PV, SP) of priority (1 & 2) historical failure cases from Shell's notification database.
2. Data exploration and analysis of failure cases to determine failure behaviour.
3. Verification of the failure behaviour with an instrumentation engineer from Shell's refinery.
4. Labelling the failure case with the correct output class.

The applications of the units vary from crude distillers to Hydrocracker units. Due to the variations in the number of valves per unit due to the number of failure cases is different for every plant. The lifespan of a control valve processing dense materials such as butamin often has a shorter lifespan and more issues than valves handling gasoline. Table 3.3 contains an overview of the five plants used for the training set.

Number	Name	Type of unit	Number of failure cases
1	ABC	Crude distiller 1	15
2	DEF	Hydrocracker unit	10
3	GHI	Hydrodesulfirization	7
4	JKL	Crude distiller 2	14
5	MNO	Platformer	5

Table 3.3: Overview of units in the refinery where failure cases have been retrieved from.

In step 2 to 4 the failure behaviour of the valve is determined and verified with instrumentation engineers, who have often been working with the machinery daily for a long time, and thus have expert knowledge on control valves. When the category is confirmed, the failure cased is labelled with the correct output class. Table 3.4 contains an overview of what the complete dataset of failure notifications looks like.

number	Functional location	Label	Date	Number of signals
1	ABC:***FC***	Fail to open	01-01-2014	3
2	ABC:***FC***	Hunting	06-02-2019	3
...	...	...	...	...
49	MNO:***FC***	Fail to close	05-09-2015	3
50	MNO:***FC***	Other	15-01-2020	3

Table 3.4: Overview of the notifications gathered from historical failure data of Shell's refinery.

3.4. Case study: Shell Gasification Hydrogen Plant

In order to get valuable results from the case study applied to a machine learning model, several requirements were set. First of all, the case study set should be kept separately from the training set. Second of all, the failures in the case study should be gathered from a new unit type to test the plant-wide operability of the model. Third of all, every failure category in the case study should contain at least two failure cases.

The unit chosen for the case study is the Shell Gasification Hydrogen Plant (SGHP). Gasification techniques have been around for several decades. However, the interest in the process is higher now than ever. This is due to environmental advantages, where atmospheric pollution from materials such as SO_x and NO_x are being reduced. Traditionally, crude oil residues were sold as a marine fuel, and however, nowadays, gasification is used to process them into clean fuel (Zuideveld and Graaf, 2005). The SGHP provides H_2 and Syngas $(CO + H_2)$ to the Hydrocracker unit (HCU) and the Power Generation Plant (PGP), during this process it reduces the amount of residue. Syngas is made by partial oxidation of the residue in the SGHP plant, afterwards H_2S and CO_2 are removed from the Syngas. Part of the Syngas is shift from CO to H_2 and CO_2 after the removal of Sulphur. The filtrate of the residue combustion from the gasifier is treated by the SARU plant.

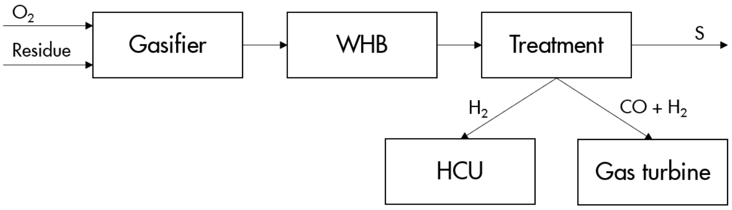


Figure 3.9: Overview of process in the SGHP unit. Gasification process of the refining residue to provide H_2 and Syngas to the HCU and PGP.

The load on control valves in the SGHP unit is exceptionally high, due to the residue materials that flow through the process equipment. Within the period 2014 - 2020, 18 failures were found with priority 1 or 2. Every behaviour category exists at least two times in the case study data. Table 3.5 contains the list with all failures and the corresponding label incorporated in the case study.

Number	Functional location	Failure behaviour	Date	Number of signals
1	PQR:***FC***	Hysteresis	06-01-2020	3
2	PQR:***FC***	Hysteresis	06-10-2014	3
3	PQR:***FC***	Fail to close	08-09-2015	3
4	PQR:***FC***	Other	17-06-2016	3
5	PQR:***FC***	Hunting	05-08-2016	3
6	PQR:***FC***	Other	15-09-2016	3
7	PQR:***FC***	Fail to close	19-05-2017	3
8	PQR:***FC***	Hunting	24-07-2017	3
9	PQR:***FC***	Hysteresis	30-05-2018	3
10	PQR:***FC***	Fail to close	15-08-2018	3
11	PQR:***FC***	Fail to open	18-01-2019	3
12	PQR:***FC***	Hunting	29-01-2019	3
13	PQR:***FC***	Hysteresis	16-01-2015	3
14	PQR:***FC***	Fail to open	17-01-2014	3
15	PQR:***FC***	Fail to open	23-03-2015	3
16	PQR:***FC***	Other	05-12-2015	3

Table 3.5: List of failures cases and related behaviour label extracted from the SGHP plant.

3.5. Conclusion

Monitoring process equipment based on its condition requires sufficient data. Shell's refinery has lots of data, such as flow sensors and controller signals with an update frequency per minute. Therefore, many possibilities are available for the implementation of predictive maintenance. However, due to the small number of historical failure cases available for valves, some assumptions have to be made on the model input data to ensure it generalises for every valve on the asset. The output of the automatic failure diagnosis consists of five failure behaviour categories: Fail to open, Fail to close, Hunting, Hysteresis, and Other. A training and test set has been prepared to use supervised ML classification. The training data, built from failure cases extracted from Shell's maintenance notification database, covers five plants with different locations and applications. The case study includes data from the SGHP unit, where the gasification process takes place on refining residue materials. The dataset of this new unit is kept separately and has a different application than the plants involved in the training set.

4 Methodology

Chapter 4 contains the methodology applied in this research for automatic failure diagnosis in control valves. First of all, the chosen approach and the reasoning behind the concept will be discussed in Chapter 4.1. The theoretical background behind the required steps in the model will be further elaborated in Chapters 4.2 - 4.4. The framework used for ML classification is shown in Chapter 4.5. Finally, the methods used to evaluate the results of the research are explained in Chapter 4.6.

4.1. Approach

Feature engineering and supervised ML classification has been chosen to diagnose failures in control valves automatically. The input data are time-series covering the failure behaviour, and the output data are the five breakdown categories discussed in Chapter 3.

4.1.1. Feature extraction

The time-series data showing the failure behaviour are compressed into single values that can be used by the supervised machine learning model. Therefore, a set of features have been chosen to distinguish the five failure behaviour categories. The first feature is the variance of the controller output (OP), measuring how far the time-series data is spread out from the average value. The goal of the OP variance is to distinguish failures showing differences in oscillations in the OP, for example, due to signs of hunting or hysteresis. The second feature is the mean of the OP, which is the average value over time. When a valve is having issues opening, the controller output often obtains substantial values for the OP contributing to a high average value. In 'Fail to close' behaviour, the opposite behaviour occurs, leading to a low average OP. Therefore, the mean OP could be a successful feature. The third, fourth, and fifth features are ratios between the variances of the three signals OP, PV, and SP. Ratios of the variances are chosen to remove the differences in measurements between the flow valves. The difference exists due to the different location, size, and application of a valve, which can have a throughput of 1000 tonnes per day, while the other valve has a throughput of 5 tonnes per day. The OP is a percentage between [-5,105] for every valve. The goal of the ratios between the variances is to distinguish differences in the volatility of the signals, which can be diverse for different behaviour. The sixth feature is build to differentiate gaps between the PV and the SP signal, which is the error that the controller tries to minimise. When the valve has trouble opening or closing as much as it desires the sixth feature will increase. The seventh, eighth, and ninth features are the Pearson's correlations between the signals. In the case of 'hunting', the OP and the PV signals are running in anti-phase, therefore the correlations between the signals should be negative. In the case of the 'other' category, the valve controls well. Therefore, the correlation between the PV and the SP should be around 1. The next three categories are using Power Spectral Density (PSD), which is the measure of the signal's power content versus frequency, used to characterise time series data. The power spectrum of a time series describes the distribution of power into frequency components composing that signal. This is obtained by Fourier analysis, which decomposes signals into contributions by several discrete frequencies. The features consist of the PSDs divided over three ranges of frequencies: low, medium, and high. Using these features, signals with low frequencies can be distinguished from high frequencies. The last feature contains the absolute difference between the OP and the PV signals to identify large inconsistencies between these time series. Table 4.1 shows an overview of the chosen features.

Feature	Short description
1	Variance of the OP
2	Mean of the OP
3	Variance SP / Variance PV
4	Variance SP / Variance OP
5	Variance PV / Variance OP
6	(PV-SP)/SP
7	Correlation SP PV
8	Correlation SP OP
9	Correlation PV OP
10	PSD low frequencies
11	PSD medium frequencies
12	PSD high frequencies
13	Absolute difference PV OP

Table 4.1: Summary table of the created features for extraction.

4.1.2. Feature selection

The feature selection methods discussed in Chapter 2.4 have different qualities and specialities, therefore, for the data available from Shell's refinery, several criteria have been set to determine the proper approach. First of all, the statistical test should be able to handle positive and negative values. Second of all, both a linear and a non-linear approach have to be applied to obtain an optimal feature selection mechanism. Finally, the method has to be able to manage numerical data input and categorical data output. Table 4.2 checks the four criteria for the discussed methods in Chapter 2.4.

	Pearson	Spearman	ANOVA	Kendal Tau	Mutual information	Chi-squared
Negative values	Yes	Yes	Yes	No	Yes	Yes
Linear/non-linear	Linear	Non-linear	Linear	Non-linear	Non-linear	Linear
Numerical input	Yes	Yes (ordinal)	Yes	Yes (ordinal)	Yes	Yes
Categorical output	No	No	Yes	Yes	Yes	No

Table 4.2: Summary table of the statistical tests available for feature selection and the criteria to rate the tests on.

The ANOVA F-test is a suitable linear statistical test, and Mutual information is a suitable non-linear test which can be applied for the feature selection. Chapter 4.3 gives a proper insight on the theory behind the two methods.

4.1.3. Supervised machine learning classification

The supervised ML classification methods discussed in Chapter 2.7 all have certain characteristics depending on the data type, application, and desired output. Several criteria have been defined to test which classifier is optimal to perform the failure diagnosis for control valves using the data from Shell's refinery. First of all, the model should be highly interpretable for the engineers so no uncertainties about the classification decision will exist. Second of all, the model should not be too prone to overfitting to prevent accurate results on the training set, but poor results on the case study set. Third of all, the model should be able to perform well on the data by obtaining highly accurate classification results. Fourth of all, due to the possible non-linearities in the relations within the feature, data models should preferably be able to detect non-linear behaviour. Finally, the implementation, consisting of the training and tuning of the model, should be straightforward regarding Shell's data. In summary, these five criteria will be used to determine the optimal models for failure diagnosis in control valves:

- Interpretability
- Prone to over fitting
- Accuracy

- Non-linear
- Implementation

Multi-criteria analysis with evenly distributed weights for the criteria has been chosen to select the best model for this application. Table 4.3 contains the analysis regarding the supervised ML classifier.

	Linear regression	Logistic regression	ANN	RF	SVM	NB
Interpretability	3	3	1	2	3	3
Prone to overfit	1	3	1	2	1	2
Accuracy	1	2	3	3	3	3
Non-linear	1	1	3	3	2	3
Implementation	3	3	1	3	1	3
Total score	9	12	9	13	10	14

Table 4.3: Multi-criteria analysis on supervised ML classification models regarding the data of Shell's refinery in Pernis.

The methods RF, and NB score the highest on the criteria. Therefore, these three supervised ML classification models will be implemented.

4.1.4. Conclusion

The complete framework chosen for failure diagnosis in control valves for Shell's refinery consists of the following steps, shown in Figure 4.1. The data input consists of the time-series of the control valves showing failure behaviour. In the first step, 13 features are extracted from the data input. The second step contains two selection methods used to choose the features that are good predictors, resulting in two feature sets. In the third step, three ML classification models are used to predict failure behaviour on new incident data. This results in 6 different prediction outcomes which are found in the output data.

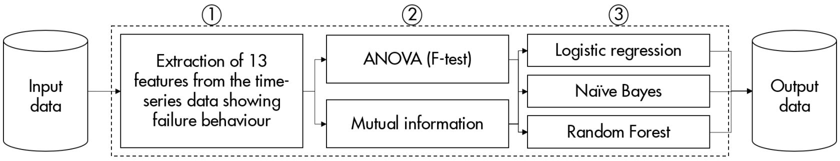


Figure 4.1: Final framework used for failure diagnosis in control valves for Shell's refinery in Pernis.

4.2. Feature extraction

The first step of the framework is creating features from the signals showing the failure behaviour in the control valve. The time-series input data consists of three signals OP, PV, and SP denoted as:

Signal	Symbol	Set
OP	x_i	i = 1,..,n
PV	y_i	i = 1,..,n
SP	z_i	i = 1,..,n

Where n is the length of the time-series chosen to represent the failure behaviour. If $n = 60$, the data represents 60 minutes, since the update frequency of the signals is per minute.

The first feature, measuring how far the time-series data is spread out from the average value, is the variance of the OP represented by Equation (4.1).

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2 \tag{4.1}$$

Equation 4.2 contains the second feature which is the average value of the OP.

$$F_2 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{4.2}$$

The goal of Equation (4.3) - (4.5) is to distinguish differences in volatility of the signals by representing the variances between the controller signals. Ratios of the variances are chosen to remove the differences in measurements between the flow valves.

$$F_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\bar{y} - y_i)^2} \tag{4.3}$$

$$F_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2} \tag{4.4}$$

$$F_5 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{y} - y_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2} \tag{4.5}$$

The sixth feature is created to differentiate behaviour where large gaps between the PV and SP signals exist. This is done by subtracting the SP signal from the measured PV, and afterwards dividing it through SP to remove differences in measurements.

$$F_6 = \frac{\sum_{i=1}^{n}(y_i - z_i)}{\sum_{i=1}^{n}(z_i)} \tag{4.6}$$

Equation (4.7) - (4.8) are characterizing linear Pearson's correlation between the three signals.

$$F_7 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2}} \tag{4.7}$$

$$F_8 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{4.8}$$

$$F_9 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{4.9}$$

The next three features consists of the PSDs of the OP signal divided into the lower, medium, and high frequencies using Welch's method (Welch, 1967). This approach estimates the power spectral density of the signals using fast Fourier transform. The record is sectioned, modified periodogram of these sections are taken, and the modified periodogram are averaged.
The implementation of the Welch's method is done in Python using the signal processing package from SciPy.org (2019). The output of the Welch's method are a list of frequencies between 0 and 1 in Hz, and a list of corresponding PSDs in dB. The list of PSDs is divided into three groups belonging to the low, medium, and high frequencies answering to the frequencies 0 – 1/3Hz, 1/3 – 2/3Hz, and 2/3 – 1Hz.

$$F_{10} = \sum_{i=0}^{n/3} S_x^W(\omega_x) \tag{4.10}$$

$$F_{11} = \sum_{i=n/3}^{n(2/3)} S_x^W(\omega_x) \tag{4.11}$$

$$F_{12} = \sum_{i=n(2/3)}^{n} S_x^W(\omega_x) \tag{4.12}$$

The last feature determines the absolute value of the difference between the signals OP and PV. Since the signals OP and PV have different units, subtracting them could lead to useless results, due to differences in measurement of the control valve. Therefore some transformations to the signals are performed to make the time-series comparable. First of all, the first order differencing method is applied to help stabilize the mean of the time series by removing changes in the level of a time series, and so eliminating trend and seasonality. The first order difference is performed by subtracting the previous observation from the current observation. Equation 4.13 shows the first order differencing method for the OP signal, similar action is performed on the PV signal.

$$x_i = x_i - x_{i-1} \tag{4.13}$$

After differencing to remove the dependence on time to eliminate trends and seasonality the signal is standardized. Standardizing of features is done by removing the mean and scaling to unit variance. Equation 4.14 shows standardization of the OP signal, similar action is performed on the PV signal.

$$x_i = \frac{x_i - \bar{x}_i}{\sigma_x} \tag{4.14}$$

After the two transformation steps the absolute difference between the OP and PV signal is calculated using Equation 4.15.

$$F_{13} = \sum_{i=0}^{n} \mathrm{abs}(y_i - x_i) \tag{4.15}$$

4.3. Feature selection

This section describes the feature selection methods implemented in the failure diagnosis model for flow control valves. The first method is ANOVA, comparing the variability between and within variables. The second method is Mutual information, using entropy to determine mutual dependence. The last part describes the significance tests used to determine the number of features used as input for the supervised ML model.

4.3.1. ANOVA (F-test)

The Analysis of Variance (ANOVA) test measures the degree of linear dependency between two random variables. The method can be used to check whether the means of two variables are significantly different from each other, allowing the feature to be a good predictor for the output. The ANOVA test calculates an F-value with the following equation (Kumar et al., 2015):

$$F = \frac{\text{Variability between groups}}{\text{Variability within group}} \tag{4.16}$$

Figure 4.2 shows two examples of probability density functions for two random variables X. The variability within the group is the same for both examples. On the left, the distance between the means of variable 1 and 2 is small, therefore, the variability between the groups is low, resulting in a low F-value. On the right, the distance between the means is considerable, which produces a high F-value.

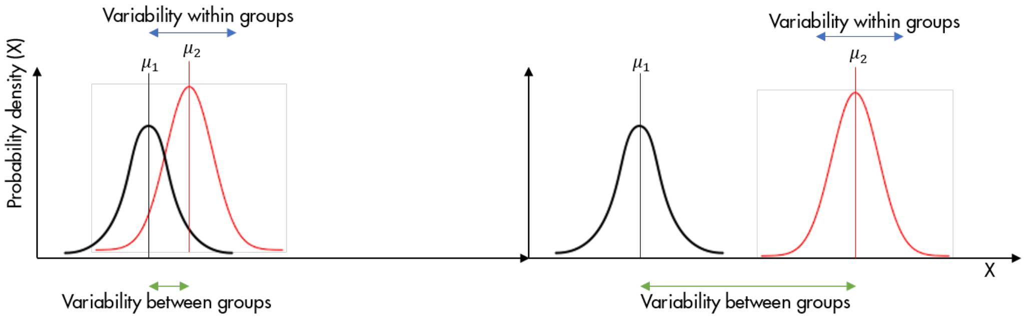

Variability within groups
μ_1
μ_2
Probability density (X)
Variability between groups
Variability within groups
μ_1
μ_2
X
Variability between groups

Figure 4.2: Probability density functions of two random variables. Left: example of low F-value, due to small distance between the means of the variables. Right: example of high F-value, due to large distance between the means of the variables.

Since for the ANOVA test, the variance within and between the groups is required the Total Sum of Squares is applied between the feature and predictor, and within the feature and predictor. Furthermore, degrees of freedom are used, which refers to the maximum number of logically independent values that have the freedom to vary. The notation is performed according to research performed by Elssied et al. (2014).

$$s_j^2 = \frac{\sum_{i=1}^{N_j}(x_{ij} - \bar{x})^2}{N_j - 1} \tag{4.17}$$

Where N_j is the number of cases with Y = j, for our specific case Y represents the different failure behaviour categories. Furthermore, $\bar{x_j}$ is the sample mean of predictor X for target class Y = j, and s_j^2 is the sample variance of predictor X for target class Y=j.

$$\bar{g} = \frac{\sum_{j=1}^{J} N_j \bar{x_j}}{N} \tag{4.18}$$

Where g is the grand mean of predictor X. Finally, the F-value between the predictor and output can be calculated using Equation (4.19).

$$F = \frac{\dfrac{\sum_{j=1}^{J} N_j(\bar{x_j} - g)^2}{(J-1)}}{\dfrac{\sum_{j=1}^{J}(N_j - 1)s_j^2}{(N-1)}} \tag{4.19}$$

Where N-1 is the degrees of freedom within the groups, which is the sum of degrees of freedom for all groups. J-1 is the degree of freedom between the groups.

4.3.2. Mutual information

Mutual information is used to determine the mutual dependence between two variables by quantifying the 'amount of information' obtained about one random variable as a result of observing the other random variable. This is a widely used method to measure non-linear dependency between variables and can vary between [0,1], where 0 is completely unrelated or independent, and 1 is completely related or dependent. Mutual information is calculated using formula (2.3):

$$I(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{4.20}$$

Where p(x,y) is the joint probability density function of the input feature and output, and p(x) and p(y) are the marginal density functions of the input feature and output. In our specific case y represents the different failure behaviour categories, and x corresponds to one of the extracted features from the time series data showing failure behaviour.

Mutual information is based on the statistical property, entropy, which is a measure of information or uncertainty. In statistics, the 'Shannon' entropy is calculated using the following formula:

$$H(X) = H(p_1,..,p_n) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{4.21}$$

Where X is an event with n possible outcomes and probabilities $p_1,...,p_n$. An example of entropy can be shown using a coin flip, where the possible outcomes are heads or tails. When plotting the probability of the event heads happening versus the entropy of heads Figure 4.3 is obtained.

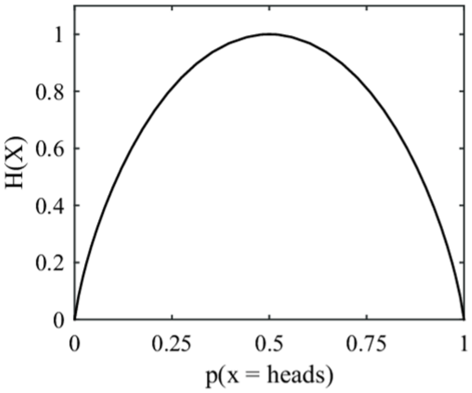


Figure 4.3: Entropy example using a coin flip. Probability of heads occurring is plotted versus the measure of uncertainty, Entropy.

When the probability of heads occurring is 0.5, the entropy is maximised, due to the uncertainty being at a maximum. When the probability of heads developing is 1, there is no uncertainty, therefore, the entropy is 0. Conditional entropy quantifies the amount of information required to describe the outcome of a random variable Y when it is given that random variable X is established.

$$H(Y|X) = \sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x,y)}{p(y)} \tag{4.22}$$

Using the entropy and conditional entropy mutual information, which indicates the amount of uncertainty or information gained, can be determined using the following formula:

$$I(X,Y) = H(X) - H(Y|X) = H(Y) - H(Y|X) \tag{4.23}$$

When two outcomes are completely independent, such as for two events of a fair coin flip, the calculated mutual information value is 0, since no information is gained by the outcome of the first coin flip. Figure 4.4 hows a visualization of mutual information and its relation to entropy, and conditional entropy.

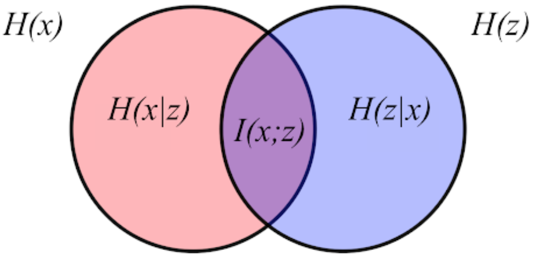


Figure 4.4: Overview of how Mutual information can be determined using entropy and conditional entropy.

Mutual information can be calculated using the underlying probability of the two data sets. However, when this is not the case, Mutual information must be estimated. When both datasets are discrete, frequencies can be calculated of combinations of pairs by counting the times each pair occurs in the data. However, for continuous data sets, this is less straightforward, therefore, 'binning' methods are proposed where continuous variables are lumped into discrete 'bins'. In the failure diagnosis model for control valves, Mutual information is implemented using the estimation method proposed by Ross (2014) and Kraskov et al. (2004), which was originally introduced by Beirlant et al. (1997). Ross (2014) describes an accurate, non-binning MI estimator for the case of one discrete data set and one continuous data set. The method is based on the nearest neighbour method using Equation 4.23.

The model uses a continuous dataset x, and a discrete dataset y, which underlying distribution can be estimated by looking at how the points are clustered. Each data point i computes a number I_i based on its nearest-neighbours in the continuous variable x. The k-th closes neighbours to point i among those N_{y_i} data points whose value of the discrete variable equals y_i using a distance metric, which is defined as d distance to this kth neighbour. Afterwards, the number of neighbours m_i is counted that lie in the area d to point i. Based on these results, I_i can be computed:

$$I_i = \psi(N) - \psi(N_{y_i}) + \psi(k) - \psi(m_i) \tag{4.24}$$

Where psi(−) is the digamma function. To estimate Mutual information from the dataset, the I_i is averaged over all data points, resulting in Equation (4.25).

$$I(X,Y) = \psi(N) - \langle\psi(N_{y_i})\rangle + \psi(k) - \langle\psi(m)\rangle \tag{4.25}$$

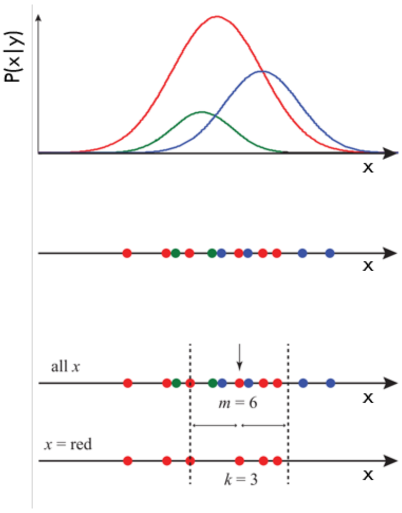


Figure 4.5: Mutual information estimation procedureRoss (2014). Top graph shows an example probability density where y can take three values (colours) and y is a continuous variable. Middle graph shows a set of (x,y) data pairs sampled from the distribution. Lower graph shows the computation of the I_i method.

4.3.3. Significance test

The value for ANOVA and Mutual information is determined in the previous sections. However, only the features x_i that are significant predictors for the output should be applied in the model.

Without a significant difference, for the ANOVA, between the groups, the F value will be close to one. When the F value is large, there is a significant difference between the groups. The hypothesis of the ANOVA test is used to check the significance of the difference. The hypothesis test is stated as follows:

- H_0: Means of all groups are equal, therefore, feature x_i is no significant predictor for the output y.
- H_1: Mean of at least one group is different, therefore, feature x_i is a significant predictor for the output y.

The critical F-value can be found, given the degrees of freedom of the nominator and the chosen confidence interval (α), in the F-distribution table (von Storch and Zwiers, 2010). Suppose, the feature has an F-value above the critical F-value, the null hypothesis is rejected. This means variance exists between the groups, which shows the feature has an impact on the output and therefore is a significant predictor for the output. In this case, the feature is adopted in the model. When the F-value is below the critical F-value, the null hypothesis is not rejected, and the feature will be removed from the input of the ML model.

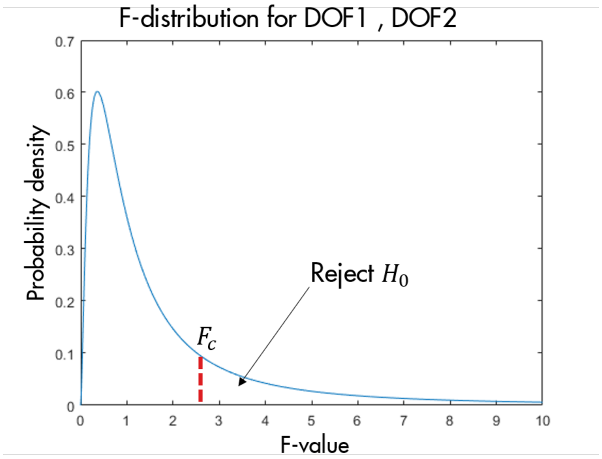


Figure 4.6: F-distribution for certain degrees of freedom for the ANOVA test. The graph shows an example of the critical F-value, which can be determined for α, given the F-distribution. In the area right of the critical F-value the null hypothesis is rejected.

The significance of the features as a result of the Mutual Information method is tested differently. One thousand random normally distributed noise variables ($\mu : 0, \sigma^2 : 1$) are created and tested using the Mutual information test. Afterwards, depending on the determined confidence interval, the quartile value of the Mutual information outcomes is chosen as the boundary. The hypothesis test is stated as follows:

- H_0: The feature x_i contains less or equal information on the output as the 95th quartile of thousand noise variables, therefore, feature x_i is no significant predictor for the output y.
- H_1: The feature x_i contains more information on the output as the 95th quartile of thousand noise variables, therefore, feature x_i is a significant predictor for the output y.

The null hypothesis is rejected if the mutual information value of the feature is higher than the determined boundary, which means the feature is a significant predictor depending on the chosen confidence interval.

4.4. Supervised machine learning classification

First of all, this section contains more information on the methodology behind the chosen supervised ML classification methods. Second of all, the framework of applying ML classification to the problem is described. Finally, the model evaluation methods used to assess the performance of the classifier are discussed.

4.4.1. Logistic regression

Logistic regression is a classifier, widely used in supervised ML problems, that determines the probability that a specific value of the predictor belongs to a particular class or category using a logistic function. The logistic function, with an S-curve shown in Figure 2.17, uses a function shown in Equation (4.26).

$$\phi(x) = \frac{1}{1+e^{-x}} \tag{4.26}$$

The goal of applying logistic regression in classification is to determine the probability, between 0 and 1, of a set of observations belonging to a certain class given the values of the observations in the features. The logit function can be written as the logistic function shown in Equation (4.27).

$$\mathrm{P}(c|X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1...+\beta_n X_n}} \tag{4.27}$$

Where $\mathrm{P}(c|X)$ is the probability that a certain class occurs given predictor $X = x_0,...,x_n$. β_0 is the intercept coefficient. $\beta_1,...,\beta_n$ are the regression coefficients. After some manipulation the logistic function can be written as the logg-odds logit, shown in Equation (4.28).

$$\log\left(\frac{\mathrm{P}(c|X)}{1-\mathrm{P}(c|X)}\right) = \beta_0 + \beta_1 X_1 ... + \beta_n X_n \tag{4.28}$$

The coefficients required in logistic regression are estimated using the maximum likelihood function. The likelihood function corresponding to a binary classification problem is shown in Equation (4.29).

$$L(\beta_0,...,\beta_n) = \prod_{i,c=0} \mathrm{P}(x_0,...,x_1) \prod_{i,c=1} \mathrm{P}(x_0,...,x_1) \tag{4.29}$$

The goal of the maximum likelikhood estimation is to obtain parameters for the model that maximize the likelihood, which is done by minimizing the cost function corresponding to the log loss function (Geron, 2017).

The decision surface resulting from the logistic regression classifier are linear, shown in Figure 4.7. Logistic regression can be used for binary classification, but also for multi-class classification using, for example, one-versus-rest or entropy loss (James et al., 2013), the difference is shown in Figure 4.7.

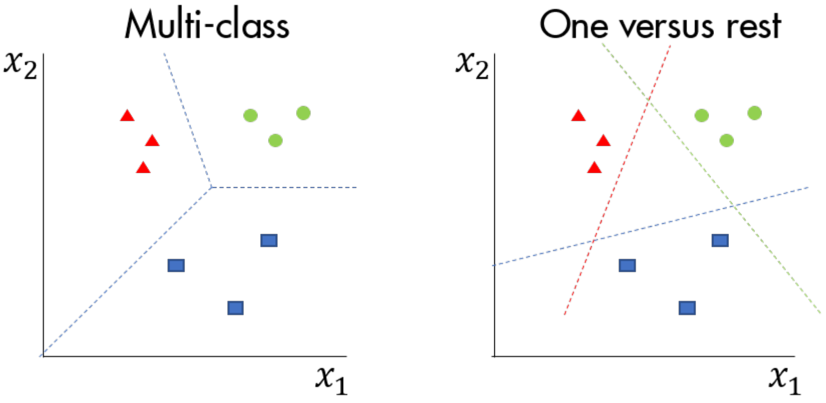


Figure 4.7: Example of classification strategies used in logistic regression. Left: multi-class classification. Right: one versus the rest binary classification.

In the failure diagnosis model for control valves, multi-class classification is performed, since there are five failure categories available. The maths behind the multinomial regression, using softmax regression, is implemented according to the book of Bishop (2006).

In the LR model, the regularisation parameter 'C' can be alternated, which is calculated with the following formula: $C = \frac{1}{\lambda}$, where λ controls the trade-off between the complexity of the model on the training set. High values of λ corresponds to a simple model, which tends to underfit. With low values of λ, the complexity of the model increases, possibly causing overfitting on the training data. The 'C' value works precisely the other way around.

4.4.2. Gaussian Naive Bayes

The Naive Bayes classification algorithm is a probabilistic classifier used in ML, which is based on Bayes theorem. Bayes Theorem is named after Reverend Thomas Bayes, who proposed to apply conditional probability that uses evidence to determine limits on an unknown parameter. The method calculates the posterior probability, using the likelihood, class prior probability, and the predictor prior probability. This means the probability of the observation being in a specific class given the observation of the input features (Zhang, 2004). Various variations to Bayes theorem classification algorithms are available, such as Bernoulli, Multinominal, and Gaussian NB (GNB). Bernoulli NB is used for data distributed according to multivariate Bernoulli distribution. Multinominal NB is an algorithm for multinominally distributed datasets. GNB is used for predictors with continuous values, which are assumed to be sampled from a Gaussian distribution (Sulzmann et al., 2007). The GNB is used for failure diagnosis in control valves.

The model consists of the following parameters: classes $(c_1, ..., c_n)$ and features $(x_1, ..., x_n)$. First of all, the prior probability $\mathrm{P}(c)$ is determined using Equation (4.30), which is the probability of the event computed before the collection of new data.

$$\mathrm{P}(c) = \frac{\text{Number of instances of class c}}{\text{Total number of instances in the dataset}} \tag{4.30}$$

Afterwards the likelihood $\mathrm{P}(x_i|c_i)$ which is the likelihood of an evidence if the hypothesis is true, is calculated using Equation 4.31

$$\mathrm{P}(x_i|c_i) = \frac{1}{\sqrt{2\pi\sigma_{c_i}^2}} e^{\left(-\frac{(x_i - \mu_{c_i})^2}{N}\right)} \tag{4.31}$$

Where μ_c and σ_c are estimated using the maximum likelihood. The Gaussian Naive Bayes classifier assumes that all the features are independent of each other, therefore, Equation (4.32) is valid.

$$\mathrm{P}(x|c) = \prod_{i=1}^{n} \mathrm{P}(x_i|c) \tag{4.32}$$

Figure 4.8 shows an example of how the likelihood is used to fit the Gaussian distributions of class c for the observations of features x_1 and x_2.

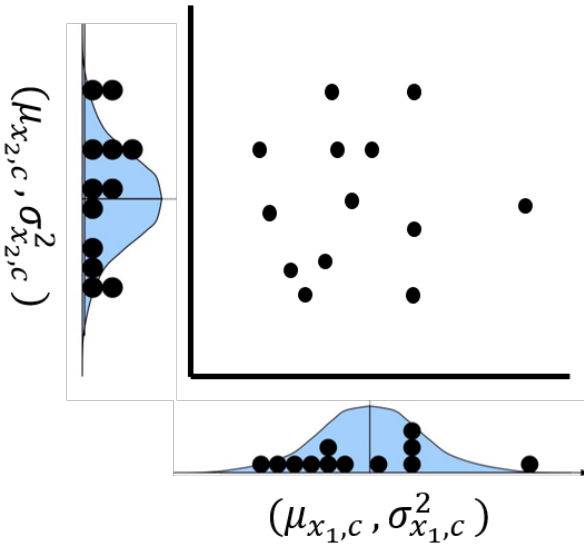

$(\mu_{x_2,c}, \sigma^2_{x_2,c})$
$(\mu_{x_1,c}, \sigma^2_{x_1,c})$

Figure 4.8: Example of how observations of features x_1 and x_2 are used to fit the Gaussian distribution for class c.

When the likelihood, prior probability of the class, predictor prior probability, and the likelihood is known, the posterior probability of every class c_i given the observations $(x_i, ..., x_n)$ can be determined using Equation (4.33).

$$P(c_i|x) = \frac{P(x|c_i) * P(c_i)}{P(x)} = \frac{P(x|c_i) * P(c_i)}{\sum_i^n P(x|c_i) * P(c_i)} \tag{4.33}$$

When applying Gaussian Naive Bayes on a two-dimensional dataset, the algorithm will separate the classes with a parabolic shape shown in Figure 4.9.

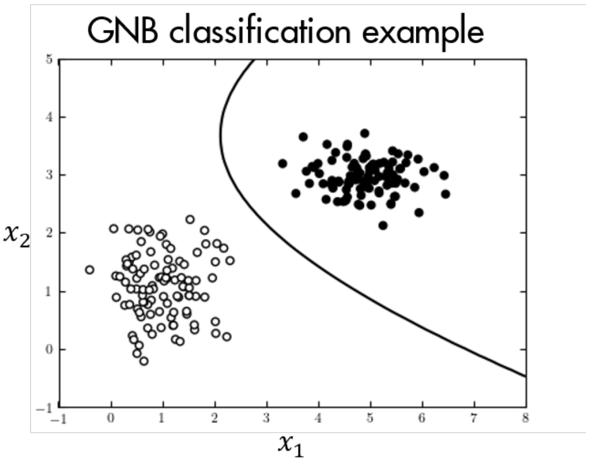


Figure 4.9: Example of the application of GNB on a two-dimensional dataset.

4.4.3. Random Forest

The supervised ML classification algorithm Random Forest is a meta estimator that fits a set of classification decision trees on various randomly drawn samples from the dataset. Each decision tree determines a class prediction, and the class, which is the most popular becomes the output of the model prediction.

The methodology behind decision trees should be known before going into the theory of the RF. Classification and Regression Trees (CART) is a definition introduced by Leo Breiman which introduces predictive classification or regression problems that can be solved by the decision tree algorithm. Classification trees attempt to distinguish classes by asking questions, which can be answered by true or false, related to one feature and one split point. Split points return two nodes, which can be pure or impure. Pure nodes do not need further splitting because the samples have been classified correctly. Impure nodes require more splits to gain more information on the data and classify the samples properly. The split of the decision tree on the feature corresponds to information gain (IG), which should be maximised to reduce the uncertainty as quickly as possible. Decision trees can apply various methods to determine how to split the observations, also referred to as the impurity criterion, such as the Gini index or entropy. IG is calculated, with the information of the parent and children node, using Equation (4.34).

$$\text{IG} = \text{information before splitting (parent node) - information after splitting (children node)} \tag{4.34}$$

When writing this in a mathematical formulation, we can define the information gain formula for binary splits with Equation 4.35.

$$\text{IG}(D_p, f) = I(D_p) - \frac{N_{\text{left}}}{N} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N} I(D_{\text{right}}) \tag{4.35}$$

Where f are the features to be split on. D_p is the dataset of the parent node, D_{left} is the dataset of the left child node, and D_{right} is the dataset of the right child node. I is the impurity criterion, such as the Gini index or entropy. N is the total number of samples, N_{left} is the number of samples in the left child node, and N_{right} is the number of samples in the right child node.

This research applies the Gini index, defined as the measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset, as impurity criterion. The Gini index is calculated using Equation (4.36).

$$I_G = 1 - \sum_{i=1}^{n} p_i^2 \tag{4.36}$$

Where p is the proportion of samples that belongs to class c for a particular node. When a node is pure, the impurity index is 0.

When going deeper into the decision tree, it becomes increasingly complex, which often leads to overfitting. Therefore, the tree depth, which is a measure of how many splits a tree can make before achieving a prediction, can be set to a maximum which stops the growth of the decision tree.

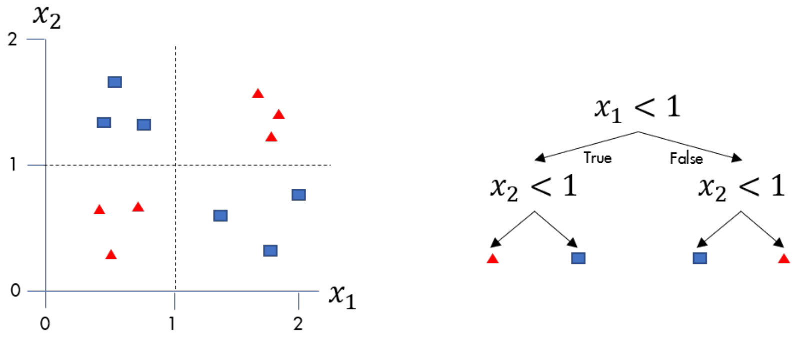


Figure 4.10: Example of how a decision tree classifies two categories. Left: Decision surface for two features $x_1 \& x_2$ and two categories red triangle, and blue square. Right: corresponding tree to the decision surface.

The RF algorithm is described by Breiman (2001) in Definition 1.

Definition 1. *A random forest is a classifier consisting of a collection of tree-structured classifiers h(x,*Θ_k*), k = 1,.. where* Θ_k *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.*

For the k-th tree, a random vector Θ is generated, which is independent of the past consisting of random vectors $\Theta_1, ..., \Theta_{k-1}$. Using the training set and the random vector Θ a tree can be grown, resulting in classifier $h(x, \Theta_k)$ where x is the input vector. For every split in the decision tree Θ consists of several independent random integers between 1 and K, which are a random set of input features. After a large number of trees has been created, the majority vote is chosen as the model output.

The steps used in the RF algorithm for a dataset with multiple features and observations are as follows. First of all, bootstrap samples, which refers to sampling with replacement, of the dataset, are randomly drawn. Second of all, decision trees are trained, where for every split a randomly chosen set of features is used to divide the samples. Third of all, the decision trees vote for their prediction. Finally, the most popular vote is chosen as the model output. Figure 4.11 shows an example of the steps taken in the RF classification.

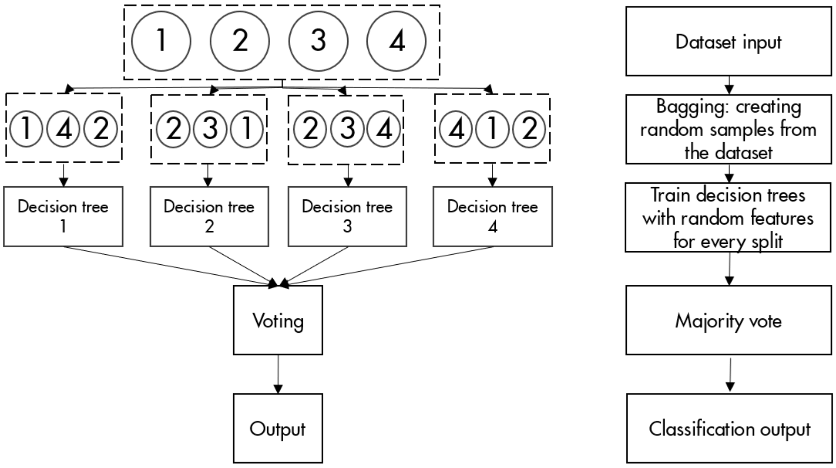


Figure 4.11: Example of the steps taken in the classification using the RF algorithm. Left: example of how a dataset is sampled, used to build decision trees which vote for their prediction, and how the most popular vote is chosen as the model output. Right: description of steps taken during the process on the left.

In the RF algorithm, two random selections are performed. First of all, random bootstrap samples are drawn from the training set, also referred to as bagging. Second of all, random subsets of features are drawn which can be chosen for every split to train the individual trees. The purpose of using two sources of randomness is to decrease the variance of the forest estimator, which reduces overfitting. Since individual decision trees typically show a high variance and tend to overfit. Since RF algorithms draw random samples and trains with random features for every tree, the predictions can be more apart, which can increase the bias somewhat more. However, since many decision trees are trained, and the majority vote for the classification output is applied the error will be reduced, allowing RF to obtain high accuracies.

4.5. Classification framework

In supervised ML classification problems, several steps are required to obtain optimal results given the dataset and algorithm chosen. First of all, the model is trained using k-fold cross-validation. Cross-validation is a resampling method used to assess the performance of ML models which have limited data available before using them on new data. After the cross-validation, depending on the results, the data input or parameters of the algorithm can be changed.

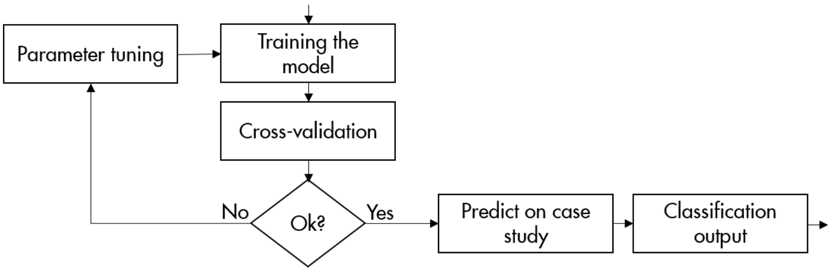


Figure 4.12: Stepwise framework of the application of supervised ML on classification problems after an algorithm has been chosen.

In k-fold cross-validation, the parameter k, which refers to the number of groups that a given data sample will be split in, can be chosen depending on the dataset size. The method is primarily applied to assess how the ML model will perform on unseen data. Cross-validation is used to prevent overfitting because the data is shuffled and randomly divided into k folds with the same size. Of the k folds, one sample is chosen as the test set, and k-1 folds are chosen as the training set. This process is repeated k times, where each of the k groups is used as a test set exactly once. After k estimations have been complete the results can be averaged into one single estimation (Schaffer, 1993).

In the failure diagnosis model for control valves, the size of the time-series t can be varied. Shorter time windows can improve the model performance for some categories. However, for others, it is the other way around. In order to obtain the optimal results depending on the wishes of the user, the effect of time-series length on the classification will be tested in Chapter 5. Parameters of the models that can be changed vary per algorithm. The most important parameters that can be changed in the RF model are the maximum number of features chosen when looking for the best split, the maximum depth of the decision trees, and the number of trees used in the forest. The metric used to evaluate the k-folds cross-validation is accuracy, which is the mean of all the k accuracies, shown in Equation 4.37.

$$\mu_{\text{cross-validation}} = \frac{1}{k}\sum_{i=1}^{k}\frac{\text{number of correct preditions}_i}{\text{total number of predictions}_i} \tag{4.37}$$

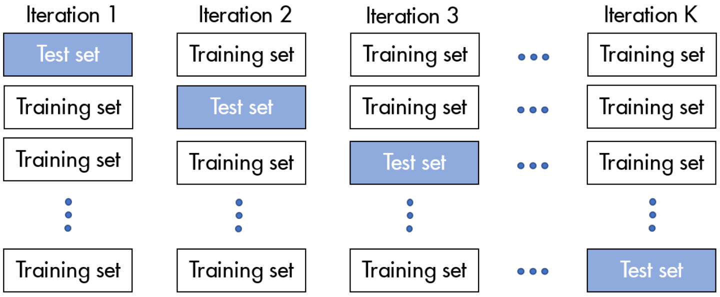


Figure 4.13: Example of how k-fold cross validation divides the dataset into k-1 training set and one test set.

After the quality of the model has been assessed several times using cross-validation, and the parameters and data input of the model has been changed, the optimal settings can be found. Once these are found, the model can start predicting new data in the form of a case study. These prediction results from the case study are also evaluated using several performance metrics found in Chapter 4.6.

4.6. Model evaluation

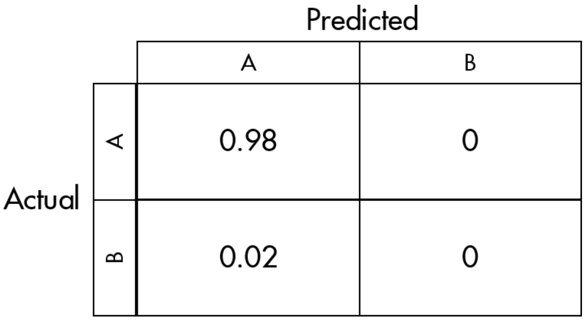
After the prediction of the classes using the supervised ML model has been made, the results can be analysed in order to assess the performance of the classifier. This evaluation is completed using several performance metrics: accuracy, log-loss, precision, and recall.

4.6.1. Accuracy

The first metric used to measure the performance of the supervised ML model is accuracy, which can be calculated using Equation 4.38.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4.38}$$

Which results in a value for accuracy between 0 and 1, where 0 means no correct predictions, and 1 is perfect prediction. Problems arise when using accuracy as an evaluation metric in a problem where the classes are very imbalanced for the dataset. For example, when class A consists 98 times in the data, and class B is represented only two times in the data. When just classifying all the 100 data points as class A, an accuracy of 98% would be achieved, which is very accurate. However, this result is deceiving, because when looking more closely, class B would not be represented in the predictions. Therefore, a confusion matrix can be applied, which visualises the performance of a classification model of which the true values are known. On the horizontal axis are the predicted classes, and on the vertical axis are the actual classes. The predictions in the confusion matrix are normalised by dividing the values through the sum of the observations available in the data. Figure 4.14 shows the normalized confusion matrix for the example described above. Using the diagonal of the confusion matrix, the accuracy of each class can be determined to give a better overview of the actual model performance.



		Predicted	
		A	B
Actual	A	0.98	0
	B	0.02	0

Figure 4.14: Example of a normalized confusion matrix, where the accuracies of the individual predictions can be found on the diagonal axis. The example shows how a model with 98% accuracy can deceive the user of the model, since class B is completely misclassified.

The failure diagnosis model for control valves uses accuracy as a performance metric. However, in combination with confusion matrices to ensure correct model evaluation.

4.6.2. Log-loss

The second performance metric used is the Logarithmic loss, known as a measure of the goodness of probability estimates, is similar to cross-entropy measuring the difference between two probability distributions for a given observation of a random variable. The Log-loss value increases, when the probability of the prediction diverges from the actual label, by taking the uncertainty of the prediction into account. Therefore, users of a classification model will strive to obtain a log-loss performance of near zero (Ferri et al., 2009).

$$\text{Log-loss} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log(p_{ij}) \tag{4.39}$$

Where N is the number of observations, c is the total number of classes, y_{ij} is a binary value 0,1 indicating whether observation i belongs to class j, and p_{ij} is the probability that observation i belongs to class j.

Equation 4.39 requires the estimation of the probability that i belongs to class j. How this estimation works depends on the supervised ML model applied. LR uses a logistic function to estimate the probabilities of an observation belonging to a specific class. GNB estimates the probabilities of an observation belonging to each class using the likelihood of the Gaussian distributions. RF estimates the probabilities using another method, it determines the percentage of individual outcomes of the tree for a particular class and divides that through the total amount of trees resulting in the probability of that class.

4.6.3. Precision

When using failure diagnosis to classify failures automatically, different strategies can be applied. The goal of the classification can be to diagnose the behaviour into one of the five categories. However, it can also be the intention to diagnose for one specific failure type which should be prevented at all times. When using this strategy, the performance of the model can be assessed using Precision and recall. Precision is the fraction of relevant predictions divided by the total amount of predictions. In the case of failure diagnosis, when looking for specific failure behaviour. The precision would be determined by dividing the correct predictions for a specific failure type by the total amount of predictions for that specific failure type. The recall will be discussed in the next section. When creating a confusion matrix, shown in Figure 4.15, four classification types can be determined: true positives, false positives, false negatives, and true negatives. On the horizontal axis, the predicted values can be found, and on the vertical axis are the actual values.

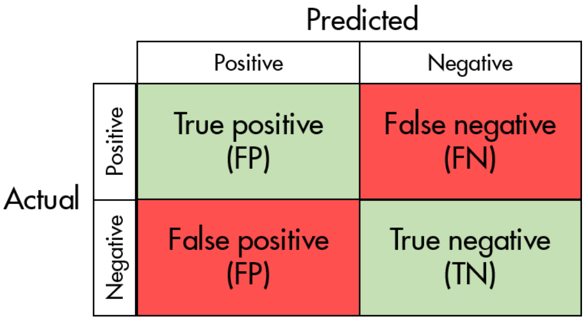


Figure 4.15: Example of confusion matrix with the corresponding classification types. On the horizontal axis are the predicted values, and on the vertical axis are the actual values.

When calculating precision using the classification types and looking for the diagnosis of a specific failure, Equation 4.40 is used.

$$\text{Precision} = \frac{\text{correctly predicted positive predictions}}{\text{Total number of positive predictions}} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.40}$$

This function can be tranformed into a mathematical formulation using Equation 4.41.

$$P(y_l, \hat{y}_l | l \in L) \tag{4.41}$$

Where y_l is the subset of the true classes with label l, and $\hat{y}_l$ is the subset of the predicted classes with label l. L is the set of labels, and $P(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|\hat{y}_l|}$.

4.6.4. Recall

Recall, also referred to as the hit-rate, measures the ability of a classification model to find all the positive samples, determined with the fraction of the relevant instances that were actually retrieved. When for a certain valve, a specific failure type should be diagnosed at all times, the recall should be maximised. The value for recall can be determined using the classification types denoted in Figure 4.15, shown in Equation 4.42.

$$\text{Recall} = \frac{\text{correctly predicted positive predictions}}{\text{correctly predicted positive predictions + number of missed failures}} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.42}$$

This can be transformed in a more mathematical formulation using Equation 4.43.

$$R(y_l, \hat{y}_l | l \in L) \tag{4.43}$$

Where y_l is the subset of the true classes with label l, and $\hat{y}_l$ is the subset of the predicted classes with label l. L is the set of labels, and $R(y_l, \hat{y}_l) = \frac{|y_l \cap \hat{y}_l|}{|\hat{y}_l|}$.

4.7. Conclusion

Several features, describing the failure behaviour with single values, are extracted from the time-series data to distinguish between classes. Ratios, correlations between signals and other features are created. The features are all logically based on analysis of the historical failure behaviour data, and therefore the outcome can be supported with reasoning on the value outcome. However, to prevent the occurrence of noise in the model due to a surplus of features, two selection methods have been determined: the linear ANOVA test, and the non-linear Mutual Information test. These tests are applied to determine which features are significant predictors for the output.

After the selection, the supervised ML classification will be exercised using three algorithms, which are determined using several criteria dependent on the data available and optimal performance situations for the models. The first classification algorithm is LR, which uses a logistic function to determine the probability of a class given the observation. The second classifier is GNB, which determines the posterior probabilities of every class given an observation. The third classification model is RF, consisting of a set of decision trees, where the wisdom of crowds is the fundamental concept. Two selection and three model options have been chosen, which will result in six different prediction outcomes. The use of ML classification models, after the algorithm has been chosen, requires two steps: First of all, cross-validation is performed on the training set to choose the optimal model, and tune the parameters and data input for the operation. Second of all, the optimal model is applied to new data with a case study. The predictions in the model output will be assessed using performance metrics: accuracy and log-loss. Depending on the strategy used in the failure diagnosis, the quality of the model will be evaluated with these indicators.

5

Results

Chapter 5 consists of the performance tests on the classification of control valve failures in five behaviour categories. Several tests will be applied to determine the achievements made by the model depending on the wanted outcome by the user. In Chapter 5.1, the training set will first be used with cross-validation to choose and tune the model for optimal performance. Chapter 5.2 contains the test results of the case study discussed in Chapter 3.4. Finally, the performance results will be validated with a comparison to the diagnosis process by hand and the current state of the literature in Chapter 5.4.

5.1. Optimal method determination

The time-series data used as input requires preparation before being able to classify the behaviour. Furthermore, the parameters of the classifiers should be adjusted to obtain optimal conditions measured using cross-validation. Therefore this section contains more information on the preparation of the training set, which is used to test the case study in Chapter 5.2.

5.1.1. Input overview

Chapter 3.3 describes the shape and content of the training set consisting of 50 critical failure cases ranging from a period between 2014 and 2020. Every failure case delivers three signals: OP, PV, and SP, bringing the total amount of signals used in the model to 150. The data points where the control valve is in manual mode are rejected from the model input.

Time window t [minutes]
200
1000
5000

Table 5.1: Size of time series used as input for the feature extraction in minutes.

The output categories $(y_i, .., y_N)$, where N is the total number of samples, represented by the failure categories, are not evenly distributed in the training set.

Output category	Number of samples
Fail to close	15
Fail to open	11
Hunting	12
Hysteresis	7
Other	5

Table 5.2: Size of time series used as input for the feature extraction in minutes.

5.1.2. Extraction

The extraction of features from the input data is performed according to Chapter 4.2. The result of the feature extraction is a 50 by 13 matrix x_{ij} for $(i = 1,..,50)$ and $(j = 1,..,13)$ representing the failure behaviour of each sample.

5.1.3. Selection

The features that are significant predictors for the output are selected in the model. This selection and is made using the ANOVA and Mutual information tests described in Chapter 4.3, and the significant predictors are adopted in the model. A confidence interval of 95% is chosen to determine the significance.

The critical F-value can be determined using an F-table, which requires the degrees of freedom of the sum of squares within and between the variables, and the confidence interval. The one-sided ANOVA test is used. Therefore, $\alpha = 0.025$ is used in the table. The degrees of freedom between the variables is $N-1$, which corresponds to the (number of categories - 1): $5-1 = 4$. The degrees of freedom within the variables is $5*(50-1) = 245$. When using $\text{df}_1 = 4$, $\text{df}_2 = 245$, and $\alpha = 0.05$ the critical F-value is 2.41. When the F-value between the variable and output is above this value, the null hypothesis is rejected, and the feature is a significant predictor.

The critical value from the Mutual information test, above which the feature is a significant predictor, is determined with a set of 1,000 noise variables. The Mutual information test is performed on the noise variables with a normal distribution, and since a confidence interval of 95% is chosen, the 95-quartile value of the results is chosen as the critical Mutual information value for the features. The determined value is $I = 0.154$. All the features that obtain a Mutual information value higher than this are significant predictors for the output, thus adopted in the model.

Figure 5.1 and 5.2 contain an overview of the two feature selection tests for a time window input of 200 minutes. The blue lines contain critical values. The ANOVA test selects six features: Mean OP, (PV-SP)/SP, Correlation SP PV, Correlation PV OP, Correlation SP OP, and Abs difference PV OP. The Mutual information test adopts seven features: Mean OP, Variance SP/Variance PV, (PV-SP)/SP, Correlation SP PV, Correlation PV OP, Correlation SP OP, and Abs difference PV OP The full results of the feature selection tests can be found in Appendix B.1 and B.2.

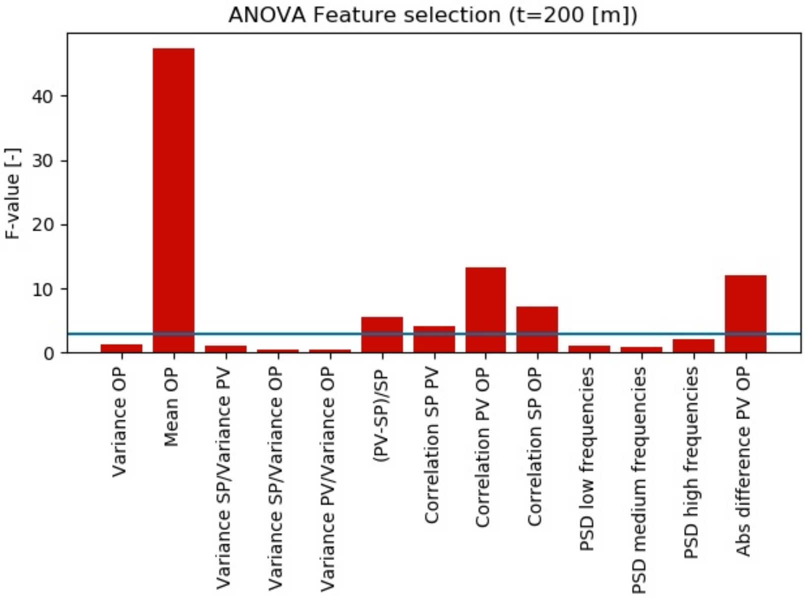


Figure 5.1: Results of ANOVA test used for feature selection using a input time-window of $t = 200$ [s]. Red bar plots show the F-values for the 13 features. Blue horizontal line shows the critical F-value: $F_{crit} = 2.4$. Features with an F-value above the blue line are adopted in the supervised ML model.

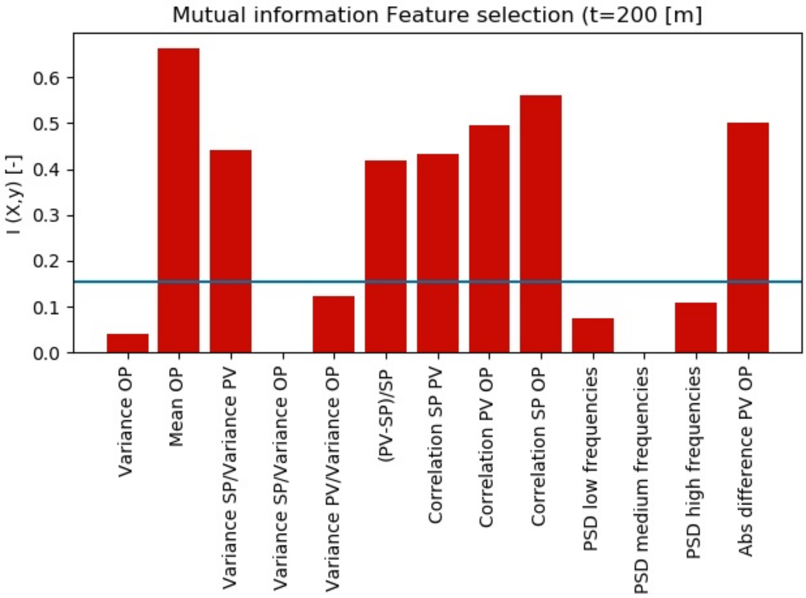


Figure 5.2: Results of the Mutual information test used for feature selection using a input time-window of $t = 200$ [s]. Red bar plots show the F-values for the 13 features. Blue horizontal line shows the critical I value $I = 0.15$. Features with a higher Mutual information value than the blue line are adopted in the supervised ML model.

The remaining features of the two selection methods are inspected using simple 2d scatter plots before being used as input of the ML classifier to ensure interpretability of the classification outcome. Figures 5.3 and 5.4 show these scatter matrices, where the first one consists of six features and the second one of 7 features. The most important scatter plots, where a clear difference between the failure behaviour in the features data, are highlighted and discussed afterwards.

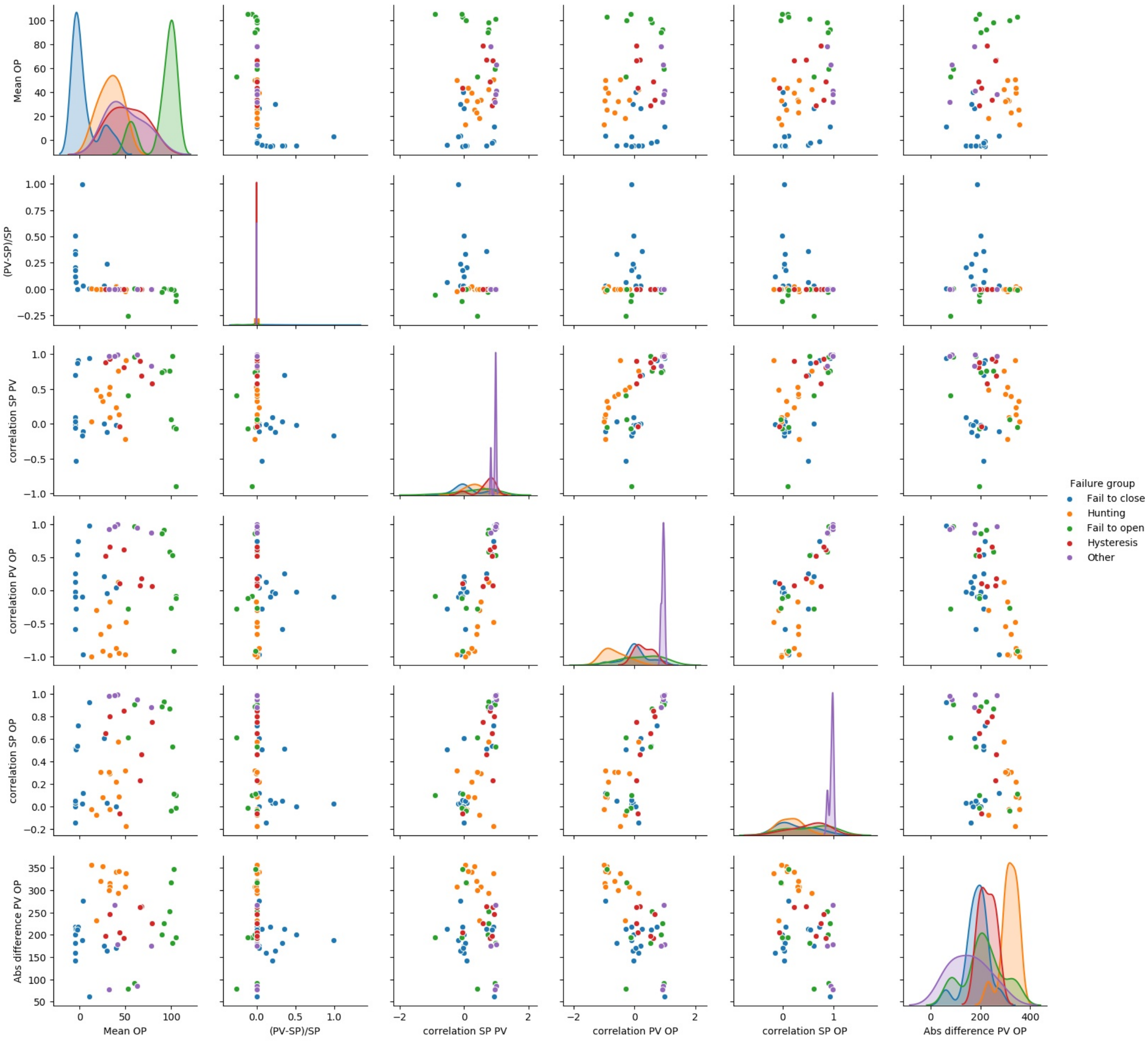


Figure 5.3: Scatter matrix of containing the features after the ANOVA selection method. Six features adopted after picking the significant predictors: mean OP, (PV-SP)/PV, 3 correlations between the signals, absolute difference PV OP.

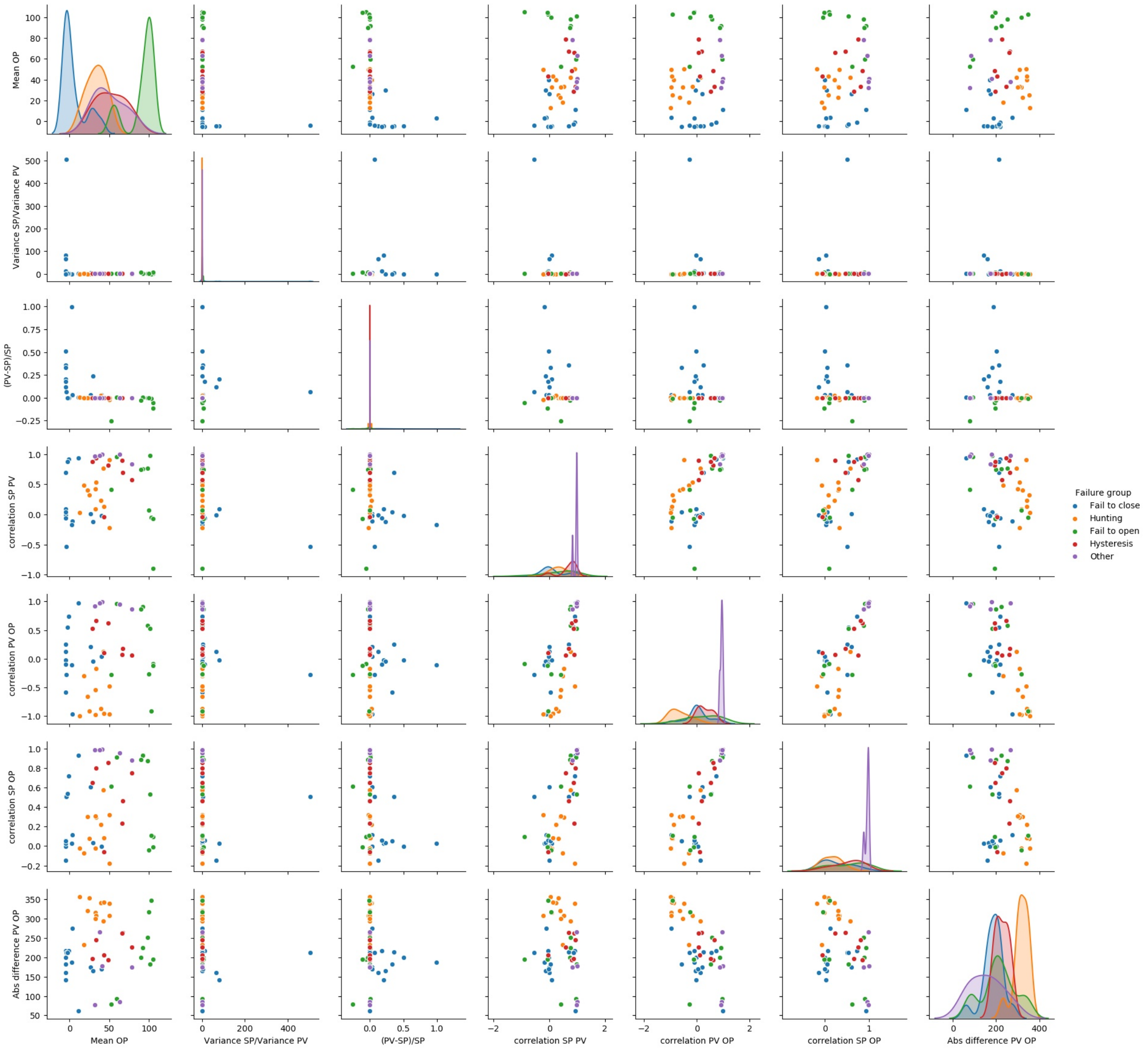


Figure 5.4: Scatter matrix of containing the features after the Mutual information selection method. Seven features adopted after picking the significant predictors: mean OP, variance SP / variance PV, (PV-SP)/PV, 3 correlations between the signals, absolute difference PV OP.

The first highlighted scatter plot, shown in Figure 5.5 is the mean OP versus the ratio between the PV and the SP (PV-SP/SP). The blue oval presents the majority of the ’Fail to open’ behaviour, where the average value for the OP is relatively high compared to the other categories. Furthermore, the ratio between the PV and the SP is negative due to the inability of the controller to open the valve fully, causing the gap between the two signals. The red oval shows the majority of the ’Fail to open’ category, where the average OP value is lower than the others. Moreover, the ratio between the PV and the SP is positive, due to the valve failing to close properly. The rest of the failure behaviour categories mostly show average OP values between 20 and 80. Also, the ratio between the PV and the SP is mostly close to zero.

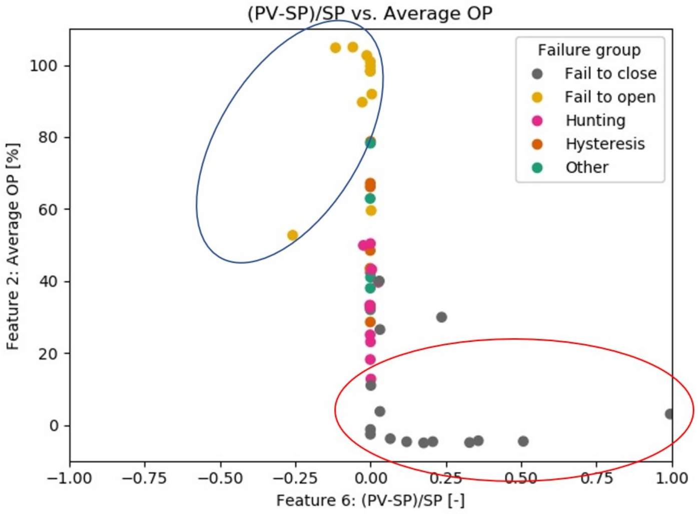


Figure 5.5: Scatter plot mean OP versus (PV-SP/SP). Blue oval shows the majority of ’Fail to open’, red oval shows the majority of ’Fail to close’.

The second scatter plot, shown in Figure 5.6 compares the correlation of the PV and the OP to the absolute difference between the PV and the OP. In the blue oval area are the majority of the ’Hunting’ data points, due to the low correlation and big absolute difference between the PV and the OP signal. The ’Hysteresis’ cases, in the red oval, have an intermediate absolute difference and correlation between the PV and OP. The ’Other’ cases in the green show an extremely high correlation between the PV and OP, due to the fact it controls well; however, it is broken.

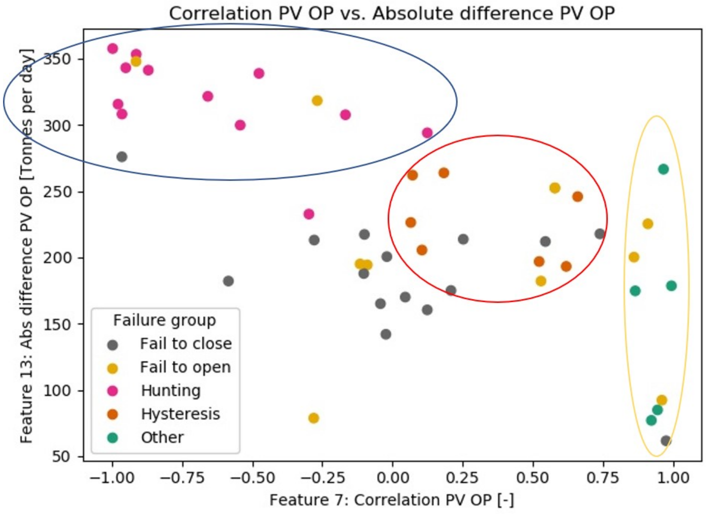


Figure 5.6: Scatter plot of the correlation between the PV and OP, and the absolute difference between the PV and OP. Blue oval shows the majority of ’hunting’, red oval shows the majority of ’Hysteresis’, and the yellow oval shows the majority of ’Other’ behaviour.

The third scatter plot, shown in Figure 5.7, compares the correlation between the PV and OP to the mean OP. For the majority of each failure category, a decent separation can be made when assessing these two features. The low and high average values of the OP are the 'Fail to close' and 'Fail to open' categories. The middle section with an intermediate PV OP can be split into three sections due to the large difference in correlation between the OP and PV.

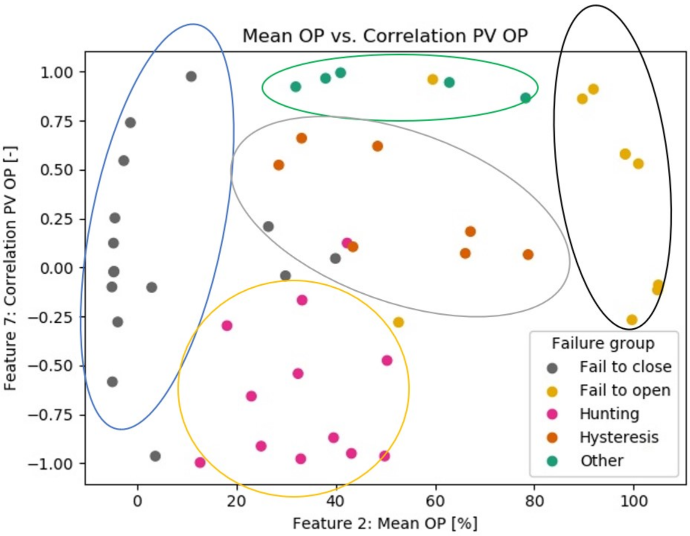


Figure 5.7: Scatter plot of the correlation between the PV and OP, and the mean OP. Blue oval shows the majority of 'Fail to close', green oval shows the majority of 'Other', yellow oval shows the majority of 'Hunting', grey oval shows the majority of 'Hysteresis', and the black oval shows the majority of 'Fail to open' behaviour.

The fourth scatter plot, shown in 5.8, compares the correlation between the PV and OP and the SP and OP. In this plot, it is clearly proved that the 'Other' category has values close to one for both of these features, due to the regular performance of the valve control. The yellow oval highlights the majority of the 'Other' category.

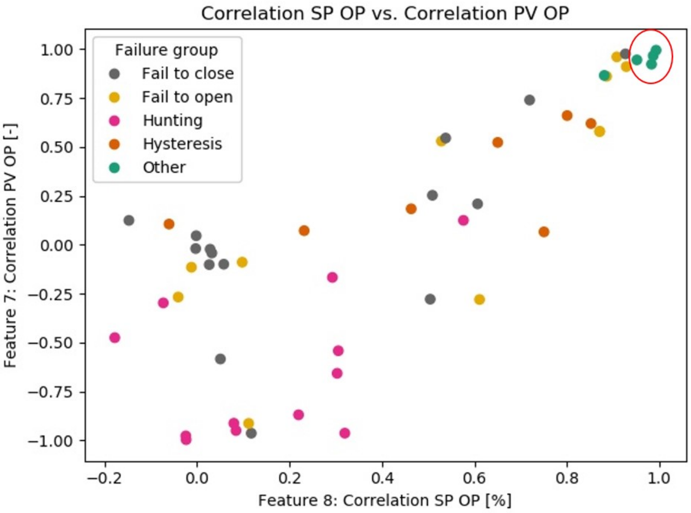


Figure 5.8: Scatter plot of the correlation between the SP and OP, and the correlation between the PV and OP. The green oval highlights the 'Other' behaviour.

5.1.4. Cross-validation performance evaluation

The results of the selection methods develop multiple input sets for the classifier models LR, GNB, and RF. This section evaluates the performance results of the training set using cross-validation to tune the parameters, length of the data input, and choose the best optimal model for the automatic failure diagnosis using the KPIs: accuracy and log loss. The time-window of the signal input is varied between 200, 1000, and 5000 minutes. The features used for every input are mentioned in the Tables in Appendix B.1.

The tuning parameters of the three models used are different. In the LR model, the regularisation value 'C' can be set, as described in 4.4. The GNB model does not require parameter tuning. In the RF model, the maximum depth of the decision trees can be altered, which changes the complexity of the model, and can prevent overfitting on the training set as described in 4.4. Tables 5.3 and 5.4 contain the results of the accuracies determined using cross-validation on the training sets with the ANOVA and MI methods. The accuracies shown are the mean of the cross-validation accuracies, and the standard error of the mean of the cross-validation accuracies: $\mu_{\text{cross-validation}} \pm \frac{\sigma_{\text{cross-validation}}}{\sqrt{N}}$.

Time input	Number of features	LR (C=0.1)	LR (C=1)	LR (C=100)	GNB	RF (max depth = 5)	RF (max depth = 10
t = 200	6	0.74 ± 0.03	0.74 ± 0.06	0.70 ± 0.03	0.76 ± 0.03	0.84 ± 0.03	0.88 ± 0.06
t = 1000	5	0.76 ± 0.07	0.74 ± 0.08	0.76 ± 0.09	0.74 ± 0.05	0.80 ± 0.06	0.82 ± 0.06
t = 5000	5	0.74 ± 0.06	0.74± 0.04	0.74 ± 0.03	0.79 ± 0.02	0.76 ± 0.03	0.76 ± 0.03

Table 5.3: Accuracy results of the cross validation on the training set for the ANOVA selection.

Time input	Number of features	LR (C=0.1)	LR (C=1)	LR (C=100)	GNB	RF (max depth = 5)	RF (max depth = 10
t = 200	7	0.74 ± 0.03	0.76 ± 0.05	0.66 ± 0.02	0.62 ± 0.01	0.76 ± 0.03	0.72 ± 0.01
t = 1000	9	0.68 ± 0.09	0.64 ± 0.1	0.62 ± 0.11	0.58 ± 0.06	0.75 ± 0.02	0.74 ± 0.03
t = 5000	6	0.78 ± 0.04	0.78 ± 0.02	0.78 ± 0.02	0.68 ± 0.07	0.80 ± 0.01	0.81 ± 0.03

Table 5.4: Accuracy results of the cross validation on the training set for the Mutual information selection.

In the case of LR the cross validation accuracies of the results using ANOVA and Mutual information are very similar, with a wide confidence interval for t=1000 minutes, especially for the Mutual information results. An explanation for these large confidence interval at the given time window can be explained by the large number of features used in the model. Changing the regularization coefficient does not affect the accuracy results significantly. Using the GNB model, the accuracies of the ANOVA method are significantly higher, due to the higher amount of features present that cause noise to the classification results. In the RF model results, the differences in accuracy are not very large, which can be related to the RF model being less sensitive to noise features. The best result for accuracy, using RF, is obtained when using a time window of 200 minutes and a maximum tree depth of 10. In order to visualize the cross validation accuracies of the model options, two heatmaps have been created, shown in Figures 5.9 and 5.10.

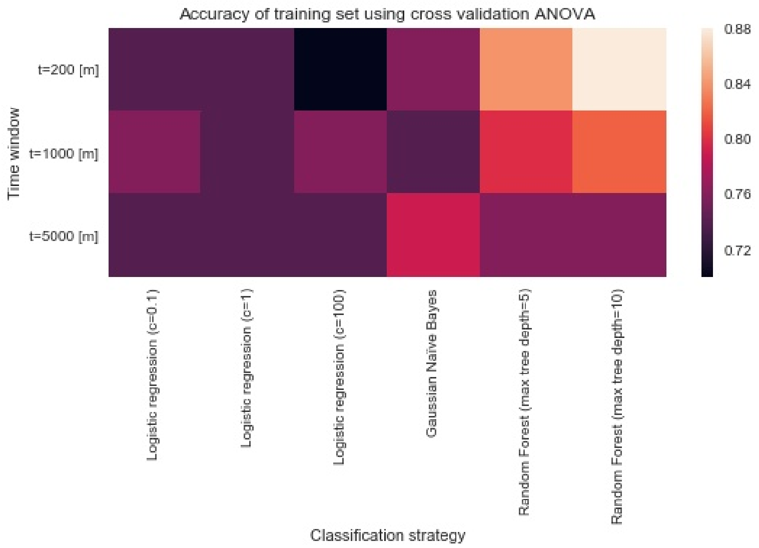


Figure 5.9: Heat map of the accuracy determined using cross validation with the features resulting from the ANOVA test.

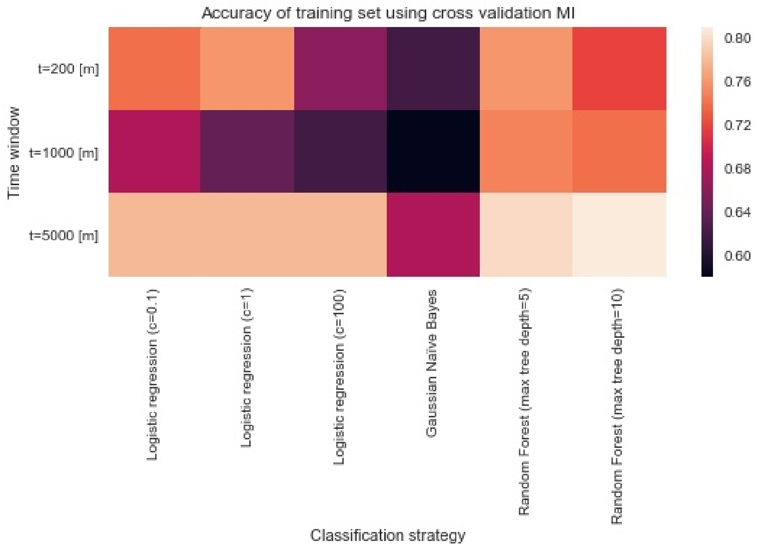


Figure 5.10: Heat map of the accuracy determined using cross validation with the features resulting from the Mutual information test.

The second KPI used to analyse the cross-validation is log loss, which has been discussed in 4.6. The log loss metric measures the uncertainty in the classification probabilities and calculates a value between 0 and 1. Where 0 means no uncertainty in the model and thus, complete classification. Tables 5.5 and 5.6 contain the results of the log loss performance using cross-validation.

Time input	Number of features	LR (C=0.1)	LR (C=1)	LR (C=100)	GNB	RF (max depth = 5)	RF (max depth = 10
t = 200	6	0.17	0.13	0.08	0.035	0.43	0.49
t = 1000	5	0.12	0.087	0.045	0.06	0.47	0.47
t = 5000	5	0.23	0.22	0.16	0.11	0.56	0.55

Table 5.5: Log loss results of the cross validation on the training set for the ANOVA selection.

Time input	Number of features	LR (C=0.1)	LR (C=1)	LR (C=100)	GNB	RF (max depth = 5)	RF (max depth = 10
t = 200	7	0.17	0.16	0.047	0.01	0.59	0.58
t = 1000	9	0.12	0.049	0.028	0.03	0.58	0.57
t = 5000	6	0.48	0.51	0.5	0.11	0.49	0.48

Table 5.6: Log loss results of the cross validation on the training set for the Mutual information selection.

The log loss results, show that for LR and GNB the uncertainty in the model increases due to the increasing time window, which can be explained by noise in the time series due to normal behaviour mixing up with failure behaviour. When increasing the regularization coefficient C the uncertainty increases due to the variance of the model increasing and the bias reducing. In the RF model the uncertainty increases for the ANOVA method when the time window grows. For the Mutual information method it is the other way around. Figures 5.11 and 5.12 show heat maps of the results to visualize the trends.

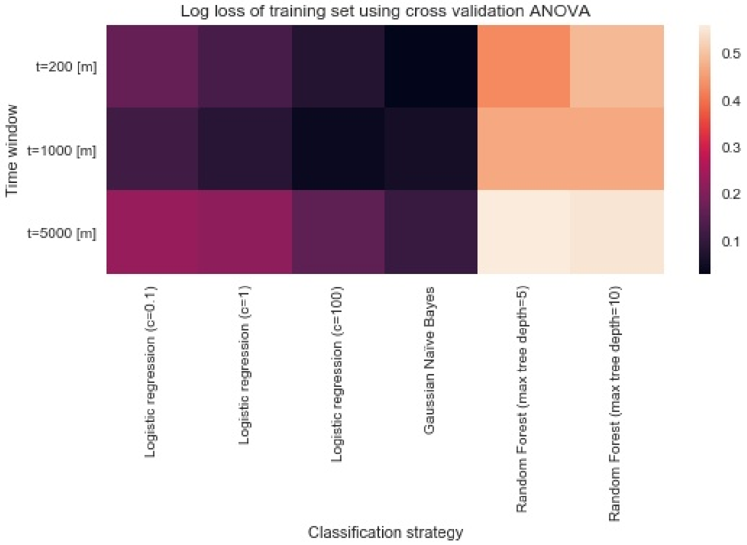


Figure 5.11: Heat map of the log loss determined using cross validation with the features resulting from the ANOVA test.

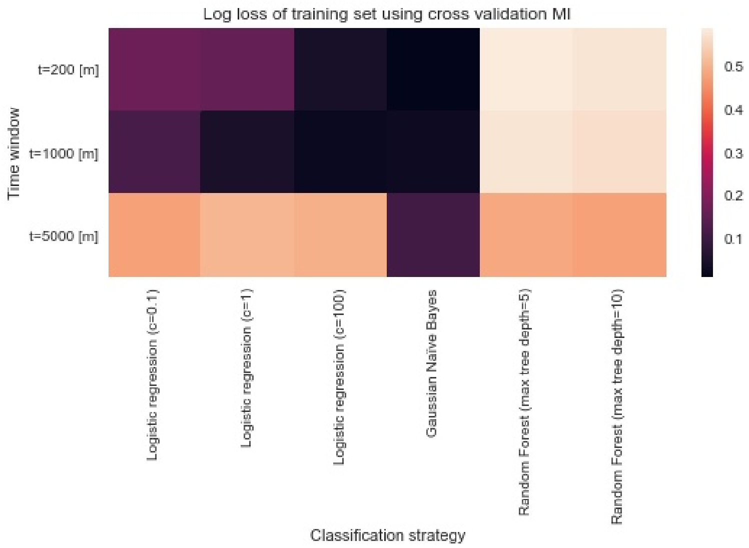


Figure 5.12: Heat map of the log loss determined using cross validation with the features resulting from the Mutual information test.

5.1.5. Conclusion

In order to find the best time window, model parameters, and optimal model for the classification problem, the algorithms are evaluated using cross-validation on the training set with the two KPIs: accuracy and log loss. The criteria for an optimal model and corresponding settings are a high accuracy from the cross-validation without the uncertainty being too large in the model. The RF model with a maximum tree depth of 10 using the ANOVA selection method for a time window of $t = 200$ minutes is chosen as the best model for automatic failure diagnosis for control valves due to several reasons. First of all, the model obtains an accuracy which is significantly higher than the other results. The uncertainty is larger than the other algorithms. However, this is due to the mechanism of the algorithm where two random draws are made from the samples and features for every decision tree. Second of all, we chose a smaller time window as there is no significant improvement in the accuracy for larger time windows and a shorter time window reduces the risk of indulging normal operations. Since for the case of automatic failure diagnosis, it is known that the failure behaviour occurs, the time window does not have to be longer than 200 minutes. Therefore, RF using ANOVA feature selection has been chosen as the optimal method for the classification and prediction of failures in flow control valves.

5.2. Case-study

Due to the results of cross-validation on the training set, the RF method, in combination with the ANOVA feature selection method, has been chosen as the optimal method for automatic failure diagnosis in control valves. This section contains the classification on the case study test data using the ANOVA feature selection and RF classifier. First of all, the model input overview will be given. Second of all, the performance will be evaluated using the determined KPI's. Finally, the performance will be validated and the last chapter consists of a discussion on the results.

5.2.1. Overview

The failure cases used in the case study are described in Chapter 3.4, which sums up to 16 incidents. Table 3.5 shows the specific failures and corresponding location summing up to the distribution of samples, displayed in Table 5.7.

Failure behaviour	Number of samples
Fail to close	3
Fail to open	3
Hunting	3
Hysteresis	4
Other	3

Table 5.7: Distribution of samples in the five failure behaviour categories.

In Chapter 5.1 the input length of the time series $t = 200$ minutes has been determined. With the ANOVA method, the corresponding six features are: mean OP, (PV-SP/SP), correlation SP PV, correlation SP OP, correlation PV OP, and the absolute difference between the PV and the OP.

The classification output performance will be evaluated using KPI's. Furthermore, the classification probabilities, which denotes the probabilities of an observation belonging to a particular class, are used to assess the decisions of the classification model. Appendix B.3 contains scatter plots of the selected features in the cases study.

5.3. Performance results

In the RF model, a set of 100 decision trees are used to train and predict the failure behaviour of the control valves. Since every decision tree randomly draws samples for every tree from the dataset and the best feature out of a random set features for every node, every tree is different. However, Figure 5.13 shows one of the decision trees trained on a sample of the training set.

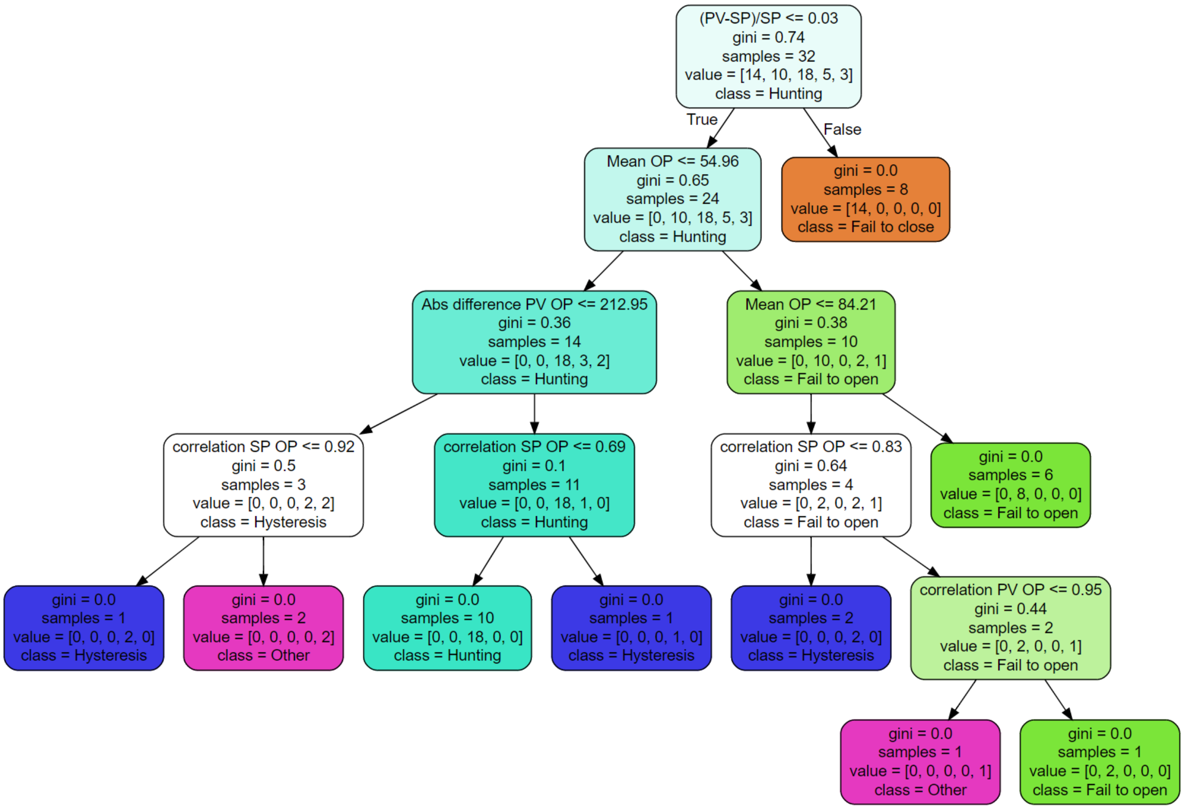


Figure 5.13: Example of a decision tree present in the RF classification model.

Features used more often on top of the decision trees, where more information can be gained. Thus more samples can be classified of higher importance to the model. Analysing these features is used to determine which features are more or less critical to the prediction outcomes. The results of this analysis are shown in Figure 5.14, where the importance is a fraction that sums up to 1. The average OP feature is the most important, with a fraction of almost 0.3. The correlation between the SP and PV signal is used the least to make important splits in the decision trees with a fraction of under 0.10.

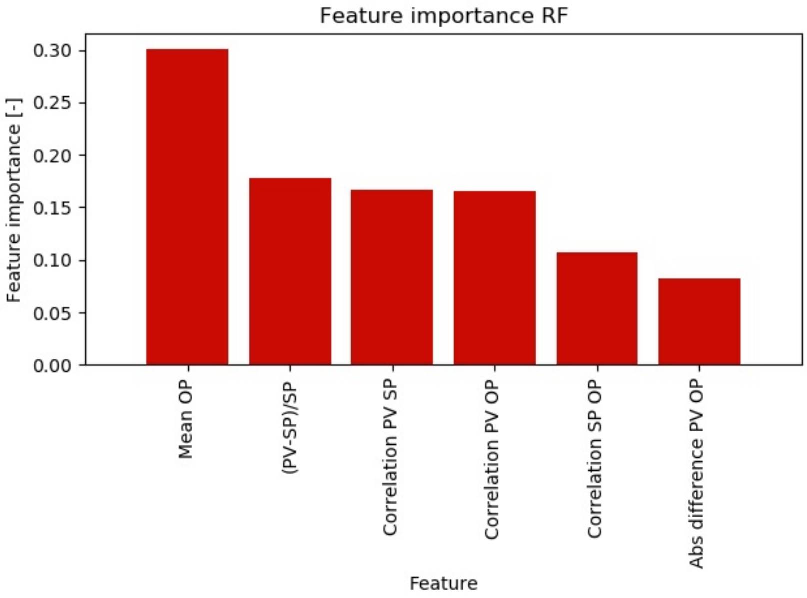


Figure 5.14: Feature importance, shown as a fraction that sums up to 1, of the RF model used with ANOVA selection.

The performance of the case study is assessed using the metrics discussed in Chapter 4.6: accuracy, log-loss, precision, and recall.

First of all, the accuracy and log-loss of the model are assessed for the GNB classifier with the ANOVA selection method. Second of all, the precision and recall values of the outcomes will be discussed. Finally, the classification results will be evaluated in more detail. The results of the accuracy and log-loss are shown in Table 5.8.

	Accuracy	Log-loss
RF	0.81	0.46

Table 5.8: Results of classifier on accuracy and log-loss.

The RF classifier obtains an accuracy of 0.81 on the classification and prediction of the case study results, which is slightly lower than the accuracy performance of the cross-validation. Possibly due to the SGHP being an unseen plant in the training set. The log loss value on the case study set is 0.46, somewhat better than the result on the training set, and thus less uncertainty in the classification. The case study is also done using three output categories: 'Fail to close', 'Fail to open', and 'Other'. This increases the accuracy of the predictions to 0.88.

The results of the KPIs precision and recall on the new incident data in the case study are shown in Table 5.9

	Precision	Recall
Fail to close	1	1
Fail to open	0.6	1
Hunting	0.67	0.67
Hysteresis	1	0.75
Other	1	0.67
Weighted	0.86	0.81

Table 5.9: Results of classifier on precision and recall.

The precision of the RF classifier on the categories 'Fail to close', 'Hysteresis', and 'Other' is 1, which means that these three classes are never predicted when another failure is present, thus the model classifier is very precise for these categories. The categories 'Fail to open' and 'Hysteresis' are sometimes mixed up by the classification model.

The recall is the ability of a classification model not to miss a particular class. In the recall results, of the RF classifier, the 'Fail to close' and 'Fail to open' category score 1. This result is satisfying, if the goal would, for example, be to prevent a valve failing to close at all times the model would pass at all times given these results. The results show that for every class, the model classifies the failure behaviour correctly for at least 67% of the time.

Apart from the KPI's class predictions can be assessed individually using a confusion matrix, where on the horizontal axis are the predictions and actual labels on the vertical axis. Figure 5.15 shown a confusion matrix of RF classification using the ANOVA selection. On the diagonal axis, are the fractions of correct predictions. For the classes 'Fail to close' and 'Fail to close' the prediction is 1, shown in Table 5.9. In order to analyse the misclassification's, the scatter plots of Appendix B.3. are used. The category 'Hysteresis' is classified as 'Fail to open' in prediction number 1. Using 5.10, the probability of predicting 'Hysteresis' was 27%, which means the classification model was on the right track. In prediction 1, this misclassification is done the other way around for 'Hunting', with a probability for 'Hunting' of 30%. These two categories are mixed up when the mean OP has an average value. Therefore, the features obtain similar values. The last misinterpretation is number 16, where 'Other' and 'Hunting' are misclassified. When inspecting the scatter matrices, we can conclude that one is on the border of 'Hunting' and 'Other' when looking at the features correlation PV OP and the absolute difference between the PV and OP. This is presumably where the mistake is. 'Other' does still get a probability of 0.20. The conclusion on this more in-depth analysis on the mistakes shows that RF, while it obtains an accuracy of 0.81, gives a prediction probability of at least 0.20 to the actual class.

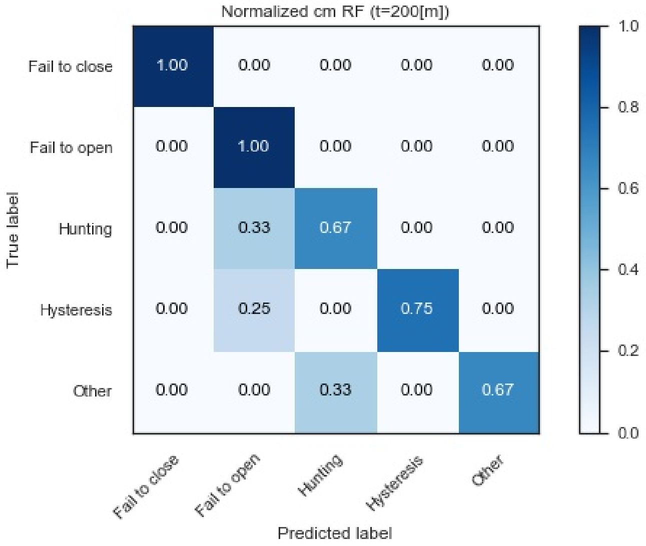


Figure 5.15: Confusion matrix of the case study using the RF classifier in combination with the ANOVA feature selection method. On the horizontal axis are the prediction labels and on the vertical axis are the actual labels.

	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0	0.65	0.05	**0.27**	0.03
2	Hysteresis	Hysteresis	0.18	0.09	0.23	**0.48**	0.02
3	Fail to close	Fail to close	**0.79**	0.05	0	0.01	0.15
4	Other	Other	0.03	0.34	0	0.19	**0.44**
5	Hunting	Fail to open	0.01	0.68	**0.3**	0.01	0
6	Other	Other	0.04	0.37	0	0.12	**0.47**
7	Fail to close	Fail to close	**0.79**	0.06	0	0.15	0
8	Hunting	Hunting	0.06	0.08	**0.86**	0	0
9	Hysteresis	Hysteresis	0.23	0.07	**0.19**	0.45	0.06
10	Fail to close	Fail to close	**0.77**	0.02	0.19	0.45	0.06
11	Fail to open	Fail to open	0.04	**0.71**	0.21	0.03	0.01
12	Hunting	Hunting	0.02	0.07	**0.9**	0.01	0
13	Hysteresis	Hysteresis	0.19	0.08	0.18	0.52	0.03
14	Fail to open	Fail to open	0.06	0.48	0.05	**0.41**	0
15	Fail to open	Fail to open	0.03	0.68	0	0.21	0.08
16	Other	Hunting	0.02	0.02	**0.69**	0.07	0.2

Table 5.10: Classification probabilities of the case study using RF with ANOVA selection.

The computation time of the failure diagnosis method has also been determined in order to benchmark the findings against the current state of the failure diagnosis. The method requires 0.0069 [seconds] to diagnose the failure modes of the 16 cases in the flow control valves. The computation time is negligible, and therefore will be treated as 0 [minutes].

5.3.1. Sample size of training set

In this section, the impact of the number of samples present in the training set on the accuracy of diagnosis on the test set. In order to test the influence, the number of samples will be reduced with ten resulting in 5 accuracies.

Number of samples	Accuracy on test set [%]
10	56
20	69
30	75
40	75
50	81

Table 5.11: The effect of the samples present in the training set on the classification accuracies on the test set.

The results shown in Table 5.11 show that the number of samples present in the training set positively impacts the accuracy achieved on the case study test set.

5.4. Validation of the results

This section validates the results achieved on the test set from the previous Chapter. First of all, the computation time of the proposed method will be benchmarked against the time the current diagnosis process requires. Finally, the accuracy performance achieved on the test set will be benchmarked with the identified minimum of the engineers using a survey and concerning the current state of the literature.

5.4.1. Required time

The current process of failure diagnosis is done by hand after the fault detection model raises an alarm. In order to determine the magnitude of the time reduced by the implementation of the automatic failure diagnosis, the case study should be performed by engineers and benchmarked against the computation time from Chapter 5.3.

Two engineers have successfully diagnosed the 16 failures from the case study, and the results are shown in Table 5.12.

Engineer	Time required for diagnosis [minutes]	Accuracy [%]
1	116	100
2	122	100
Average	119	100

Table 5.12: Results of failure diagnosis of case study by hand performed by two Shell engineers.

These findings show that on average, it takes the engineers roughly 119 minutes to determine the failure mode of the valve using the time series data. The corresponding accuracy, however, was 100% for both of the engineers. To conclude, in the case study, 119 minutes of diagnosis time can be reduced for 16 failures, which corresponds to approximately 7.5 minutes per failure.

5.4.2. Accuracy

The goal of the research is to develop an effective automatic failure diagnosis method where the required time reduced with sufficient accuracy. Therefore, the accuracy resulting from the test set is benchmarked to determine the quality of the proposed method. First of all, engineers that perform the diagnosis will be surveyed to determine the aspired accuracy level. Second of all, a method described in the literature is replicated and tested on the Shell Pernis data.

This section contains more information on the by Shell engineers identified minimum accuracy level. The accuracy of the method should be above the minimum to be sufficient for use in the refinery. Several engineers from the Shell Pernis refinery, who work with the control valves daily, have been interviewed. The survey consisted of three questions:

1. How useful would you rate the impact of implementing the automatic failure diagnosis method for flow control valves?

2. What level of accuracy would you require from an automatic failure diagnosis method with five fault categories that reduces 7.5 minutes per incident?

3. What level of accuracy would you require from an automatic failure diagnosis methods with three fault categories that reduce 7.5 minutes per incident? (Fail to close, Fail to open, and Other)

Question 1 is ranked with a score between 0 and 5. The engineers can give a percentage score between 0 and 100 for Questions 2 and 3. Three engineers have been interviewed. The average scores can be found in Table 5.13.

Question	Average score
1	4.7
2	73%
3	83%

Table 5.13: Survey results of automatic failure diagnosis impact and aspirations.

These findings show that the accuracy results for five failure categories when diagnosing new incident data and reducing 7.5 minutes, should be above 73%. The results of the case study show that the RF with ANOVA selection method can achieve 81% accuracy on unseen data. Furthermore, the accuracy of the three mentioned categories must be above 83% in order to be useful for automatic failure diagnosis. The findings of the case study present an accuracy of 88% when applying the three failure categories. Finally, the survey shows a large impact, 4.7 out of 5, that can be achieved by implementing the proposed method.

In this section the proposed method will be benchmarked against the current state of the literature on failure diagnosis for flow control valves. It is hard to compare these methods in literature due to various reasons. First of all, most of the research is done with simulated data. This research, however, is performed on actual industry data. Second of all, different failure types are found and distinguished in other applications of control valves, which makes it hard to do a proper benchmark. Finally, the proposed method can diagnose failures of flow control valves all over the plant while most of the research is done on a single valve. Implementing single failure diagnosis methods on all the valve in the Shell refinery in Pernis would require too much time. Nevertheless, due to the lack of available data, we replicated the paper on failure diagnosis using ANN on flow control valves with similar failure categories to benchmark our proposed method to the current state of the literature. Marciniak et al. (2003) describes the application of feed-forward neural network architecture on simulated data. However, due to the lack of code available from the research and the missing clarity of the description, the method is reproduced to the best of our knowledge. The method first detects failures and afterwards isolates three failure modes (F1, F2, and F3) and a normal state (F0). The first failure (F1) is described as a restricted movement of the rod, also known as a blockage. The second failure is related to the bypass valve, not closing properly. The third failure is related to a broken flow meter caused by wiring or electronic issues. After detection, the accuracy of the failure classification is determined using leave-one-out cross-validation. The method uses several measured time series to detect and isolate the failures: upstream and downstream pressures, fluid temperature, rod displacement, control reference, and the flow. This behaviour of the time series is described using feature extraction methods such as correlation functions and power spectral analysis. The classification method is replicated using a Multi-Layer Perceptron (MLP) NN, which is a class of feed-forward ANNs. The architecture is built with five hidden layers and optimised using stochastic gradient descent described by Kingma and Ba (2015). The evaluation method is like k-fold cross-validation, where k is equal to the total number of samples in the set. In order to benchmark the methods, the neural network architecture was replicated and applied on the data available from the Shell refinery in Pernis. Two failure categories can be compared. The first failure mode (F1), related to blockage, can be compared to the behaviour categories ‘Fail to close’ and ‘Fail to open’. Furthermore, due to the similar causes, we combine the ‘Hunting’, ‘Hysteresis’, and ‘Other’ modes into one category named ‘Other’. The confusion matrix results of the leave one out cross-validation using Random Forest with ANOVA are shown in Figure 5.17 and using the NN method is found in Figure 5.16.

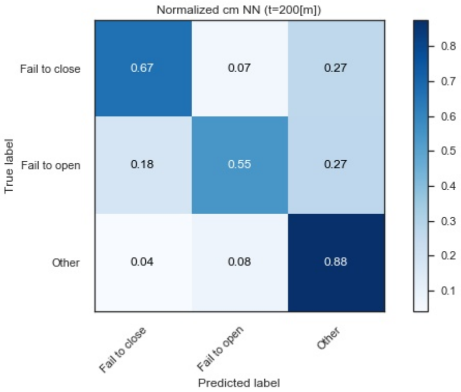


Figure 5.16: Benchmark results of NN method using leave-one-out cross validation on the training set.

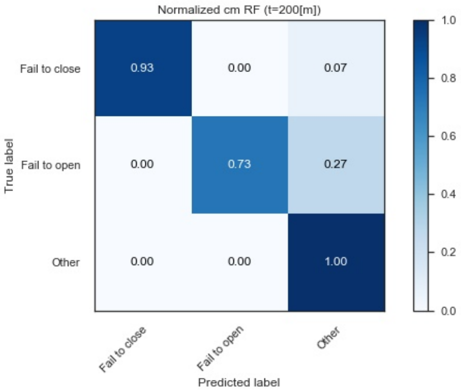


Figure 5.17: Benchmark results of RF method using leave-one-out cross validation on the training set.

These findings show that in all the failure modes, the proposed supervised ML classification model Random Forest scores better on leave one out cross-validation. Therefore, we can conclude that the proposed method obtains an increase in accuracy of 19% on average regarding the three failure categories.

5.5. Conclusion

The training set, represented by five plants within the refinery of Pernis, is used to determine what model, parameters, and length of time series input are optimal for automatic failure diagnosis in control valves. First of all, different features are adopted in the model, using the ANOVA and Mutual information method. Afterwards, the models and its parameters are compared using the two KPIs with cross-validation. Cross-validation on the training set prevents overfitting and allows us to draw proper conclusions on the model performance since the set consists of data from multiple plants. The optimal model for this problem should have high accuracy, without too much uncertainty in the model. The RF classifier, in combination with the ANOVA selection method, scores the best on both of the criteria and, therefore, is the method which is applied to the new incident data from the case study.

The RF classifier method obtains an accuracy of 0.81 on the new incident data of the case study, which is a less than when using the cross-validation, however, an excellent performance on unseen data from a plant that is not represented in the training set. The corresponding log loss value is 0.46, which means there is still quite some uncertainty in the prediction probabilities; however, it scores better than when using cross-validation on the training set. Furthermore, the model shows to never miss a classification on the categories ’Fail to close’ and ’Fail to open’, which is shown in the recall results. If the classification probabilities and mistakes are analysed more thoroughly, it is shown that the actual class obtains a probability of at least 0.20. The mean of the OP signal is the most important feature used in the model, meaning it makes the most critical decisions in the trees. The correlation between the signals PV and SP is the least important feature resulting from the classification model. The computation time of the method classifying 16 failures is less than a second. Finally, the findings show that increasing the sample size of the training set improves the accuracy results on the test set.

The results of the case study are validated to determine the quality of the method. Both the time required for diagnosis and the accuracy achieved are benchmarked with the current process and the current state of the literature. Regarding the required time, approximately 7.5 minutes can be reduced per failure using the proposed method. However, the accuracy drops from 100% to 81% on the test set. To determine if this accuracy is sufficient, the results are compared to the current state of the literature, and interviews have been done with Shell engineers to determine the minimum required accuracy. The accuracy performance results are better than both the current state of the literature (an increase of 19%) and the by Shell engineers determined minimum (73%).

5.6. Discussion

The results show that using the proposed novel method for automatic failure diagnosis in control valves, incident behaviour in control valves can be classified and predicted using ML classification on industry data. The optimal method, RF in combination with ANOVA selection, is tested on new incident data in the case study on the gasification plant. The performance results are significant for the application of large scale failure diagnosis in an industrial plant, and approximately 7.5 minutes can be reduced using the automatic failure diagnosis method. The accuracy achieved on unseen data exceeds the performance minimum set by the Shell engineers. Furthermore, accuracy is benchmarked against the current state of the literature. The accuracy of the proposed failure diagnosis method scores 19% better using leave one out cross-validation compared to a NN on the available data from the Shell Pernis refinery. However, there are some limitations to the proposed method.

First of all, the automatic diagnosis model uses selected input data based on information from the fault detection model. Therefore, it is highly reliant on a predictive failure detection method. The input dataset only contains actual failure behaviour data, so there is no standard behaviour data available. Without predictive failure detection, the automatic failure diagnosis can not be performed. However, this could be initiated in our framework by classifying normal behaviour as a category if successful the automatic failure diagnosis model can be used as incident detection monitoring tool. Further research should be performed in order to test whether the classification method would be suitable for this application.

Second of all, this research has been approached from an engineering perspective where the failure types were analysed beforehand. What if unsupervised learning methods such as clustering techniques with algorithms like SVMs are used to define the optimal decision points between samples? Data-driven approaches could result in different classes, which might also give exciting insights for engineers using the automatic diagnosis of failures in flow control valves. More research can be done here using the data-driven approach.

Third of all, the best performing method on the training set using cross-validation is the RF model with the ANOVA feature selection. However, more research has also been done to test the outcomes of the other classification models on the case study, shown in Appendix C.1. The additional research shows that the GNB method also obtains an accuracy 0.81 on the case study, which is similar to the RF classifier. However, on the training set cross-validation, the accuracy of the RF method scored significantly better than the GNB classifier. Does this imply that the population of the training set, retrieved from valves of five plants, is different from the SGHP plant used in the case study? This new result shows that given the current training set, the model choice can not be based on just the training set when incidents from other plants are analysed. Furthermore, more case studies should be performed on new plants, and more plants should be added to the training population in order to obtain a set on which we can ultimately make decisions based on the cross-validation results.

Finally, the goal of this research is to build a classification algorithm and selection method to automate the failure diagnosis process for engineers. Findings show a minimum prediction probability of 0.20 for the actual label. Therefore, when the diagnosis strategy is to prevent a certain failure for a control valve at all costs setting a probability threshold that increases the importance of the alarm when breached is extremely useful. An example, when the particular valve controls the fuel flow to a furnace, the valve should be able to close at all times, since too much fuel could overheat the whole mechanism. The model can give the alarm the highest priority when the threshold of 0.20 is breached. Furthermore, taking into account the accuracy of the first or second position could solve this problem. Additional research, shown in C.4 of the additional research, highlights the RF method as the most suitable for this technique.

6 Conclusion and recommendations for further research

Chapter 6 contains the conclusions and recommendations for further research. The conclusion chapter briefly answers the main and sub research questions based on the findings. In Chapter 6.2 the recommendations are stated containing several suggestions for future researchers working on failure diagnosis in general and in the ML classification field.

6.1. Conclusion

This chapter will give a conclusion to the research questions, as stated in Chapter 1, based on the results, described in Chapter 5, obtained by the classification and prediction of failures in control valves.

Sub research question 1: *What are the available failure types in Shell's data and can these be distinguished successfully in flow control valves using the available input of data?*

The data containing information on the behaviour of the control valves are the three signals of the controller: OP, PV, and SP. These time series, showing explicit failure behaviour, are used as input for the model since the failures can be distinguished with them. The signal input is updated every minute by the controller and stored in a cloud database. Using Shell's historical failure database seven critical failure types were identified. However, these failure types could show similar characteristics in the signal. Therefore, we have determined five failure behaviour categories. The result is the output of the automatic failure diagnosis which consists of five incident behaviour categories: Fail to open, Fail to close, Hunting, Hysteresis, and Other.

The training data, built from 50 failure cases extracted from Shell's maintenance notification database, covers five plants with different locations and applications. The case study includes data from the SGHP unit, where the gasification process takes place on refining residue materials. The dataset of this new unit is kept separately and has a different application from the plants involved in the training set.

Sub research question 2: *What features can be extracted from the model input to distinguish failures, and which of these are significant predictors for the output?*

The input data is compressed into single values with feature extraction, where the signals are summarized using statistical and time series analysis methods. The results are 13 features, created especially for flow control valves and the signals available. The features range from average values of the OP to high range frequencies of the power spectral density using Fourier transforms. In order to determine the significant predictors for the output, two statistical tests, meeting the requirements for the available data, have been used: ANOVA, and Mutual information. Hypothesis testing assesses the significance of features with a confidence interval of 95%. The null hypothesis states that the features are no significant predictors to the output. The features for which the null hypothesis was rejected are adopted in the model. The results of the case study also show the feature importance obtained from the fit of the training set. The mean OP feature was found to be the most important to the classification of failure behaviour.

Sub research question 3: *Which supervised machine learning algorithm performs best on the classification and prediction of cross-validation of the training set using performance metrics?*

After the selection of the features using the ANOVA and Mutual information methods, we apply three classification algorithms: Linear regression, Gaussian Naive Bayes, and Random Forest. Cross-validation has been applied to the training set, using the KPIs accuracy and log loss. This method is used to choose the best model and selection method for the diagnosis, tune the model parameters and determine the optimal length of the time series input. The RF classifier with the ANOVA feature selection method has been determined as the optimal classifier with a maximum tree depth of 10. The algorithm scores significantly higher on the accuracy with 0.88 ± 0.06 and performs reasonably on the average sample log loss with an outcome of 0.49. The length of the signal input, showing the failure behaviour, is chosen as $t = 200$. The shortest time window has been chosen due to the results of the KPI's, and to reduce the chance of normal operating behaviour, acting as noise, being present in the input signals.

Main research question: *What performance can be achieved in the classification and prediction of failure types in flow control valves using supervised machine learning classification, and what is the impact on the failure diagnosis process?*

The quality of the supervised ML classifier is measured using several performance metrics on the case study of the gasification plant, containing data not represented in the training set. The results are shown in Table 6.1.

	RF
Accuracy	0.81
Log loss	0.46
Precision (weighted)	0.86
Recall (weighted)	0.81

Table 6.1: Distribution of samples in the five failure behaviour categories.

Results of the case study show that we can automatically classify and predict failure behaviour on new incident data with an accuracy 0.81 using unsupervised ML classification on unseen data from a different plant. The log loss result shows there is quite some uncertainty in the model caused by the type of algorithm. However, this also delivers opportunities. The results section shows that the correct class always obtains a probability of at least 0.20 in the classification probabilities. When the diagnosis strategy is to prevent critical failure type at all times for a specific control valve, a classification probability threshold can be set. Furthermore, the model shows to never miss a classification on the categories 'Fail to close' and 'Fail to open', which is shown in the recall results. Therefore, the research presents the possibility to diagnose these two failures at all times correctly.

The issue that Shell faces is the lack of information in the output of the available fault detection model, resulting in an overload of unresolved alarms. Therefore, the goal of this research was to develop an effective automated failure diagnosis method for flow control valves in the refinery of Pernis. Effective in terms of reduced time required for diagnosis and sufficient accuracy. On unseen data, the automatic failure diagnosis method can reduce approximately 7.5 minutes per fault, and obtain an accuracy of 81%. This accuracy is sufficient when compared to the identified minimum by the Shell engineers (73%), and when benchmarked with the current state of the literature (19% increase). Therefore, the goal of this research has been achieved.

The implementation of the proposed method will have an impact on the current failure diagnosis process. First of all, due to the ability to classify and predict within several seconds, the diagnosis time will reduce considerably. Currently, step two of the failure diagnosis process takes about 7.5 minutes on average for an engineer. The whole process will shorten to less than a second using the automatic failure diagnosis. Second of all, engineers can use the model to classify alarms, which will help them in actively prioritizing and searching for issues that are critical and urgent.

This research has shown that the diagnosis time required for failures in control valves can be reduced with 7.5 minutes per unit on unseen data from a different plant processing various materials in an oil refinery. The automatic failure diagnosis is possible with an accuracy of 81%, shown in the performance results of the case study.

6.2. Recommendations for further research

This chapter describes recommendations that can be further explored, following from this research on automatic failure diagnosis in control valves. Chapter 5.6 has opened up the discussion on the results and their limitations. During this work, various possible improvements were discovered; these are stated in this Chapter. First, some recommendations on the classification model are given. Afterwards, some new directions are proposed to research the impact on the failure diagnosis process for Shell.

6.2.1. Classification model

First of all, the thirteen features created from the time series input showing the failure behaviour allow the model to obtain an accuracy of over 80% on new incident data. The results, however, do show that the model sometimes has problems in distinguishing the failure groups 'Fail to open', 'Hysteresis', and 'Hunting'. Therefore, future research should be done in adding more features to the model that show the individual characteristics of the classes and improve the ability of the classifier to predict the outcomes of these categories with various similarities correctly. The results obtained from the classification are not flawless. However, the advantage of the use of supervised ML classification in automatic failure diagnosis is the continuous improvement of the model after more training data is added to the model. Furthermore, new features and failure cases can be added without having to change the architecture of the automated failure diagnosis fundamentally. Moreover, due to the implemented selection algorithm, just the significant predictors will be adopted in the model. Also, the addition of more samples to the training set would improve the quality and reliability of the model, as shown in Chapter 5.3.1.

Second of all, the possibility of large scale classification of normal behaviour as an output category should be explored in order to apply the model for fault detection and diagnosis. There are two options of using this technique: classifying alarms when the model is showing normal behaviour as false alarms, and using the model to monitor normal behaviour and raise the alarm when the model shows characteristics other than the trained normal behaviour. The first option would reduce the number of false alarms, and thereby reduce the overload of unresolved alarms. The second proposal would require substantial changes to the model since it transforms the approach of the mechanism. When the model would be used to monitor the valve behaviour continually, a significant imbalance in the data input will be created, due to the large amount of normal behaviour in comparison to the failure characteristics. Therefore, the possibilities of suitable ML models should be explored first to determine the optimal model available.

Third of all, as mentioned in the discussion, this research has been approached from an engineering perspective, where the failure categories are divided into groups based on different failure types. Afterwards, the data is used to build the features and afterwards distinguish the classes. However, it is not known if these behaviours are now distributed over the perfect categories given the input data. Therefore, the opportunities for using unsupervised learning, such as clustering with an SVM, should be researched. SVM based clustering clusters data with no a priori knowledge of the input classes (Winters-Hilt and Merat, 2007). The results of the clustering techniques can be benchmarked versus the current classification models to look for the optimal division in the failure categories and look for the possibilities of separating more classes given the data.

6.2.2. Failure diagnosis process

This research has been performed on a brownfield site, with aged equipment and sensors. The implementation of new hardware for control valves, such as smart valve positioners, would provide many opportunities in the detection and diagnosis of control valves. The additions of smart positioner bring several advantages such as automatic calibration and configuration of the positioner, real-time diagnostics, and improved process control. When the actual position of the valve opening is known, many new potential failures can be predicted. Furthermore, smart control valves often have built-in diagnostic capabilities in order to continually be up to date on the current state of the control valve.

Bibliography

Adhikari, P., Rao, H. G., and Buderath, M. (2018). Machine Learning based Data Driven Diagnostics & Prognostics Framework for Aircraft Predictive Maintenance. *10th International Symposium on NDT in Aerospace, October 24-26, 2018, Dresden, Germany*, (Ml):1–15.

Aerd Statistics (2018). Spearman's Rank-Order Correlation using SPSS Statistics. `https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php`. [Online; accessed 28-January-2020].

Bacci Di Capaci, R., Scali, C., Pestonesi, D., and Bartaloni, E. (2013). *Advanced diagnosis of control loops: Experimentation on pilot plant and validation on industrial scale*, volume 46. IFAC.

Bala, R. and Kumar, D. (2017). Classification Using ANN: A Review. *International Journal of Computational Intelligence Research ISSN*, 13(7):973–1873.

Beirlant, J., Dudewicz, E. J., Györfi, L., and Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Bishop, C. (2006). *Pattern recognition and Machine learning*, volume 27. Springer, Singapore.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Calado, J. M., Sá da Costa, J. M., Bartys, M., and Korbicz, J. (2006). FDI approach to the DAMADICS benchmark problem based on qualitative reasoning coupled with fuzzy neural networks. *Control Engineering Practice*, 14(6 SPEC. ISS.):685–698.

Choudhury, M., Thornhill, N., and Shah, S. (2004). A Data-Driven Model for Valve Stiction. *IFAC Proceedings Volumes*, 37(1):245–250.

Cortes, C. and Vapnik, V. (1994). Support-Vector Neworks. *Machine Learning*, 20(3):273–297.

Elssied, N., Ibrahim, O., and Osman, A. (2014). A novel feature selection based on one-way ANOVA F-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3):625–638.

Emerson Automation Solutions (2017). *Control valve handbook*. Fisher, 5 edition.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Gary, J., Handwerk, G., and Kaiser, M. (1988). *Petroleum refining technology and economics*, volume 67. Fifth edition.

Geron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems.*

Huang, X., Liu, J., and Niu, Y. (2010). Fault Detection of Actuator with Digital Positioner Based on Trend Analysis Method. *Fault Detection*.

Huang, X. and Yu, F. (2008). A simple method for fault detection of industrial digitial positioners. *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, pages 6858–6862.

Institute, M. G. (2017). Artificial intelligence – the next frontier in IT security? *Network Security*, 2017(4):14–17.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics.

Jones, D. (2006). *Process equipment in petroleum refining*, pages 877–1070. Springer Netherlands, Dordrecht.

Kent State University (2017). SPSS tutorials: Pearson correlation. `https://libguides.library.kent.edu/SPSS`. [Online; accessed 01-December-2019].

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *InforProceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies-matica 31*, 31(1):249–268.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(6):16.

Kumar, M., Rath, N. K., Swain, A., and Rath, S. K. (2015). Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor. *Procedia Computer Science*, 54:301–310.

Ling, B., Zeifman, M., and Liu, M. (2007). A practical system for online diagnosis of control valve faults. *Proceedings of the IEEE Conference on Decision and Control*, pages 2572–2577.

Marciniak, A., Bocăială, C. D., Louro, R., Sa Da Costa, J., and Korbicz, J. (2003). Pattern recognition approach to fault diagnosis in the DAMADICS benchmark flow control valve. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 36(5):861–866.

Mathur, N., Asirvadam, V. S., Abd Aziz, A., and Ibrahim, R. (2019). Fault Tree Analysis for Control Valves in Process Plants by using R. *Proceedings - 2019 IEEE 15th International Colloquium on Signal Processing and its Applications, CSPA 2019*, (March):152–156.

Minckler, D. (1995). Bias, Correlation, and Causation. *Ophthalmology*, 102(4):531–532.

Mitchel, T. (1997). *Machine Learning*. McGraw-Hill Science.

Nwaoha, C., Onyewuenyi, O. A., Holloway, M. D., and Holloway, M. D. (2012). *Process Plant Equipment : Operation, Control, and Reliability*. John Wiley & Sons, Incorporated, Somerset, UNITED STATES.

Pareti, P. (2010). Mining unexpected behaviour from equipment measurements. *Department of Information Technology*.

Prabakaran, K., Mageshwari, U., Prakash, D., and Suguna, A. (2013). Fault Diagnosis in Process Control Valve Using Artificial Neural Network. *International Journal of Innovation and Applied Studies*, 3(1):138–144.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2).

Samdani, G. (1992). Predictive Maintenance. *Chemical Engineering*, 41(4):38–43.

Samuel, A. L. (1969). Some studies in machine learning using the game of checkers. II-Recent progress. *Annual Review in Automatic Programming*, 6(PART 1):1–36.

Santi, G. B., Assis, A. A., Rebli, V. N., Ciarelli, P. M., Rauber, T. W., and Munaro, C. J. (2019). Feature extraction in an ensemble of multiple local classifiers for fault diagnosis in industrial processes. *IFAC-PapersOnLine*, 52(1):293–298.

Scali, C., Matteucci, E., Pestonesi, D., Zizzo, A., and Bartaloni, E. (2011). *Experimental characterization and diagnosis of different problems in control valves*, volume 44. IFAC.

Schaffer, C. (1993). Technical Note: Selecting a Classification Method by Cross-Validation. *Machine Learning*, 13(1):135–143.

SciPy.org (2019). Welch's method from the SciPy signal processing package. `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.signal.welch.html#r145`. [Online; accessed 19-November-2019].

Seo, M. and Jun, H.-B. (2019). A Study on the Diagnostics Method for Plant Equipment Failure. In Ameri, F., Stecke, K. E., von Cieminski, G., and Kiritsis, D., editors, *Advances in Production Management Systems. Production Management for the Factory of the Future*, pages 701–707, Cham. Springer International Publishing.

Statistics Solutions (2019). Kendall's Tau and Spearman's Rank correlation coefficient. `https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient/`. [Online; accessed 28-January-2020].

Sulzmann, J. N., Fürnkranz, J., and Hüllermeier, E. (2007). On pairwise naive bayes classifiers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4701 LNAI:371–381.

Swanson, L. (2001). Linking maintenance strategies to performance. *International journal of production economics*, 70:237–244.

Syfert, M., Patton, R., Bartyś, M., and Quevedo, J. (2003). Development and application of methods for actuator diagnosis in industrial control systems (DAMADICS): A benchmark study. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 36(5):843–854.

Trunzer, E., Weis, I., Folmer, J., Schrufer, C., Vogel-Heuser, B., Erben, S., Unland, S., and Vermum, C. (2018). Failure mode classification for control valves for supporting data-driven fault detection. *IEEE International Conference on Industrial Engineering and Engineering Management*, 2017-Decem:2346–2350.

Van Hulse, J., Khoshgoftaar, T. M., Napolitano, A., and Wald, R. (2009). Feature selection with high-dimensional imbalanced data. *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, pages 507–514.

Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186.

von Storch, H. and Zwiers, F. W. (2010). The F Distribution . *Statistical Analysis in Climate Research*, pages 424–430.

Wang, H., Chai, T. Y., Ding, J. L., and Brown, M. (2009). Data driven fault diagnosis and fault tolerant control: Some advances and possible new directions. *Zidonghua Xuebao/Acta Automatica Sinica*, 35(6):739–747.

Welch, P. D. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Trans. Audio and electroacoustic*, 15(AU-15):70–73.

Winters-Hilt, S. and Merat, S. (2007). SVM clustering. *BMC Bioinformatics*, 8(SUPPL. 7):1–12.

Xie, C., Yang, D., Huang, Y., and Sun, D. (2015). Feature extraction and ensemble decision tree classifier in plant failure detection. *Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM*, pages 727–735.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224.

Zhang, H. (2004). The optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, 2:562–567.

Zhang, L., Xia, C., Cao, J., and Zheng, J. (2012). Physical-based modeling of nonlinearities in process control valves. *Proceedings - 2012 International Conference on Control Engineering and Communication Technology, ICCECT 2012*, pages 75–78.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., 1st edition.

Zuideveld, P. and Graaf, J. D. (2005). Overview of Shell Global Solutions ' Worldwide Gasification Developments. pages 1–7.

A
Scientific paper

Automatic Failure Diagnosis for Flow Control Valves

E. Ruijs[1], Dr. X. Jiang[1], Dr. T. Park[2], Prof. R. Negenborn[1]

[1] *Delft University of Technology, 2628CD, Department of Transport Engineering & Logistics, Delft, Netherlands*
[2] *Shell Global Solutions International BV., 1031HW, Amsterdam, Netherlands*

April 8, 2020

The increased implementation of digitalisation all over the world has led to an exponential growth of available data across various industries. Consequently, there is a large growth of Machine Learning (ML) techniques being applied to process data. Predictive maintenance is a digital strategy using condition-based monitoring techniques to track the performance of equipment to detect possible defects in advance. In oil refineries, various types of process equipment are used, of which flow control valves are essential to regulate the throughput of heavy, possibly dangerous material. Control valve failures can lead to production loss and increase maintenance costs. This paper addresses the use of time-series data of the valve controller for automated failure diagnosis of flow control valves. Statistical features are extracted from the time series and the significant predictors for the output are adopted in the model using the ANOVA test. The classification and prediction of the failure behaviour are performed using Random Forest (RF) classification. The performance of the diagnosis is measured using the indicators: accuracy and log loss. Findings show that five failure behaviour categories can be predicted, using new and unseen data for the model, with high accuracy(81%).

Keywords: Control valves, Predictive maintenance, Failure diagnosis, Machine learning, AI

1 Introduction

Control valves are essential fixed process equipment in oil refineries, where a network of closed control loops contributes to the generation of finished products. Component failure in flow control valves should be detected early in order to avoid unexpected breakdown, causing downtime. Therefore, the application of predictive maintenance strategies on flow control valves is researched.

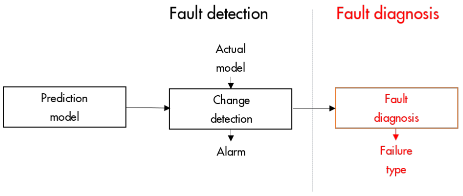


Figure 1: *Failure detection and diagnosis in flow control valves. Black: fault detection model raising an alarm whenever a change is detected between the predicted and actual model. Red: fault diagnosis model with a failure type as output when an alarm is raised by the fault detection tool.*

1.1 Research goal

Currently, Shell's refinery uses ML to detect failures in control valves up to 75 days in advance. The problem, however, is that diagnosing failures by hand requires too much time, causing an overload of unresolved alarms. This research aims to develop an effective automated failure diagnosis method for flow control valves in the refinery of Pernis when a failure is known to be present. Effective in terms of reduced time required for diagnosis and sufficient accuracy Figure 1 shows the architecture of fault detection and diagnosis for flow control valves.

1.2 Literature review

Automatic failure diagnosis can be divided into two approaches: model-based and data-driven methods. Where model-based methods focus on mathematical models to estimate the state and parameters of the system and data-driven methods concentrate on high dimensional data and are applied to highlight important information from the available data (Wang et al., 2009). Several papers describe the use of trend analysis to determine failure states in the control valve. Bacci Di Capaci et al., 2013 and Scali et al., 2011 create KPIs from the available signals and set thresholds for these KPIs to diagnose the failure. Furthermore, Huang and Yu, 2008 proposes a trend analysis method to detect actuator failures, and Trunzer et al., 2018 describes a classification table using expert knowledge. Trend analysis methods, however, do not easily generalise for valves with different applications and therefore behaviour and material throughput. Prabakaran et al., 2013 and Marciniak et al., 2003 research the use of Artificial Neural Networks (ANN) for the diagnosis of process control valves on simulated data. The downside of the use of ANN is the low interpretability of the models and the high risk of overfitting. Mathur et al., 2019 describes a diagnosis method using Fault Tree Analysis (FTA). However, this model determines the likelihood of a failure event, not the type based on incident data. Furthermore, fault diagnosis methods are used on other process equipment. Seo and Jun, 2019; Santi et al., 2019 propose the use of clustering methods to extract abnormal patterns in data. Adhikari, Rao, and Buderath, 2018 provides a framework for predictive maintenance using ML classification for aircraft. Praveenkumar et al., 2014 shows the use of Support Vector Machine (SVM) to identify faults of an automobile gearbox. This research proposes a novel framework for the classification and prediction of failures in flow control valves using supervised ML classification on actual plant data.

1.3 Paper outline

This paper will have the following outline. First of all, the data input available from the flow control valves and the desired output will be discussed. Second of all, the methodology of the feature extraction, selection, and supervised ML classification will be explained. Third of all, the results of the case study performed on the Shell Gasification Hydrogen Plant (SGHP) will be reviewed. Finally, the results and performance of the model will be discussed and compared.

2 Data

In flow control valves, the Set-Point (SP), Process Variable (PV), and Controller output (OP) are usually recorded from the feedback loop (Bacci di Capaci and Scali, 2018). The control valve aims to minimise the difference between the PV [tonnes per day] and the SP [tonnes per day] by adjusting the OP [%]. These three signals show the failure behaviour in the flow control valve and therefore, are used as time series input for the model.

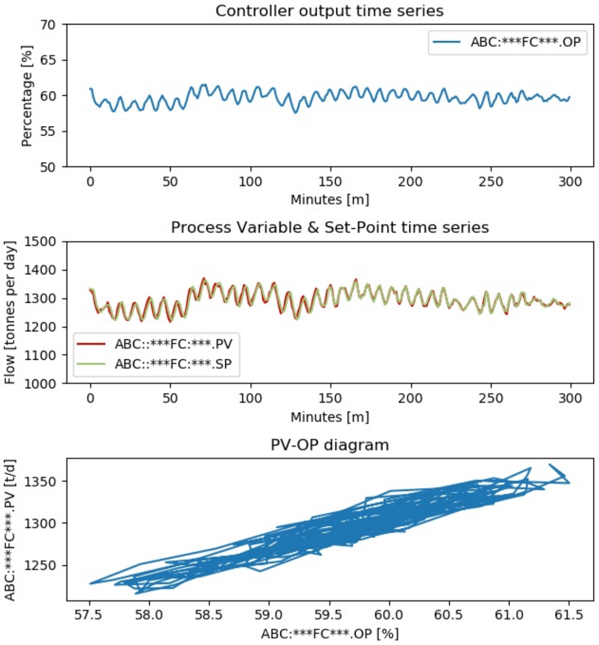


Figure 2: *Example of the model input time series data. Upper: OP time series. Middle: PV and SP time series. Lower: plot of the PV and OP.*

Several assumptions are made on the input data. First of all, the valve type does not influence the failure behaviour. Second of all, the amount and type of material thus the location and application of the valve do not influence the failure behaviour. Finally, the control mode, automatic or cascade, does not influence the failure behaviour.

The output of the model should contain the failure diagnosis of the flow control valves. The type and number of failure categories used in literature is very diverse. Trunzer et al., 2018 and Syfert et al., 2003 divide the failures over different components of the control valve. Bacci Di Capaci et al., 2013 distinguishes six failure groups: nominal, disturbance, jamming, stiction, leakage, and i/p converter malfunction. Ling, Zeifman, and Liu, 2007 describes four failure categories: leakage, blocking, deadband, and backlash. Mathur et al., 2019 isolates three categories, where the valve fails to open fully, open partially, or close fully. Using information from refinery engineers, that currently do the diagnosis by hand, and data from historical failure cases seven major failure types were classified for flow control valves: sticking, blockage, leakage, tuning issue, i/p defect, broken flow meter, external issue. However, some of these failure are hard to distinguish, due to the fact that various types can show similar failure behaviour. Based on the time series data, five failure categories are identified: Fail to close, Fail to open, Hunting, Hysteresis, and Other. The 'Fail to close' category contains failure cases where the valve has trouble closing, which results in the valve steering the OP to close but failing and therefore causing a large difference between the SP and the PV. The characteristics of the 'Fail to open' category are exactly the opposite. The 'Hunting' behaviour shows a constant overshoot of the positioner, resulting in an anti-phase between the signals OP and PV. 'Hysteresis' is the difference between the valve position on the upstroke and its position on the downstroke at any given input signal, resulting in different values of the PV for the same value of the OP. The 'Other' category show failures where no clear failure behaviour can be identified in the input data. The training set for the supervised ML classifier is build from the most critical historical cases retrieved from five different plants from Shell's refinery in Pernis using the failure notification database. The dates of the problems in the flow control valves range from 2014 to the start of 2020. Four steps were taken in order to retrieve the data input and output:

1. Manual extraction of time series data;
2. Data exploration and analysis to determine behaviour;
3. Verification of behaviour with Shell's instrumentation engineer;
4. Labelling cases with correct output class.

The result is 50 cases from two crude distillers, one hydrocracker unit, one hydrodesulfurization unit, and one platformer unit.

The case study, used as a test set for the supervised ML classification, is performed on the SGHP, where low-value heavy residues are converted into clean fuel for combined cycle power generation. The heavy materials flowing through the control valves result in 18 failure cases divided over the five behaviour categories.

3 Methodology

Given the available data, the methodology for the extraction and selection of features should be determined. Furthermore, the supervised ML classification algorithms have to be chosen.

3.1 Feature extraction

The extraction of features from time series refers to the transformation of data into formats that are suitable for a ML model. The set of features are built in order to distinguish the characteristics of the various failure behaviour categories, using statistical methods on the signals, such as the mean, variance, correlations, and ratio's. Furthermore, the Power Spectral Densities (PSD) of the signals are applied to create features. In total 13 features have been created.

3.2 Feature selection

Feature selection refers to removing unnecessary features before running the model to improve its quality in terms of run-time and performance. Several feature selection methods, with different requirements of the data, have been considered. Given the continuous input data from the created features and the categorical output, two statistical tests have been chosen: ANOVA, and Mutual information.

The ANOVA (F-test) measures the degree of linear dependency between two random variables and can be used to test whether a feature is a significant predictor to the output. The test calculates an F-value by comparing the variability between and within the groups. Large F-value, emerging from a large distance between the means of the variables, corresponds to a good predictor for the output. The mathematical formulation is gathered from work by Elssied, Ibrahim, and Osman, 2014.

$$s_j^2 = \frac{\sum_{i=1}^{N_j}(x_{ij} - \bar{x})^2}{N_j - 1} \tag{1}$$

$$\bar{g} = \frac{\sum_{j=1}^{J} N_j \bar{x_j}}{N} \tag{2}$$

Finally, the F-value between the predictor and output can be calculated:

$$F = \frac{\dfrac{\sum_{j=1}^{J} N_j(\bar{x_j} - g)^2}{(J-1)}}{\dfrac{\sum_{j=1}^{J}(N_j - 1)s_j^2}{(N-1)}} \tag{3}$$

where N_j = the number of cases with Y = j, $\bar{x_j}$ = the sample mean of predictor X for target class Y = j, s_j^2 = the sample variance of predictor X for target class Y=j, and g = the grand mean of predictor X.

The Mutual information (MI) test, based on Shannon's entropy, is used to determine the non-linear mutual dependence between two random variables by quantifying the 'amount of information' obtained about one random variable as a result of observing the other random variable. Mutual information is calculated using Equation (4) and obtains a value between 0 and 1, where completely dependent.

$$I(x, y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{4}$$

The estimation of the MI value is performed according to work by Ross, 2014.

The significance of the feature selection methods is proven based on hypothesis testing, where the null hypothesis states that feature x_i is not a significant predictor for output y. The null hypothesis of the ANOVA is rejected when the F-value is higher than a critical value determined by using the F-distribution tables with a confidence interval and degrees of freedom. When the MI value of a feature is higher than mutual dependence of the 95^{th} quartile of the output and a thousand normally distributed variables ($\mu : 0, \sigma^2 : 1$) the null hypothesis is rejected for MI. Thus the features will be adopted in the model.

3.3 Supervised ML classification

Based on criteria, related to the interpretability, risk of overfitting, accuracy, ability to detect non-linear relationships, and implementation three supervised ML classification algorithms have been implemented: Logistic regression (LR), Gaussian Naive Bayes (GNB), and Random Forest (RF).

LR is a widely used ML classifier that determines the probability that a specific value of the predictor belongs to a particular class or category using a logistic function.

$$\mathrm{P}(c|X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1...+\beta_n X_n}} \tag{5}$$

Where P$(c|X)$ is the probability that a certain class occurs given predictor $X = x_0, ..., x_n$. β_0 is the intercept coefficient. $\beta_1, ..., \beta_n$ are the regression coefficients. After some manipulation the logistic function can be written as the logg-odds logit, shown in Equation (6).

$$\log\left(\frac{\mathrm{P}(c|X)}{1-\mathrm{P}(c|X)}\right) = \beta_0 + \beta_1 X_1... + \beta_n X_n \tag{6}$$

The coefficients required in logistic regression are estimated using the maximum likelihood function (Geron, 2017). Due to the five failure categories multinomial regression is applied, using the softmax regression (Bishop, 2006). In the LR model the regularization parameter C can be alternated, which controls the trade-off between the complexity of the model. Large values of C answer to highly complex model that tend to over fit.

Naive Bayes is a probabilistic classifier, based on Bayes theorem. The algorithm calculates the posterior probability, which is the chance of a specific class given the observation, using the likelihood, class prior probability, and the predictor prior probability. The Gaussian version of the Naive Bayes, assumes the predictors are sampled from a Gaussian distribution, and therefore the likelihood of an observation given the class can be calculated using Equation (7) (Zhang, 2004).

$$\mathrm{P}(x_i|c_i) = \frac{1}{\sqrt{2\pi\sigma^2_{c_i}}} e^{\left(-\frac{(x_i-\mu_{c_i})^2}{N}\right)} \tag{7}$$

Where μ_c and σ_c are estimated using the maximum likelihood. The Gaussian Naive Bayes classifier assumes that all the features are independent of each other, therefore, Equation (8) is valid.

$$\mathrm{P}(x|c) = \prod_{i=1}^{n} \mathrm{P}(x_i|c) \tag{8}$$

When the likelihood, prior probability of the class, predictor prior probability, and the likelihood is known, the posterior probability of every class c_i given the observations $(x_i, ..., x_n)$ can be determined using Equation 9.

$$\mathrm{P}(c_i|x) = \frac{\mathrm{P}(x|c_i) * \mathrm{P}(c_i)}{\mathrm{P}(x)} = \frac{\mathrm{P}(x|c_i) * \mathrm{P}(c_i)}{\sum_i^n \mathrm{P}(x|c_i) * \mathrm{P}(c_i)} \tag{9}$$

A RF classifier consists of a collection of tree-structured classifiers h(x, Θ_k), k = 1,.. where Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman, 2001). In the RF algorithm two random selections are performed. First of all, random bootstrap samples are drawn from the training set, also referred to as bagging. Second of all, random subsets of features are drawn for every split made in the decision trees. In the RF model, the maximum decision tree depth can be varied to prevent over fitting due to large depths in trees. Decision trees, based on the Classification and Regression Trees (CART) algorithm, distinguish classes based in Information Gain (IG). The split of a decision tree attempts to maximize the information gained by the split determined using Equation (10).

$$\mathrm{IG}(D_p, f) = I(D_p) - \frac{N_{\text{left}}}{N} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N} I(D_{\text{right}}) \tag{10}$$

Where f are the features to be split on. D_p is the dataset of the parent node, D_{left} is the dataset of the left child node, and D_{right} is the dataset of the right child node. I is the impurity criterion, such as the Gini index or entropy. N is the total number of samples, N_{left} is the number of samples in the left child node, and N_{right} is the number of samples in the right child node. The Gini index is calculated using Equation (11).

$$I_G = 1 - \sum_{i=1}^{n} p_i^2 \tag{11}$$

Where p is the proportion of samples that belongs to class c for a particular node. When a node is pure, the impurity index is 0.

3.4 Model evaluation

The model performance is assessed using two metrics: accuracy and log-loss. Accuracy divides the number of correct predictions through the total number of predictions made. Log-loss is a measure of goodness of probability estimates, similar to the cross-entropy measuring the difference between two probability distributions for a given observation of a random variable. Equation 12 shows the mathematical formulation of the log-loss function Ferri, Hernández-Orallo, and Modroiu, 2009.

$$\text{Log-loss} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log(p_{ij}) \tag{12}$$

Where N is the number of observations, c is the total number of classes, y_{ij} is a binary value 0,1 indicating whether observation i belongs to class j, and p_{ij} is the probability that observation i belongs to class j.

4 Results

The optimal classification model is chosen based on k-fold cross-validation on the training set. Cross-validation is a resampling method used to assess the performance, as stated in section 3.4, using ML models which have limited data available before using them on new data (Schaffer, 1993). Furthermore, the ML algorithm parameters are tuned and the input length of the time series is varied using $t = 200$, $t = 1000$, and $t = 5000$ [minutes].

4.1 Training set

The extraction of the training set features results in matrix x_{ij} for $(i = 1, .., 50)$ and $(j = 1, .., 13)$ representing the failure behaviour of each sample. The significant features are determined using the two selection methods resulting in different subsets of predictors.

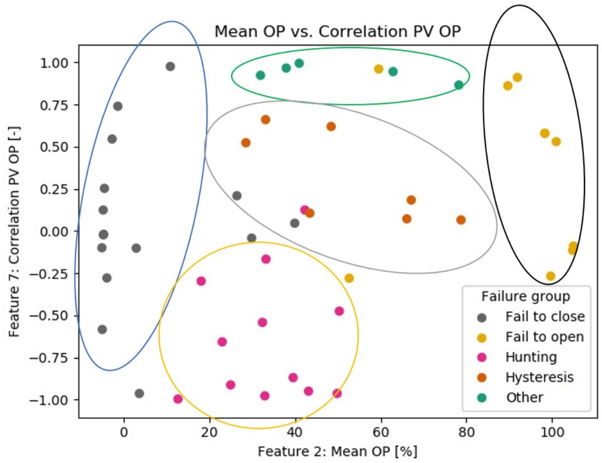


Figure 3: *Scatter plots of two features: Mean OP and Correlation PV OP. Blue oval shows the majority of 'Fail to close', green oval shows the majority of 'Other', yellow oval shows the majority of 'Hunting', grey oval shows the majority of 'Hysteresis', and black oval shows the majority of 'Fail to open' behaviour.*

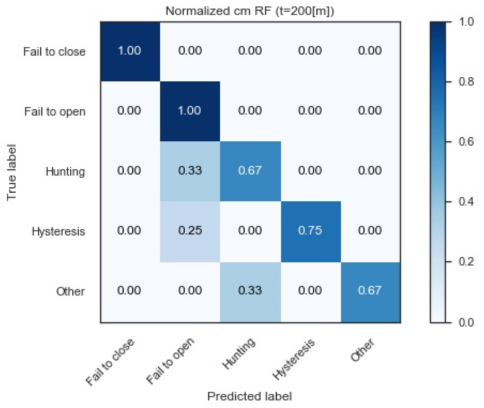
The cross-validation is applied to the training set, using the different features selected and corresponding time windows. The cross-validation accuracy metric is the mean of the cross-validation accuracies ± standard error of the mean of the cross-validation accuracies. The optimal model and settings are determined using cross-validation on the training set to prevent overfitting, and due to the large variety of plants present in the samples.

The cross-validation results of the RF model, using a maximum tree depth of 10, show a significantly higher accuracy on the training set compared to the other models when using the ANOVA selection method with a confidence interval of 95% and $t = 200[m]$. The accuracy achieved is 0.88 ± 0.06, and a log-loss value of 0.49. Next to the cross-validation results, the use of the short time-window decreases the chance of normal behaviour, possibly acting as noise, being present in the time series. The cross-validation accuracy results are shown in Table 1.

4.2 Case study: SGHP

The case study, performed on the SGHP of Shell's refinery, consists of unseen plant data not represented in the training set. The input data consists of 6 features, tested as significant with a confidence interval of 95% using ANOVA, in matrix x_{ij} for $(i = 1, .., 18)$ and $(j = 1, .., 6)$. The features adopted are: mean OP, (PV-SP)/SP, correlation SP PV, correlation PV OP, correlation SP OP, and the absolute difference between the PV and OP. When features are applied more in important decisions, corresponding to more information gained by splits higher up the decision trees, the importance grows. After the model is fit, the most important feature, measured in percentage of the total features, is the mean OP with 30%. The least important feature is the absolute difference between OP and PV.

When analysing the misclassification of the model, the actual category obtains a prediction probability of at least 0.20, which implies that the model was never completely off. The normalised confusion matrix of the results, shown in Figure 4, gives a clear overview of the predictions made by the classifier. The categories 'Fail to close' and 'Fail to open' are never misclassified and therefore obtain a recall value of 1. Most of the prediction errors are around the three categories, 'Hunting', 'Hysteresis', and 'Other'. This is due to these feature values being on the borders of the decision boundaries.


Normalized cm RF (t=200[m])

True label \ Predicted label	Fail to close	Fail to open	Hunting	Hysteresis	Other
Fail to close	1.00	0.00	0.00	0.00	0.00
Fail to open	0.00	1.00	0.00	0.00	0.00
Hunting	0.00	0.33	0.67	0.00	0.00
Hysteresis	0.00	0.25	0.00	0.75	0.00
Other	0.00	0.00	0.33	0.00	0.67

Figure 4: *Confusion matrix of the case study using the RF classifier with six features selected by the ANOVA test with $t = 200[m]$.*

	ANOVA			MI		
	LR	GNB	RF	LR	GNB	RF
t = 200	0.74 ± 0.03	0.76 ± 0.03	0.88 ± 0.06	0.74 ± 0.03	0.62 ± 0.01	0.72 ± 0.01
t = 1000	0.76 ± 0.07	0.74 ± 0.05	0.82 ± 0.06	0.68 ± 0.09	0.58 ± 0.06	0.74 ± 0.03
t = 5000	0.74 ± 0.06	0.79 ± 0.02	0.76 ± 0.03	0.78 ± 0.04	0.68 ± 0.07	0.81 ± 0.03

Table 1: *Accuracy results of the cross validation on the training set for the ANOVA and MI selection.*

	Accuracy	Log-loss
RF ($t = 200[m]$)	0.81	0.46

Table 2: *Case study results of the RF classifier on accuracy and log-loss using 6 features selected using ANOVA with* $t = 200[m]$.

5 Validation of the results

The computation time and accuracy of the case study using the method for automatic failure classification are benchmarked to the current process of diagnosis by hand, and to the current state of the literature. The required computation time is reduced with 7.5 minutes per control valve failure, however, the accuracy decreases from 100% to 81%. The replication of the NN method, proposed by Marciniak et al., 2003 on the data available from the Shell Pernis refinery, shows 19% lower accuracy results on three failure modes compared to the RF method using leave-one-out cross-validation on the training set. Finally, the obtained accuracy is sufficient when compared to the identified minimum by the Shell engineers (73%).

5.1 Discussion

The results show, using the proposed novel framework for automatic failure diagnosis in control valves, incident behaviour in control valves can be classified and predicted reducing 7.5 minutes per fault with an accuracy of 81% using supervised ML classification on actual data. However, there are some limitations to the method. The automatic failure diagnosis is completely reliant on the predictive fault detection method since failure behaviour should be present in order to classify the failure. However, the possibility of classifying normal behaviour to also detect failures should be explored. Moreover, the division of failure categories has been done from an engineering perspective. However, the categories show overlap and therefore are sometimes hard to distinguish. The results when using a data-driven approach, with an unsupervised clustering method such as SVM, could lead to new decision boundaries and distributions over the classes. The accuracy of the classifier was somewhat lower in the case study in comparison with the cross-validation results of the training set. Therefore, it can be assumed that the SGHP data set is not completely represented by the five plants present in the training set. Thus, when new plants are analysed multiple supervised ML models should be considered.

6 Conclusion

Notifications from 50 historical cases have been analysed, and five failure behaviour categories identified. The best performance results of the cross-validation are obtained by the RF classification model using six significant predictors selected using the ANOVA method. The results on the case study present accuracy of 81% on the case study with new unseen data from a plant not represented in the training set. Another finding shows the prediction probability of the actual label is higher than 20% at all times, which opens up opportunities for diagnosis strategies where critical failure types should be prevented at all times for a specific valve. Also, due to the recall values of 1 for the classes 'Fail to close' and 'Fail to open', such a prevention strategy can be realised. To conclude, on unseen industry data the automatic failure diagnosis method can reduce approximately 7.5 minutes per fault, and obtain an accuracy of 81%. This accuracy is sufficient when compared to the identified minimum by the Shell engineers (73%), and when benchmarked with the current state of the literature (19% increase). Therefore, the goal of this research has been achieved.

More research should be performed on the creation of new features, in order to test the possibilities of making a better division between the three failure categories 'Hunting', 'Hysteresis' and 'Other'. The advantage of the framework is that features can be created without having to change the architecture of the automated failure diagnosis fundamentally. Furthermore, due to the implemented selection algorithm, the significant predictors will be adopted in the model. Finally, more samples can be added to the training set to increase the quality of diagnosis.

Bibliography

Adhikari, Partha, Harsha Gururaja Rao, and Dipl.-Ing Matthias Buderath (2018). “Machine Learning based Data Driven Diagnostics & Prognostics Framework for Aircraft Predictive Maintenance”. In: *10th International Symposium on NDT in Aerospace, Octo-*

ber 24-26, 2018, Dresden, Germany Ml, pp. 1–15. URL: https://www.ndt.net/article/aero2018/papers/We.5.B.3.pdf.

Bacci Di Capaci, R. et al. (2013). *Advanced diagnosis of control loops: Experimentation on pilot plant and validation on industrial scale*. Vol. 46. 32 PART 1. IFAC, pp. 589–594. ISBN: 9783902823595. DOI: 10.3182/20131218-3-IN-2045.00107. URL: http://dx.doi.org/10.3182/20131218-3-IN-2045.00107.

Bacci di Capaci, Riccardo and Claudio Scali (2018). "Review and comparison of techniques of analysis of valve stiction: From modeling to smart diagnosis". In: *Chemical Engineering Research and Design* 130, pp. 230–265. ISSN: 02638762. DOI: 10.1016/j.cherd.2017.12.038. URL: http://dx.doi.org/10.1016/j.cherd.2017.12.038.

Bishop, Christopher (2006). *Pattern recognition and Machine learning*. Vol. 27. Singapore: Springer, p. 738. ISBN: 9780387310732.

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45, pp. 5–32. DOI: 10.1007/978-3-662-56776-0_10.

Elssied, Nadir Omer Fadl, Othman Ibrahim, and Ahmed Hamza Osman (2014). "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification". In: *Research Journal of Applied Sciences, Engineering and Technology* 7.3, pp. 625–638. ISSN: 20407459. DOI: 10.19026/rjaset.7.299.

Ferri, C., J. Hernández-Orallo, and R. Modroiu (2009). "An experimental comparison of performance measures for classification". In: *Pattern Recognition Letters* 30.1, pp. 27–38. ISSN: 01678655. DOI: 10.1016/j.patrec.2008.08.010.

Geron, Aurelien (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. ISBN: 9781491962299.

Huang, Xiaobin and Feng Yu (2008). "A simple method for fault detection of industrial digitial positioners". In: *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, pp. 6858–6862. DOI: 10.1109/WCICA.2008.4593976.

Ling, Bo, Michael Zeifman, and Ming Liu (2007). "A practical system for online diagnosis of control valve faults". In: *Proceedings of the IEEE Conference on Decision and Control*, pp. 2572–2577. ISSN: 01912216. DOI: 10.1109/CDC.2007.4434224.

Marciniak, A. et al. (2003). "Pattern recognition approach to fault diagnosis in the DAMADICS benchmark flow control valve". In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 36.5, pp. 861–866. ISSN: 14746670. DOI: 10.1016/S1474-6670(17)36600-4.

Mathur, Nirbhay et al. (2019). "Fault Tree Analysis for Control Valves in Process Plants by using R". In: *Proceedings - 2019 IEEE 15th International Colloquium on Signal Processing and its Applications, CSPA 2019* March, pp. 152–156. DOI: 10.1109/CSPA.2019.8696008.

Prabakaran, K et al. (2013). "Fault Diagnosis in Process Control Valve Using Artificial Neural Network". In: *International Journal of Innovation and Applied Studies* 3.1, pp. 138–144. ISSN: 2028-9324.

Praveenkumar, T. et al. (2014). "Fault diagnosis of automobile gearbox based on machine learning techniques". In: *Procedia Engineering* 97, pp. 2092–2098. ISSN: 18777058. DOI: 10.1016/j.proeng.2014.12.452. URL: http://dx.doi.org/10.1016/j.proeng.2014.12.452.

Ross, Brian C. (2014). "Mutual information between discrete and continuous data sets". In: *PLoS ONE* 9.2, p. 5. ISSN: 19326203. DOI: 10.1371/journal.pone.0087357.

Santi, Gustavo B. et al. (2019). "Feature extraction in an ensemble of multiple local classifiers for fault diagnosis in industrial processes". In: *IFAC-PapersOnLine* 52.1, pp. 293–298. ISSN: 24058963. DOI: 10.1016/j.ifacol.2019.06.077. URL: https://doi.org/10.1016/j.ifacol.2019.06.077.

Scali, C. et al. (2011). *Experimental characterization and diagnosis of different problems in control valves*. Vol. 44. 1 PART 1. IFAC, pp. 7334–7339. ISBN: 9783902661937. DOI: 10.3182/20110828-6-IT-1002.02755. URL: http://dx.doi.org/10.3182/20110828-6-IT-1002.02755.

Schaffer, Cullen (1993). "Technical Note: Selecting a Classification Method by Cross-Validation". In: *Machine Learning* 13.1, pp. 135–143. ISSN: 15730565. DOI: 10.1023/A:1022639714137.

Seo, Minyoung and Hong Bae Jun (2019). "A Study on the Diagnostics Method for Plant Equipment Failure". In: *IFIP Advances in Information and Communication Technology* 566. Ed. by Farhad Ameri et al., pp. 701–707. ISSN: 1868422X. DOI: 10.1007/978-3-030-30000-5_85.

Syfert, Michal et al. (2003). "Development and application of methods for actuator diagnosis in industrial control systems (DAMADICS): A benchmark study". In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 36.5, pp. 843–854. ISSN: 14746670. DOI: 10.1016/S1474-6670(17)36598-9.

Trunzer, Emanuel et al. (2018). "Failure mode classification for control valves for supporting data-driven fault detection". In: *IEEE International Conference on Industrial Engineering and Engineering Management* 2017-Decem, pp. 2346–2350. ISSN: 2157362X. DOI: 10.1109/IEEM.2017.8290311.

Wang, Hong et al. (2009). "Data driven fault diagnosis and fault tolerant control: Some advances and possible new directions". In: *Zidonghua Xuebao/ Acta Automatica Sinica* 35.6, pp. 739–747. ISSN: 02544156. DOI: 10.3724/SP.J.1004.2009.00739.

Zhang, Harry (2004). "The optimality of Naive Bayes". In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004* 2, pp. 562–567.

B

Results

B.1. Method selection

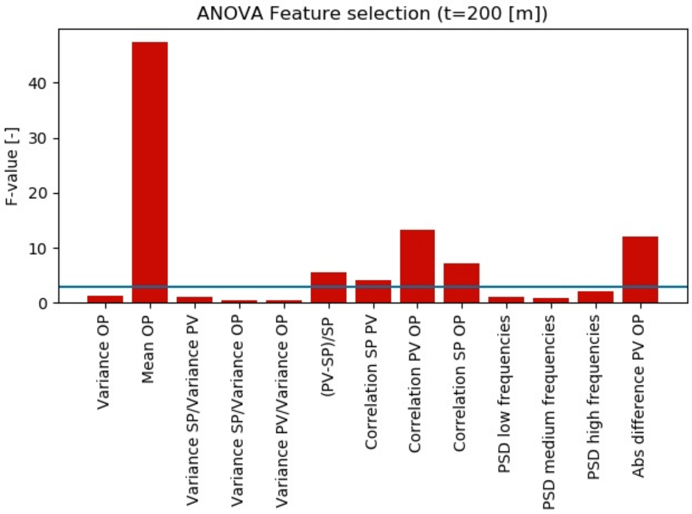
	Method 1				Method 2				Method 3			
Number	1	2	3	4	1	2	3	4	1	2	3	4
1	4	2	2	4	4	5	4	4	1	3	1	4
2	4	1	3	4	4	5	4	3	1	5	1	1
3	5	1	5	2	3	4	2	3	2	5	1	3
4	5	3	4	4	4	4	5	4	1	3	2	2
5	3	1	2	3	4	4	3	4	1	4	1	1
Average	4.2	1.6	3.2	3.4	3.8	4.4	3.6	3.6	1.2	4	1.2	2.2

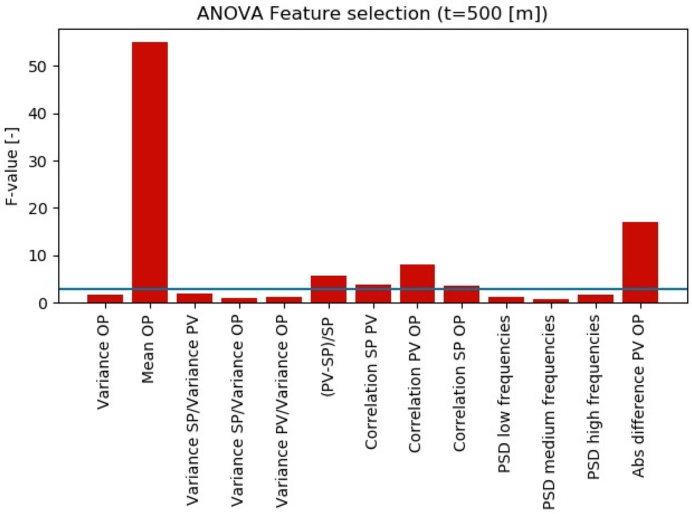
B.2. Feature selection tables

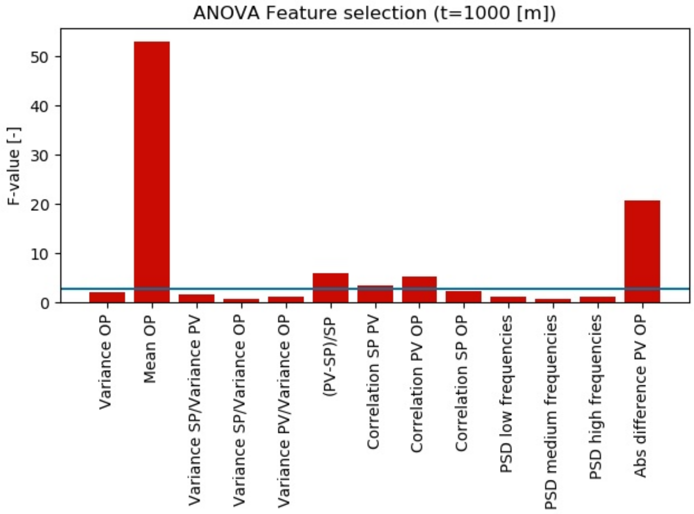
Time window	Number of features	Features
t=200 [m]	6	Mean OP, PV-SP/SP, Correlation SP PV, Correlation SP OP, Correlation PV OP, Absolute difference PV OP
t=500 [m]	6	Mean OP, PV-SP/SP, Correlation SP PV, Correlation SP OP, Correlation PV OP, Absolute difference PV OP
t=1000 [m]	5	Mean OP, PV-SP/SP, Correlation SP PV, Correlation PV OP, Absolute difference PV OP
t=2000 [m]	5	Mean OP, PV-SP/SP, Correlation SP PV, Correlation PV OP, Absolute difference PV OP
t=5000 [m]	5	Variance OP, Mean OP, PV-SP/SP, Correlation PV OP, Absolute difference PV OP

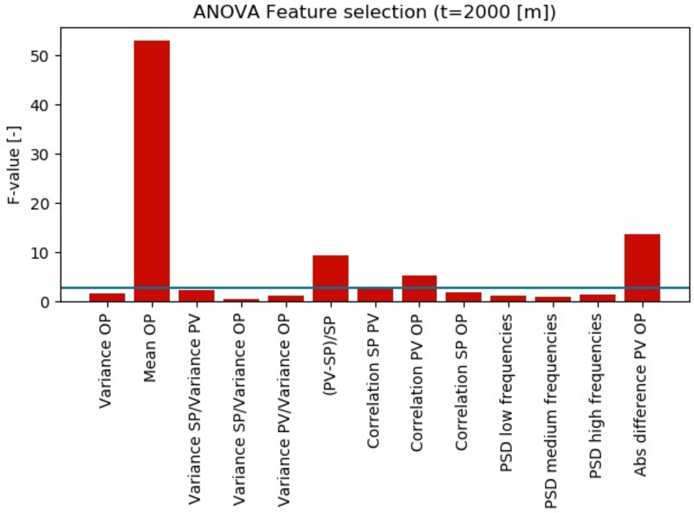
Time window	Number of features	Features
t=200 [m]	7	Mean OP, Variance SP/Variance PV, PV-SP/SP, Correlation SP PV, Correlation SP OP, Correlation PV OP, Absolute difference PV OP
t=500 [m]	8	Mean OP, Variance SP/Variance PV, PV-SP/SP, Correlation SP PV, Correlation SP OP, Correlation PV OP, PSD high frequencies, Absolute difference PV OP
t=1000 [m]	9	Mean OP, Variance SP/Variance PV, PV-SP/SP, Correlation SP PV, Correlation SP OP, Correlation PV OP, PSD high frequencies, PSD medium frequencies, Absolute difference PV OP
t=2000 [m]	8	Mean OP, Variance SP/Variance PV, PV-SP/SP, Correlation SP OP, Correlation PV OP, PSD low frequencies,PSD medium frequencies, Absolute difference PV OP
t=5000 [m]	6	Mean OP, Variance SP/Variance PV, PV-SP/SP, Correlation SP OP, Correlation PV OP, Absolute difference PV OP

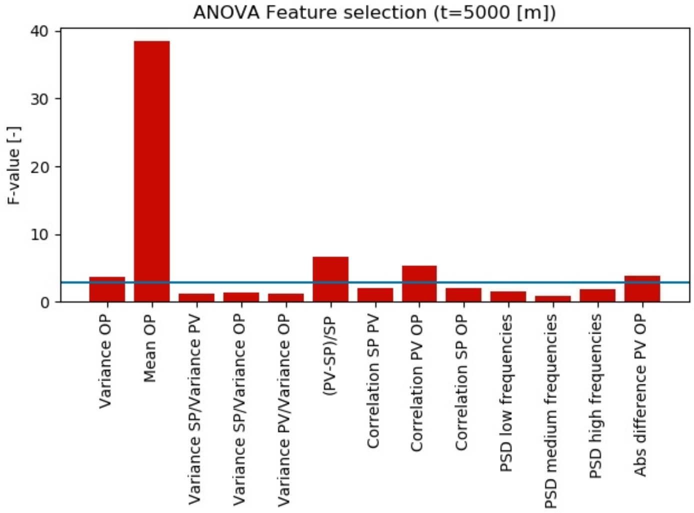
B.3. Feature selection graphs

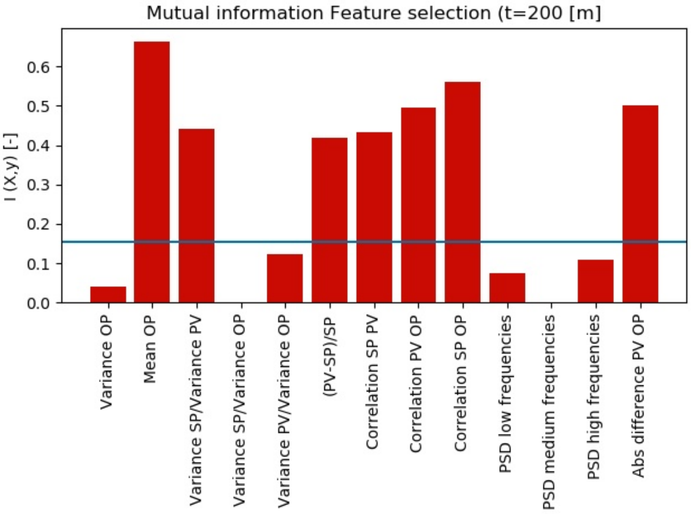


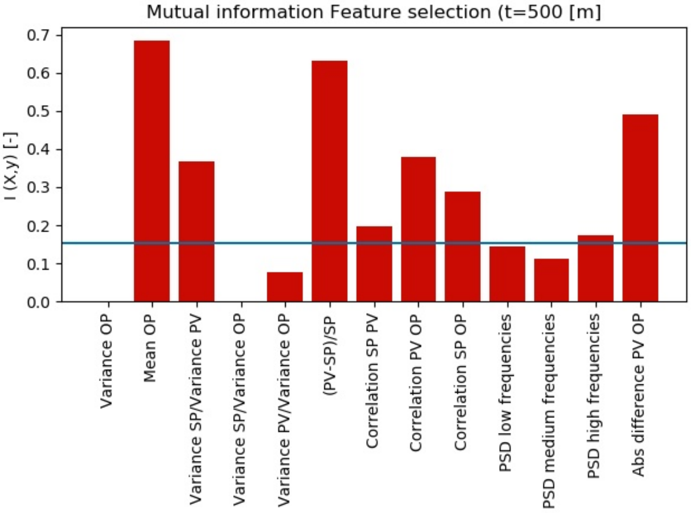


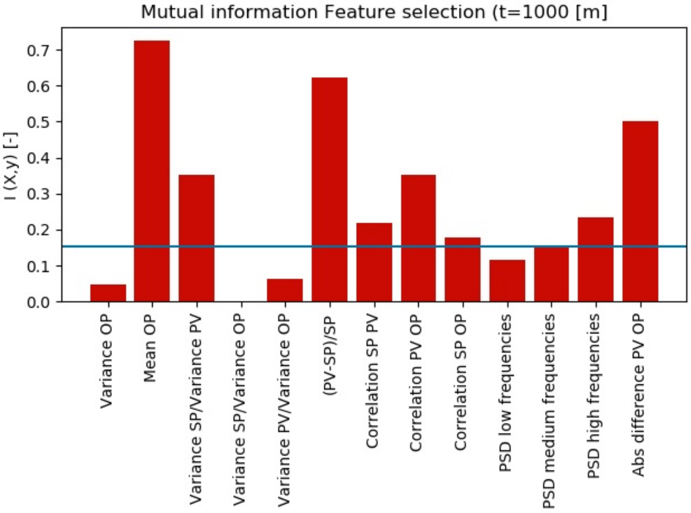

ANOVA Feature selection (t=1000 [m])
F-value [-]
0
10
20
30
40
50
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP

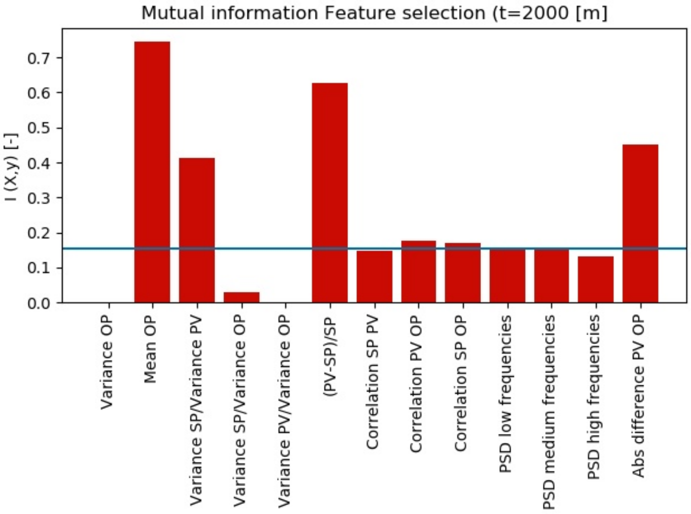

ANOVA Feature selection (t=2000 [m])
F-value [-]
0
10
20
30
40
50
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP

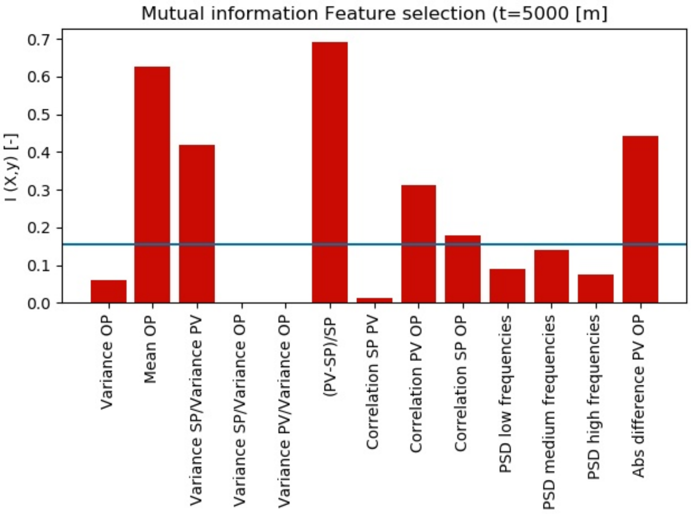

ANOVA Feature selection (t=5000 [m])
F-value [-]
0
10
20
30
40
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP


Mutual information Feature selection (t=200 [m]
I (X,y) [-]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP


Mutual information Feature selection (t=500 [m]
I (X,y) [-]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP

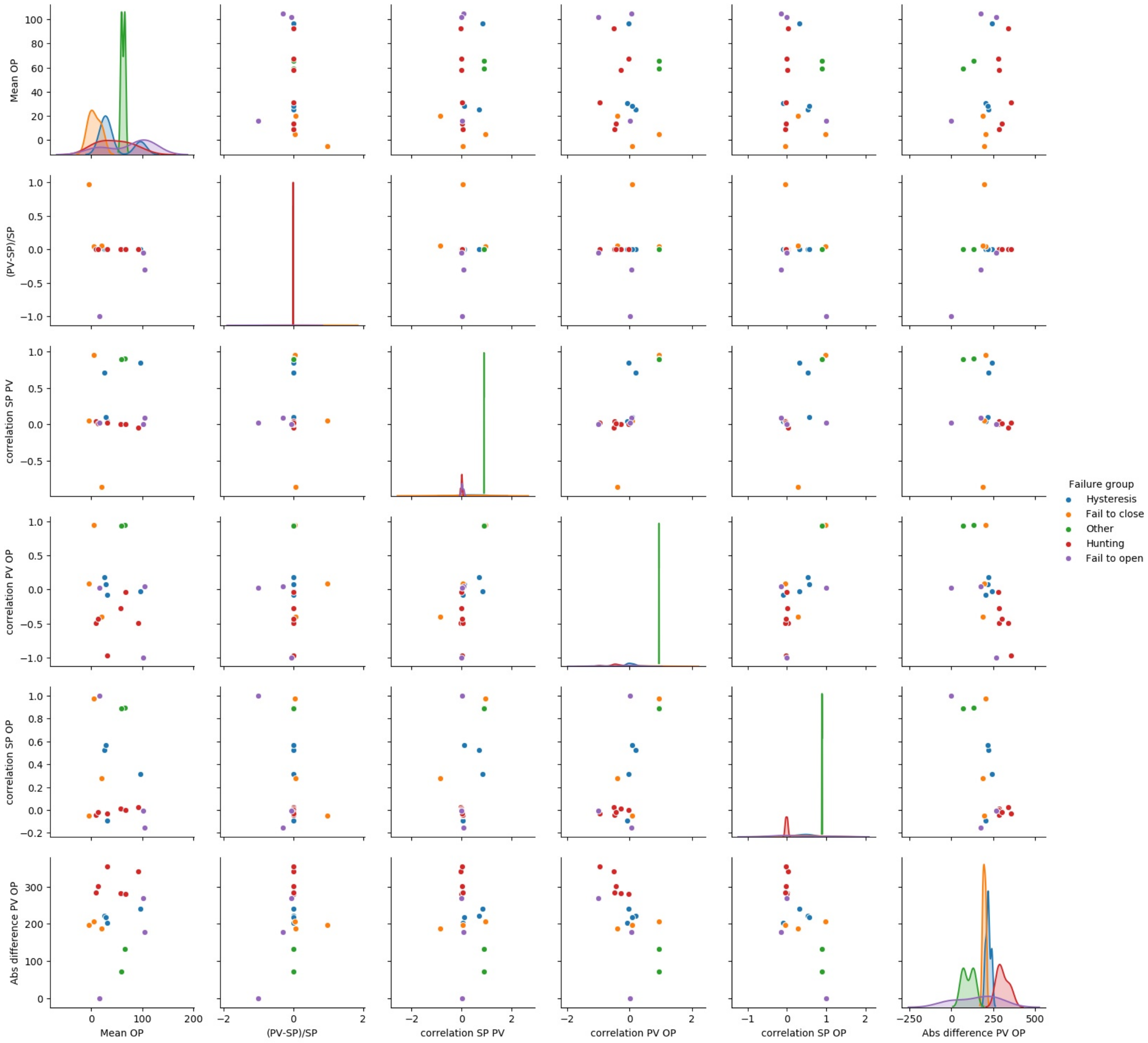

Mutual information Feature selection (t=1000 [m]
I (X,y) [-]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP


Mutual information Feature selection (t=2000 [m])
I (X,y) [-]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP

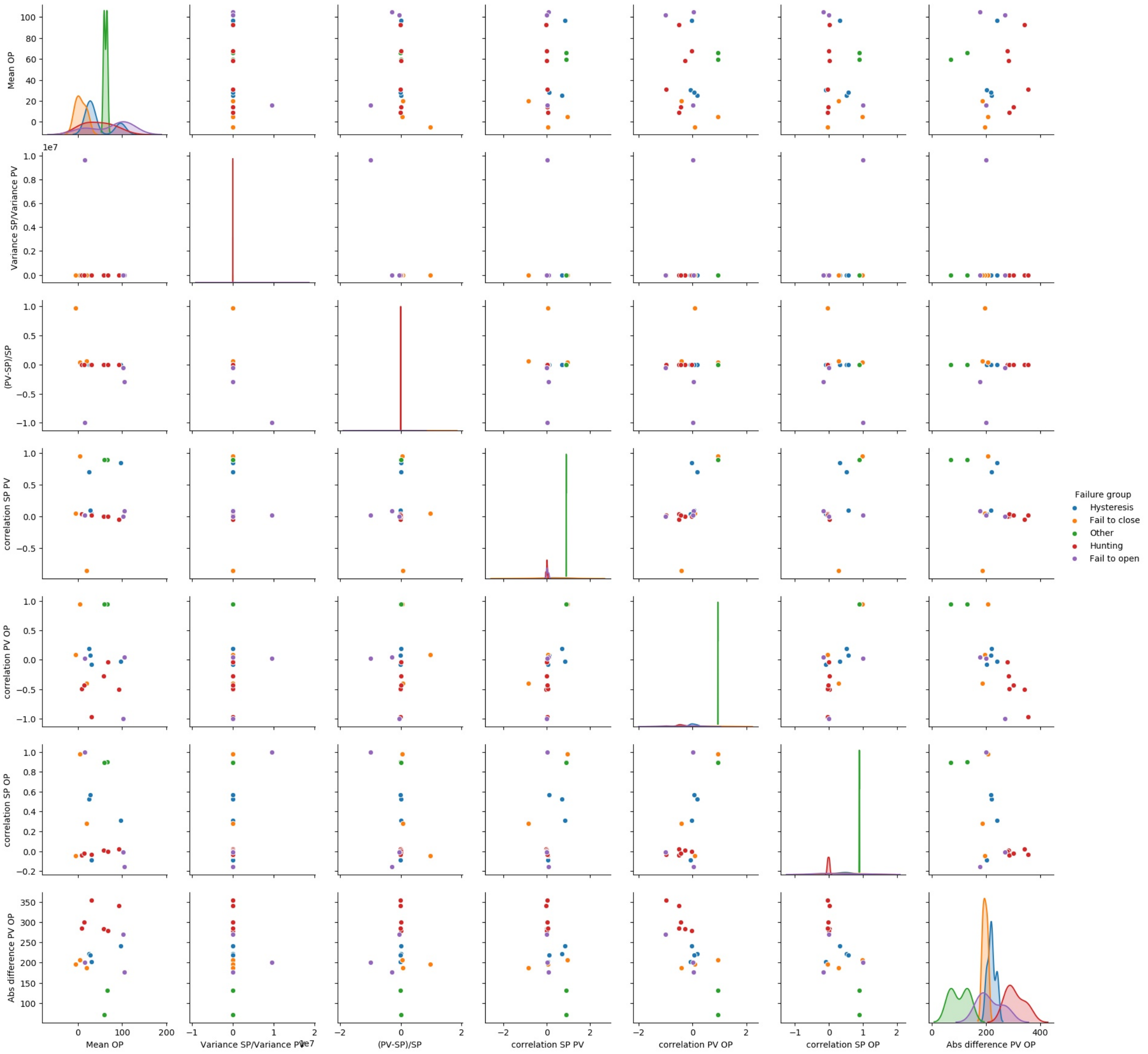

Mutual information Feature selection (t=5000 [m])
I (X,y) [-]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Variance OP
Mean OP
Variance SP/Variance PV
Variance SP/Variance OP
Variance PV/Variance OP
(PV-SP)/SP
Correlation SP PV
Correlation PV OP
Correlation SP OP
PSD low frequencies
PSD medium frequencies
PSD high frequencies
Abs difference PV OP

B.4. Case study selected feature plots

Scatter plots of the case study features selected using the ANOVA test with $t = 200[m]$.



Scatter plots of the case study features selected using the Mutual information test with $t = 200[m]$.



C

Additional research; extra case study analysis

This section contains additional research on the classification of test data using the feature selection results and the trained classifiers. First of all, the model input overview will be given. Second of all, the performance will be evaluated using the determined KPI's.

C.1. Overview

The failure cases used in the case study are described in Chapter 3.4, which sums up to 16 incidents. Table 3.5 shows the specific failures and corresponding location summing up to the distribution of samples, displayed in Table C.1.

Failure behaviour	Number of samples
Fail to close	3
Fail to open	3
Hunting	3
Hysteresis	4
Other	3

Table C.1: Distribution of samples in the five failure behaviour categories.

In Chapter 5.1 the input length of the time series $t = 200$ minutes has been determined. The corresponding six features for the ANOVA test are: mean OP, (PV-SP/SP), correlation SP PV, correlation SP OP, correlation PV OP, and the absolute difference between the PV and the OP. The Mutual information test adopted seven features in the model: mean OP, variance SP/variance PV, the ratio between PV and SP, correlation SP PV, correlation SP OP, correlation PV OP, and the absolute difference between the PV and the OP.

Furthermore, in Chapter 5.1 the optimal parameters of the classification models have been determined. The parameters of the LR model are the regularisation coefficient C, which has been set to C=1.0. The RF classifier obtains a maximum depth of the decision trees of 10.

The classification output performance will be evaluated using KPI's. Furthermore, the classification probabilities, which denotes the probabilities of an observation belonging to a particular class, are used to assess the decisions of the classification models. Appendix B.3 contains scatter plots of the selected features in the cases study.

C.2. Performance evaluation

The performance of the case study is assessed using the metrics discussed in Chapter 4.6: accuracy, log-loss.

First of all, the accuracy and Log-loss of the model are assessed for the two selection methods and three classification models. Second of all, the precision and recall values of the outcomes will be discussed. Finally, the classification results will be evaluated in more detail. The results of the accuracy and log-loss are shown in Table C.2.

	Anova			MI		
	LR	GNB	RF	LR	GNB	RF
Accuracy	0.62	0.81	0.81	0.75	0.69	0.81
Log-loss	0.24	0.1	0.46	0.28	0.15	0.52

Table C.2: Accuracy and Log-loss results from the case study.

The accuracy results when using the ANOVA selection method score better for LR and GNB. The feature added to the input of the classification model degrades the quality of the model, and thereby reduces the accuracy. In the case of the RF classifier, this does not influence due to the algorithm, which considers the features that decrease the impurity the most. The GNB and RF models obtain the highest accuracy (0.81) on the case study data set.

The Log-loss results, measuring the uncertainty in the prediction probabilities show a decrease in performance for all the classification methods when using Mutual information over the ANOVA test.

C.3. Classification output analysis

The classification outputs of the different models can be analysed more in detail to gain more knowledge on how certain predictions were made. Therefore, Appendix B.4. contains six prediction probabilities showing the chances of an observation belonging to a specific class. These results can be used to evaluate the mistakes made by the classifiers. Table C.4 contains a more in-depth analysis of the mistakes made by the classification models using the ANOVA feature selection with t=200 minutes.

LR			GNB			RF		
Number	Position	Probability	Number	Position	Probability	Number	Position	Probability
1	2nd	0.02	1	-	0	1	2nd	0.27
2	3rd	0.18	2	2nd	0.39	2	2nd	0.30
3	2nd	0.28	3	-	0	4	2nd	0.20
4	2nd	0.24						
5	-	0.0						

Table C.3: Classification position of mistakes in the predictions of ANOVA with t=200 minutes.

From the results of Table C.3 shows that for the LR model the probabilities when the correct class has not been chosen, the prediction probability of the actual class has a large variance. This allows us to conclude that in some of the predictions the LR model was completely off. When analysing the classification probabilities of the GNB classifier in Appendix B.4., the probability estimates are often binary, where the classification is perfect or completely wrong. In the RF model, it is the other way around due to the two random influences in the algorithm where random samples and random features are drawn at every decision tree. Therefore, the probability have more uncertainty, which explains the high log-loss values, however, Table C.3 shows that when a prediction mistake is made the model was often close to the correct answer resulting in a minimum probability of 0.20 for the actual class. When the diagnosis strategy is to prevent a certain failure at all costs, a specific probability threshold could be set increasing the priority of the failure alarm when the incident possibly exists. Furthermore, these results can be used to determine the accuracy if the prediction probability ends up on the first or second place. The accuracy on the first or second position is the ratio between the sum of the classes with the highest prediction probability or highest - 1 prediction probability for an observation divided by the total number of predictions. Table C.4 compares the accuracies on just the first position to the accuracy when the first and second place are allowed.

	Accuracy 1st position	Accuracy 1st or 2nd position
LR	0.67	0.88
GNB	0.81	0.88
RF	0.81	1.0

Table C.4: First and second position accuracy.

The results from Table C.4 show that the RF method in combination with the ANOVA method for 100% of the failure samples in the case study the prediction of the classifier was correct when also including the second best option of the model. When building an extremely robust failure diagnosis model, where the goal is to ensure specific failures do not occur in valves, this high second position accuracy can be very useful.

C.4. Conclusion

The case study results show six outcomes assessed using the KPI's, accuracy, log loss described in Chapter 4.6. On the metric accuracy, the RF and GNB methods in combination with the ANOVA selection method scores the highest with an accuracy fraction of 0.81. On log loss, with a value between 0 and 1 where 0 measuring the uncertainty in prediction outcomes, the GNB paired with the ANOVA scores the best.

More conclusions can be drawn from the results of the case study when looking more closely into the classification mistakes made. When new strategies for classification are applied, such as the prevention of certain failure types for valves, the first and second position accuracy or alarms above a certain probability level can be used by analysing the prediction probability outcomes. The RF method, in combination with the ANOVA selection, performs optimally on this first or second position accuracy, shown in Table C.4. Using such a strategy allows the user of the method to classify and predict control valve failures with 100% accuracy.

C.5. Classification probabilities

Classification probabilities of LR classifier, with ANOVA feature selection and t=200[m].

	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0.0	0.97	0.0	**0.02**	0.01
2	Hysteresis	Hysteresis	0.68	0.0	0.07	**0.22**	0.03
3	Fail to close	Fail to close	**0.97**	0.0	0.01	0.01	0.01
4	Other	Other	0.0	0.33	0.0	0.11	**0.56**
5	Hunting	Hysteresis	0.0	0.29	**0.18**	0.52	0.01
6	Other	Other	0.0	0.46	0.0	0.04	**0.5**
7	Fail to close	Fail to close	**0.99**	0.0	0.01	0.0	0.0
8	Hunting	Hunting	0.06	0.0	**0.94**	0.0	0.0
9	Hysteresis	Fail to close	0.5	0.0	0.17	**0.28**	0.06
10	Fail to close	Fail to close	**0.99**	0.0	0.0	0.01	0.0
11	Fail to open	Fail to open	0.0	**0.99**	0.0	0.01	0.0
12	Hunting	Hunting	0.0	0.0	**1.0**	0.0	0.0
13	Hysteresis	Fail to close	0.59	0.0	0.13	**0.24**	0.04
14	Fail to open	Fail to open	0.0	**0.83**	0.0	0.14	0.03
15	Fail to open	Fail to open	0.0	**0.99**	0.0	0.0	0.0
16	Other	Hunting	0.01	0.0	**0.99**	0.0	0.0

Classification probabilities of GNB classifier, with ANOVA feature selection and t=200[m].

	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Hysteresis	0.0	0.14	0.0	**0.86**	0.0
2	Hysteresis	Hysteresis	0.24	0.0	0.05	**0.71**	0.0
3	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
4	Other	Other	0.0	0.0	0.0	0.0	**1.0**
5	Hunting	Fail to open	0.0	**1.0**	0.0	0.0	0.0
6	Other	Other	0.0	0.0	0.0	0.0	**1.0**
7	Fail to close	Fail to close	**0.56**	0.0	0.0	0.44	0.0
8	Hunting	Hunting	0.0	0.0	**1.0**	0.0	0.0
9	Hysteresis	Hysteresis	0.0	0.0	0.0	**1.0**	0.0
10	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
11	Fail to open	Fail to open	0.0	**1.0**	0.0	0.0	0.0
12	Hunting	Hunting	0.0	0.0	**1.0**	0.0	0.0
13	Hysteresis	Hysteresis	0.01	0.0	0.0	**0.98**	0.0
14	Fail to open	Hysteresis	0.0	0.39	0.0	**0.61**	0.0
15	Fail to open	Fail to open	0.0	**1.0**	0.0	0.0	0.0
16	Other	Hunting	0.0	0.0	**1.0**	0.0	0.0

Classification probabilities of RF classifier, with ANOVA feature selection and t=200[m].

	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0.0	0.65	0.05	**0.27**	0.03
2	Hysteresis	Hysteresis	0.18	0.09	0.23	**0.48**	0.02
3	Fail to close	Fail to close	**0.79**	0.05	0.0	0.01	0.15
4	Other	Other	0.03	0.34	0.0	0.19	**0.44**
5	Hunting	Fail to open	0.01	0.68	**0.3**	0.01	0.0
6	Other	Other	0.04	0.37	0.0	0.12	**0.47**
7	Fail to close	Fail to close	**0.79**	0.06	0.0	0.15	0.0
8	Hunting	Hunting	0.06	0.08	**0.86**	0.0	0.0
9	Hysteresis	Hysteresis	0.23	0.07	**0.19**	0.45	0.06
10	Fail to close	Fail to close	**0.77**	0.02	0.19	0.45	0.06
11	Fail to open	Fail to open	0.04	**0.71**	0.21	0.03	0.01
12	Hunting	Hunting	0.02	0.07	**0.9**	0.01	0
13	Hysteresis	Hysteresis	0.19	0.08	0.18	0.52	0.03
14	Fail to open	Fail to open	0.06	0.48	0.05	**0.41**	0.0
15	Fail to open	Fail to open	0.03	0.68	0.0	0.21	0.08
16	Other	Hunting	0.02	0.02	**0.69**	0.07	0.2

Classification probabilities of LR classifier, with Mutual information feature selection and t=200[m].

	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0.0	0.96	0.0	**0.03**	0.01
2	Hysteresis	Fail to close	1.0	0.0	0.0	**0.0**	0.0
3	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
4	Other	Other	0.0	0.35	0.0	0.09	**0.56**
5	Hunting	Hysteresis	0.0	0.32	**0.13**	0.53	0.01
6	Other	Other	0.0	0.48	0.0	0.03	**0.49**
7	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
8	Hunting	Hunting	0.09	0.0	**0.91**	0.0	0.0
9	Hysteresis	Hysteresis	0.32	0.0	0.22	**0.41**	0.05
10	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
11	Fail to open	Fail to open	0.0	**0.99**	0.0	0.01	0.0
12	Hunting	Hunting	0.0	0.0	**1.0**	0.0	0.0
13	Hysteresis	Hysteresis	0.33	0.0	0.19	**0.43**	0.05
14	Fail to open	Hysteresis	0.0	**0.98**	0.0	0.0	0.02
15	Fail to open	Fail to open	0.0	**0.99**	0.00	0.0	0.01
16	Other	Hunting	0.0	0.0	0.99	0.0	**0.01**

Classification probabilities of GNB classifier, with Mutual information feature selection and t=200[m].

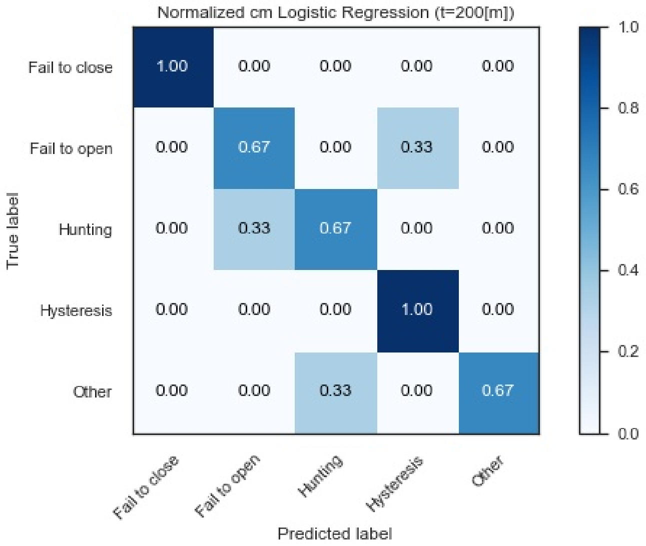
	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0.0	0.02	0.0	**0.98**	0.0
2	Hysteresis	Fail to close	0.97	0.03	0.0	**0.0**	0.0
3	Fail to close	Fail to close	**0.99**	0.01	0.0	0.0	0.0
4	Other	Other	0.0	0.0	0.0	0.0	**1.0**
5	Hunting	Fail to open	0.0	0.94	**0.06**	0.0	0.0
6	Other	Fail to open	0.0	1.0	0.0	0.0	**0.0**
7	Fail to close	Hysteresis	**0.01**	0.0	0.0	0.99	0.0
8	Hunting	Hunting	0.0	0.00.0	**0.99**	0.01	0.0
9	Hysteresis	Hysteresis	0.0	0.0	0.0	**1.0**	0.0
10	Fail to close	Fail to close	**1.0**	0.0	0.0	0.0	0.0
11	Fail to open	Fail to open	0.0	**1.0**	0.0	0.0	0.0
12	Hunting	Hunting	0.0	0.0	**1.0**	0.0	0.0
13	Hysteresis	Hysteresis	0.0	0.0	0.01	**0.99**	0.05
14	Fail to open	Fail to open	0.0	**1.0**	0.0	0.0	0.0
15	Fail to open	Fail to open	0.0	**1.0**	0.00	0.0	0.0
16	Other	Hunting	0.0	0.0	1.0	0.0	**0.0**

Classification probabilities of RF classifier, with Mutual information feature selection and t=200[m].

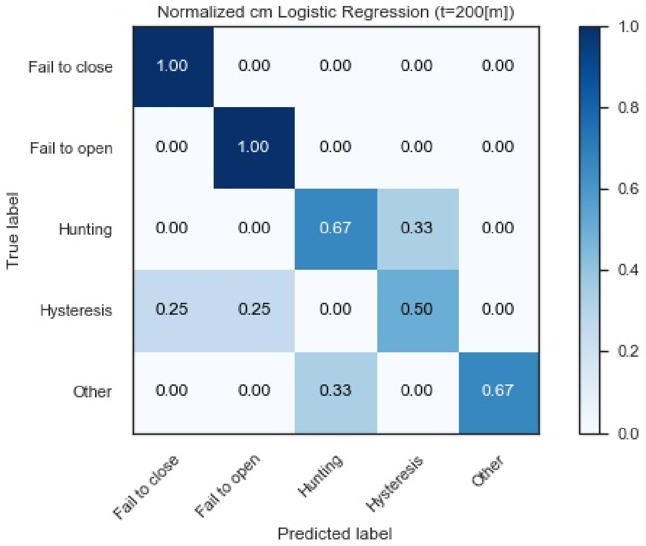
	Classification		Classification probabilities				
	Actual	Predicted	Fail to close	Fail to open	Hunting	Hysteresis	Other
1	Hysteresis	Fail to open	0.01	0.57	0.11	**0.29**	0.02
2	Hysteresis	Hysteresis	0.29	0.14	0.17	**0.39**	0.01
3	Fail to close	Fail to close	**0.73**	0.08	0.0	0.01	0.18
4	Other	Other	0.15	0.32	0.0	0.15	**0.38**
5	Hunting	Fail to open	0.0	0.53	**0.42**	0.05	0.0
6	Other	Other	0.16	0.29	0.0	0.14	**0.41**
7	Fail to close	Fail to close	**0.83**	0.02	0.0	0.11	0.04
8	Hunting	Hunting	0.12	0.09	**0.71**	0.07	0.01
9	Hysteresis	Hysteresis	0.20	0.04	**0.29**	0.47	0.0
10	Fail to close	Fail to close	**0.76**	0.08	0.13	0.03	0.0
11	Fail to open	Fail to open	0.02	**0.65**	0.21	0.12	0.0
12	Hunting	Hunting	0.02	0.08	**0.89**	0.01	0
13	Hysteresis	Hysteresis	0.21	0.04	0.27	0.48	0.0
14	Fail to open	Fail to open	0.15	0.47	0.01	**0.36**	0.01
15	Fail to open	Fail to open	0.05	**0.75**	0.0	0.12	0.08
16	Other	Hunting	0.02	0.02	**0.69**	0.07	0.20

C.6. Confusion matrices

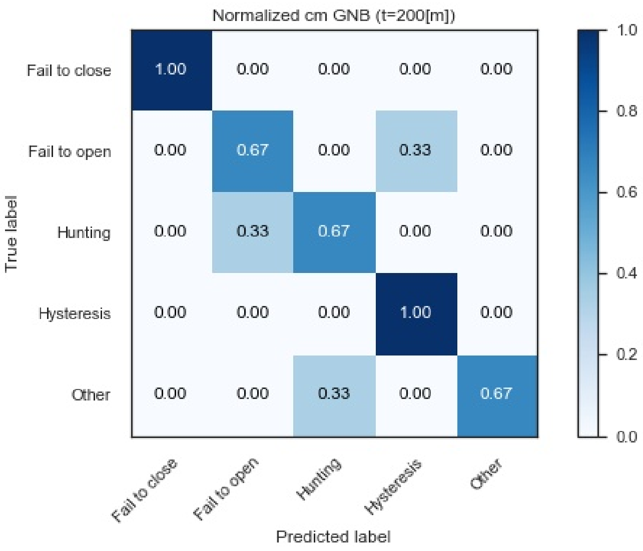
Confusion matrix of LR model using ANOVA (t=200).



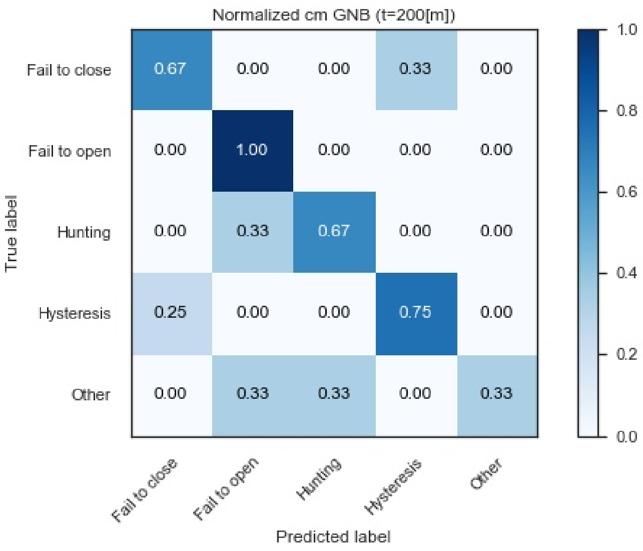
Confusion matrix of LR model using MI (t=200).



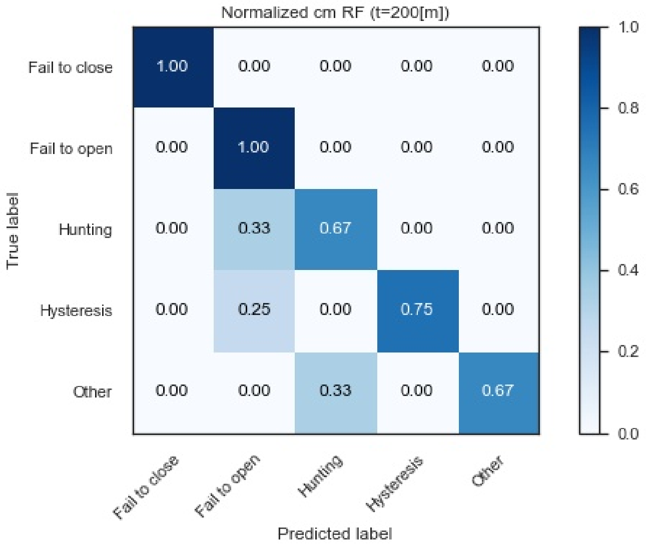
Confusion matrix of GNB model using ANOVA (t=200).



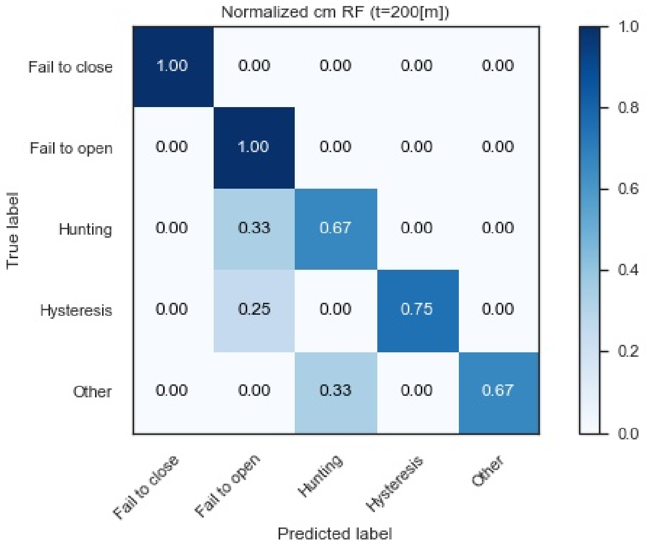
Confusion matrix of GNB model using MI (t=200).



Confusion matrix of RF model using ANOVA (t=200).



Confusion matrix of RF model using MI (t=200).


Normalized cm RF (t=200[m])
True label
Fail to close
Fail to open
Hunting
Hysteresis
Other
Predicted label