Novelty Detection from Temperature Oscillation Data

Helena Hilander

School of Electrical Engineering

Bachelor's thesis
Espoo 1.9.2020

Supervisor

Dr Markus Turunen

Advisors

MSc. Ilari Kampman

Dr Charles Gadd

Aalto University
School of Electrical
Engineering

Aalto University
School of Electrical
Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Helena Hilander

Title Novelty Detection from Temperature Oscillation Data

Degree programme Electrical Engineering

Major Bioinformation Technology **Code of major** ELEC3016

Teacher in charge Dr Markus Turunen

Advisors MSc. Ilari Kampman, Dr Charles Gadd

Date 1.9.2020 **Number of pages** 45 **Language** English

Abstract

This thesis focuses on industrial failure detection. The interest is to find a suitable machine-learning method for recognizing malfunctioning of industrial pipes based on temperature data.

The most suitable machine-learning technique for recognizing malfunctioning of industrial pipes was proven to be novelty detection. Novelty detection refers to the task of classifying unseen data as normal or abnormal. In other words, novelty detection model is trained with normal behaviour data, and new data is compared to this learned normal behaviour. New data that deviates from this learned normal behaviour, will be classified as abnormal. Novelty detection suits for industrial failure detection because no failure data is required to train the model, and industrial failure data can be very expensive to acquire.

This thesis develops a distance-based novelty detection pipeline for detecting malfunctioning based on temperature oscillation data. The pipeline includes several steps. The first steps are representing oscillation spectra by sorting them based on the magnitude of the amplitude values and extracting a median spectrum as a template to which other spectra will be compared. The hypothesis is that malfunctioning spectra differ more from this template than normal spectra. After extracting the median template, an error threshold for the difference between normally working spectra and the template will be learned. The error threshold is defined with Local Outlier Factor or DBSCAN. Spectra that differ more from the template than the error threshold will be classified as abnormal.

The pipeline with the most optimal parameters reaches 100 % accuracy in recognizing malfunctioning test spectra and 85 % accuracy in recognizing normal test spectra. The model also differentiates spectra from unsimilar pipes with 99 % accuracy, which also gives information about the ability of the model to recognise abnormal behaviour. Further improvements to the model can be made by determining malfunctioning based on several subsequent spectra rather than a single spectrum and by combining frequency domain analysis with time domain analysis.

The novelty detection pipeline will be integrated to industrial settings for further testing and validation.

Keywords Novelty detection, Data science, Frequency, Feature Engineering, Oscillation

Aalto University
School of Electrical
Engineering

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Helena Hilander

Työn nimi Poikkeamien tunnistaminen lämpötilavärähtelydatasta

Koulutusohjelma Sähkötekniikan kandidaattiohjelma

Pääaine Bioinformaatioteknologia **Pääaineen koodi** ELEC3016

Vastuuopettaja Dr Markus Turunen

Työn ohjaajat MSc. Ilari Kampman, Dr Charles Gadd

Päivämäärä 1.9.2020 **Sivumäärä** 45 **Kieli** Englanti

Tiivistelmä

Tässä kandidaatintyössä sovelletaan koneoppimistekniikoita teollisen vian tunnistamiseksi lämpötiladatasta. Lämpötiladata on hankittu antureista, jotka sijaitsevat teollisten putkien pinnalla. Putkia lämmitetään sähkönsaattopiireillä. Teollisen vian tunnistamiseksi sovelletaan koneoppimistekniikoita, jotka opetetaan normaalisti käyttäytyvän datan avulla, ilman dataa vikatilanteista. Koneoppimismalli oppii normaalin käytöksen ja luokittelee kaiken normaalista poikkeavan epänormaaliksi. Malli, joka voidaan opettaa ilman poikkeamadataa, on hyödyllinen teollisuudessa, jossa viallisen datan hankkiminen on kallista.

Kandidaatintyössä kehitetään monivaiheinen koneoppimismalli, joka tunnistaa teollisen vian lämpötilavärähtelydatan perusteella. Lämpötilavärähtelyspektrin esitysmuodolla havaitaan olevan merkittävä vaikutus mallin suorituskykyyn. Tarkimmassa koneoppimismallissa lämpötilavärähtelyspektri esitetään amplitudiarvojen suuruuksien mukaan, jonka jälkeen spektridatasta lasketaan mediaanispektri. Työn hypoteesi on, että viallinen spektridata eroaa mediaanispektristä enemmän kuin normaali spektridata. Mediaanispektrin laskemisen jälkeen normaali virhemarginaali spektrin ja mediaanimallispektrin välillä määritellään Local Outlier Factor - tai DBSCAN-menetelmällä. Spektrit, jotka poikkeavat mediaanimallispektristä opetusvaiheessa määriteltyä virhemarginaalia enemmän, luokitellaan viallisiksi.

Kandidaatintyön koneoppimismalli tunnistaa testidatan vialliset spektrit 100 %:n tarkkuudella. Normaalit spektrit tunnistetaan 85 %:n tarkkuudella, joten osa normaaleista spektreistä luokitellaan virheellisesti viallisiksi. Malli kykenee erottamaan eri putkien spektridataa toisistaan 99 %:n tarkkuudella, mikä osoittaa myös, että malli kykenee tunnistamaa poikkeavaa dataa. Poikkeamatunnistusmallia voidaan parantaa luokittelemalla teollinen vika usean ajassa peräkkäisen spektrin perusteella sen sijaan, että luokittelu tehtäisiin yhden spektrin perustella.

Työssä kehitettyä koneoppimismallia tullaan jatkokehittämään ja uudelleentestaamaan, sillä työn aikana saatavilla olevan testidatan määrä on ollut vähäinen. Jatkossa taajuustason analyysi yhdistetään aikatason analyysin kanssa, jotta myös mahdolliset viat, joissa taajuusspektri pysyy muuttumattomana, voidaan tunnistaa.

Avainsanat Taajuus, Data-analyysi, Poikkeama, Oskillaatio

Preface

This thesis is written as part of project work in Wapice Ltd.

Thank you for Ilari Kampman from Wapice Ltd, Toni Piirainen and Jari Väisänen from Planray Oy, Veli Vehviläinen and Mika Marttila from Caverion, Janne Välimäki from Forchem Oyj, and Charles Gadd from Aalto University for guidance and providing the data for this thesis.

Otaniemi, 1.9.2020

Helena Hilander

Contents

Symbols and abbreviations

Operators

$\sum_i$	sum over index i
$\mathbf{x} \cdot \mathbf{y}$	dot product of vectors $\mathbf{x}$ and $\mathbf{y}$
$\|\|\mathbf{x}\|\|$	Eucledian norm of a vector $\mathbf{x}$

Abbreviations

LOF	Local outlier factor

1 Introduction

In the last decade, the field of artificial intelligence (AI), referring to the automation of human tasks, has rapidly developed due to decreased computational costs and increased computational power. The increased efficiency of AI methods compared to human labour has recently come to the attention of different industries that have started to search for ways to apply AI. Industrial applications of AI include automatization of quality control, transportation, and failure detection, the last of which is the focus of this thesis.

Failures and malfunctioning in industrial systems can lead to mild to severe economic losses as well as unnecessary energy consumption. One type of malfunctioning which causes unnecessary energy consumption is faults occurring in processing pipes. These faults in processing pipes may be caused by insufficient insulation or heating. To prevent the economic losses and excess energy consumption of processing pipes, there is a need to find a machine-learning method which would recognise these faults in the pipes. This thesis develops a failure detection method for industrial pipes based on temperature data collected from the surface of the pipes.

Different machine learning methods for the failure detection were considered and novelty detection proved to be the most promising one. Novelty detection is defined as the task of classifying unseen data as normal or abnormal [1]. In practice, a novelty detection model is trained with normal data and the model classifies anything that differs from this learned normal behaviour as abnormal; in other words, abnormal behaviour data is not required to train the model.

A model which can be trained without malfunctioning data is a considerable advantage in industrial settings in which acquiring malfunctioning data may require a total shutdown of the machinery, leading to significant economic losses. Other machine learning models, such as multi-class classification, cannot be used to solve the problem since they require malfunctioning data for training.

As a solution for the problem of recognizing malfunctioning based on temperature data, this thesis develops a novelty detection pipeline for temperature oscillation data. The pipeline consists of several data representations and machine learning steps that together form the data analysis process. The pipeline with the most optimal parameters shows the potential to be applied in industrial settings.

The structure of this thesis is as follows: Chapter 2 is devoted to understanding the temperature data and its possible representations. After discussing the data, Chapter 3 introduces the novelty detection pipeline and the theories of the two machine learning methods applied in the pipeline: Local Outlier Factor and DBSCAN. The pipeline will be applied with temperature oscillation data in Chapters 4 and 5. Chapter 4 applies the pipeline with Local Outlier Factor and Chapter 5 with DBSCAN. Chapter 6 concludes the thesis and proposes improvements.

2 Feature engineering

Data may be represented in several ways and the selected representation affects the performance of a machine-learning model [2]. The task of finding the most optimal representation of data given the used machine-learning model is referred to as Feature Engineering [2].

The goal of this chapter is to consider the possible representations of the temperature data dealt with in this thesis. The temperature data will be introduced in time domain (Section 2.1) but most of the analysis and feature engineering is conducted with temperature oscillation spectra in frequency domain (Section 2.2).

2.1 Feature engineering in time domain

The data dealt with in this thesis is temperature data from industrial pipes that are being heated. The temperature has a manually set target that should be reached.

This thesis has two interests regarding the temperature data:

1. Learn a normal temperature behaviour of a pipe and recognise when this pipe starts to malfunction.

2. Differentiate the temperature behaviour of unsimilar pipes

The second task of differentiating data from unsimilar pipes was chosen to support the first task in situation in which not much malfunctioning testing data is available. The ability of the model to recognise unsimilar pipes describes the ability of the model to detect abnormal data and therefore suggests the ability of the model to detect malfunctioning.

The data is originally represented as a time series. Examples of the temperature time series from six pipes are shown in Figure 3. Pipe 3 (f) clearly cannot reach its target and is malfunctioning. Pipes 3 (a) and 3 (b) oscillate around their target temperature while pipes 3 (c), (d) and (e) oscillate slightly below their target. All pipes 3 (a), (b), (c), (d), (e) average less than a degree away from their target despite 3 (c), (d) and (e) oscillating slightly below their targets. Pipes 3 (b), (c) and (e) also have some outlier pikes.

To gain understanding of how the same pipe behaves in normal and malfunctioning situation, a pipe referred to as Pipe 3 was purposely malfunctioned for six hours. The temperature of Pipe 3 exhibit clean oscillation around the target temperature in situations in which the pipe is functioning correctly. This normal behaviour is shown in Figure 1. In malfunctioning situation shown in Figure 2, the temperature oscillation becomes noisier and the temperature averages below target temperature.

It should be noted that the simulation of malfunctioning was conducted with only one pipe. Pipes are not identical and therefore the malfunctioning behaviour may differ in other pipes. Also, only one type of malfunctioning was tested, and other types of malfunctioning may exhibit different behaviour even within the same pipe.

The malfunctioning simulation with Pipe 3 showed that temperature drops slightly and oscillation becomes noisier at least in one type of malfunctioning situation. A

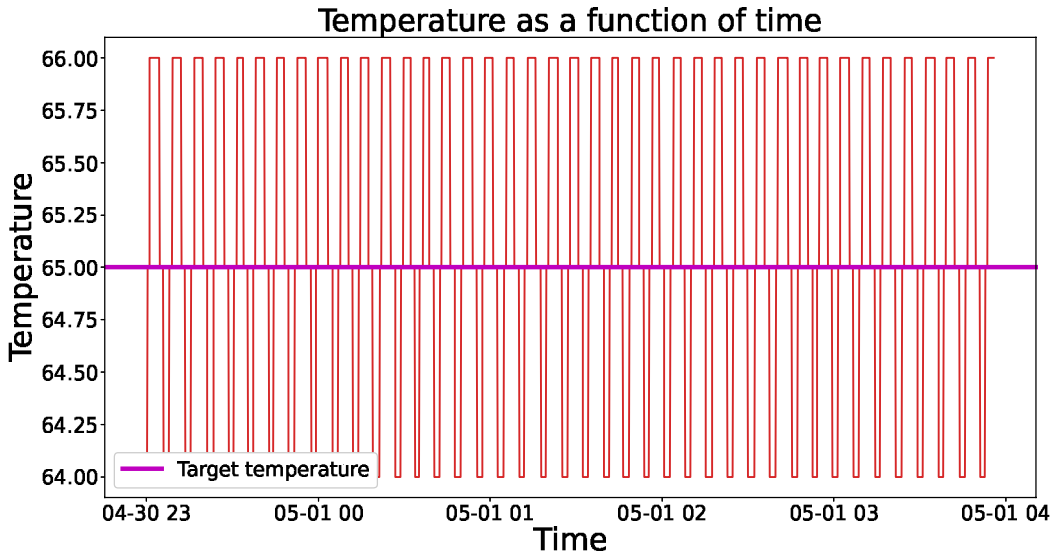


Figure 1: Timeseries from Pipe 3 working normally.

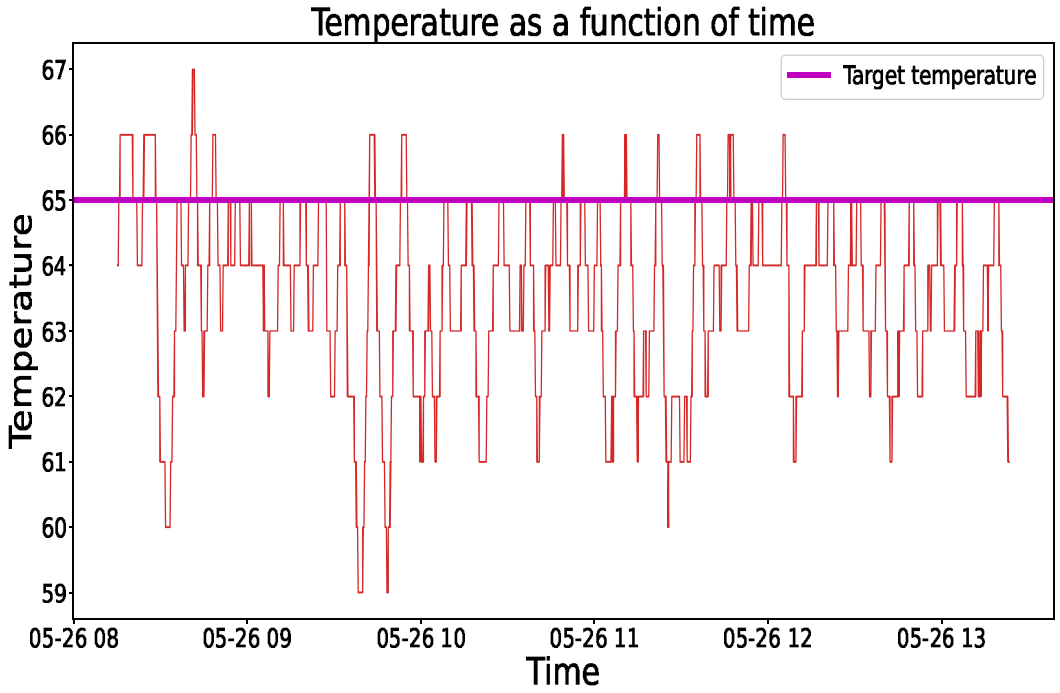


Figure 2: Timeseries from Pipe 3 malfunctioning.

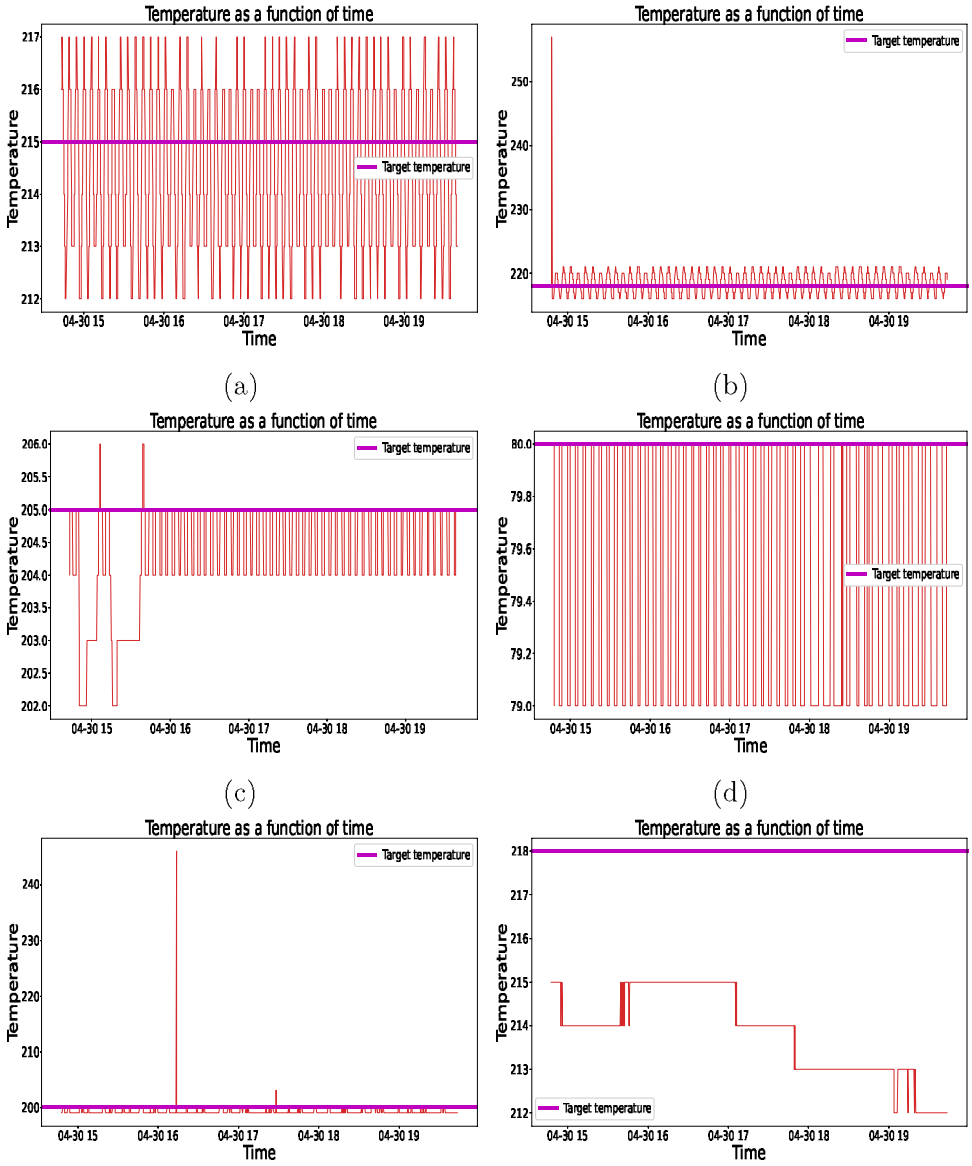


Figure 3: Temperature time series from six different pipes. The frequency domain representations of the same pipes are shown in Figure 4

very simple method for detecting this malfunctioning would be the following of the temperature median. Median over average would be preferred because median is more robust to outliers, and outliers are likely to be present in the data according to Figure 3. If the median temperature would drop around a specified threshold, pipe would be classified to malfunction. However, there is a problem with this simple approach if used alone. Figure 2 shows that malfunctioning of Pipe 3 has two characteristics: drop in the temperature and change in the oscillation pattern. If only the median temperature is followed, changes in the oscillation pattern are ignored.

Following only the median temperature would allow the detection of most malfunctioning situations but combining this with oscillation pattern analysis assures malfunctioning detection also in the possible scenarios in which only the oscillation pattern changes. Only one malfunctioning simulation is not enough to exclude the possibility that some malfunctioning situations exhibit only one of the two characteristics: drop in the temperature but no changes in the oscillation pattern or changes in the oscillation pattern but no drop in the temperature.

As a conclusion, the most accurate novelty detection model will be achieved by combining time and frequency domain analysis. Rest of this thesis will focus on developing the novelty detection model in frequency domain. The justification for focusing on frequency domain analysis is that time domain analysis would be a rather simple calculation of the median temperature and the customer of this thesis project already has a good offset to follow the median temperature.

2.2 Feature engineering in frequency domain

The focus of this thesis is on frequency domain feature engineering and novelty detection with the goals of

1. learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction and

2. differentiating the temperature behaviour of unsimilar pipes.

The original time series representation of the temperature data is being converted to frequency domain with Fast Fourier Transform. As an example, the frequency domain representations of six pipes are shown in Figure 4. The time domain representations of the same pipes are shown in Figure 3. Comparison of the frequency and time domain representations show that pipes oscillating around the target temperature seem to have less noisier spectra than pipes oscillating slightly below the target, but no conclusions should be made with such a small amount of data.

As mentioned in the previous section, originally there was no data of malfunctioning and normal behaviour from the same pipe, but Pipe 3 was purposely malfunctioned for six hours. An example spectrum of the normal behaviour of Pipe 3 is shown in Figure 5 and an example of the malfunctioning spectrum in Figure 6. Based on 178 normal spectra and 24 malfunctioning spectra of Pipe 3, there is a clear increase in the noise and the DC component in a malfunctioning situation.

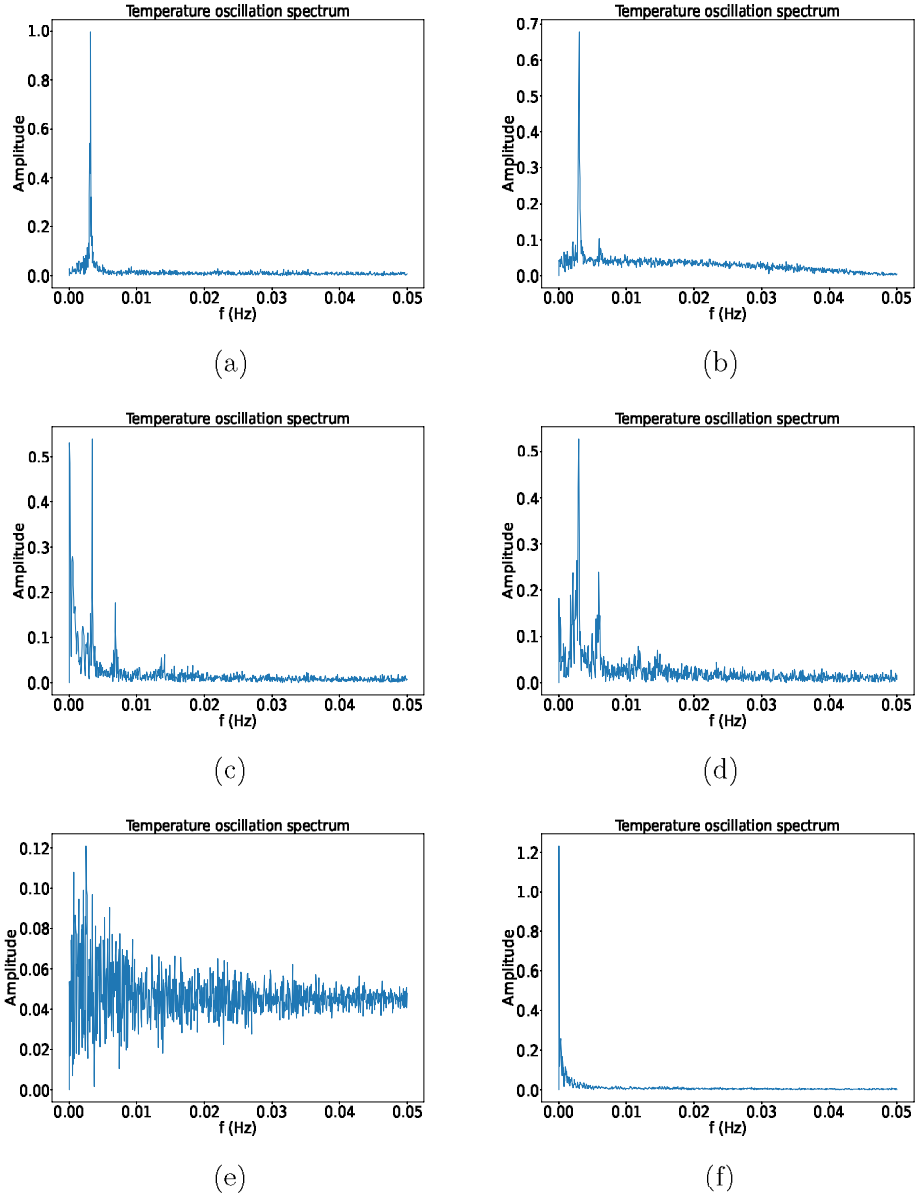


Figure 4: Temperature frequency spectra from six different pipes. The time domain representation of the same pipes is shown in Figure 3

The malfunctioning simulation of Pipe 3 proposes the differences between normal and malfunctioning pipes and may be used to develop the failure detection model. However, some pipes may show behaviour not observed in this data and further testing and possible development of the model is important after initial implementation to the industrial setting in order to increase the accuracy.

Normal and malfunctioning data of Pipe 3 will be used to train and test the novelty detection pipeline developed in this thesis. Testing the model with Pipe 3

data describes the ability of the model to recognise malfunctioning behaviour. If the pipeline performs well in this first task, second test is performed of how well this model differentiates the normal spectra of Pipe 3 from spectra of other machines. The second test supports the first: if a machine-learning model is able to differentiate two similar pipes, it is more likely to perform better in recognizing a malfunctioning pipe.

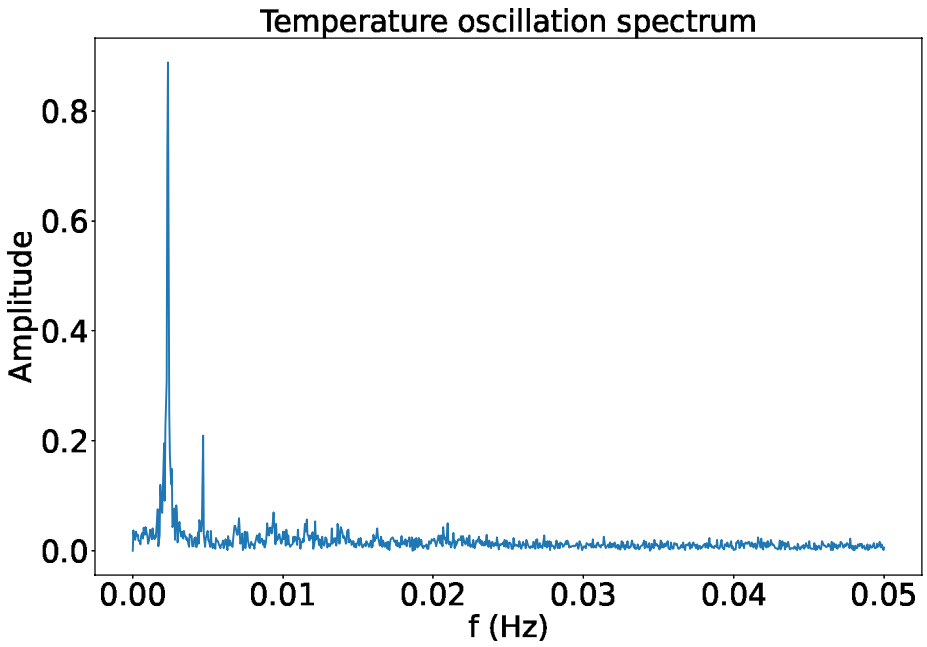


Figure 5: Spectrum from Pipe 3 working normally.

Other spectra compared to Pipe 3 in the second task are from pipes named Pipe 1, Pipe 2 and Pipe 7. Pipe 1 is known to malfunction (Figure 7). Pipe 2 (Figure 8) exhibit similar behaviour to Pipe 3 (Figure 5). The spectra of Pipe 7 (Figure 9) are noisier compared to Pipe 3.

In time domain, Pipe 2 oscillates around the target temperature (Figure 10 a) similarly to Pipe 3 (Figure 1). Pipe 7 (Figure 10 b) oscillates slightly below its target temperature while Pipe 1 has a temperature above the target (Figure 10 c).

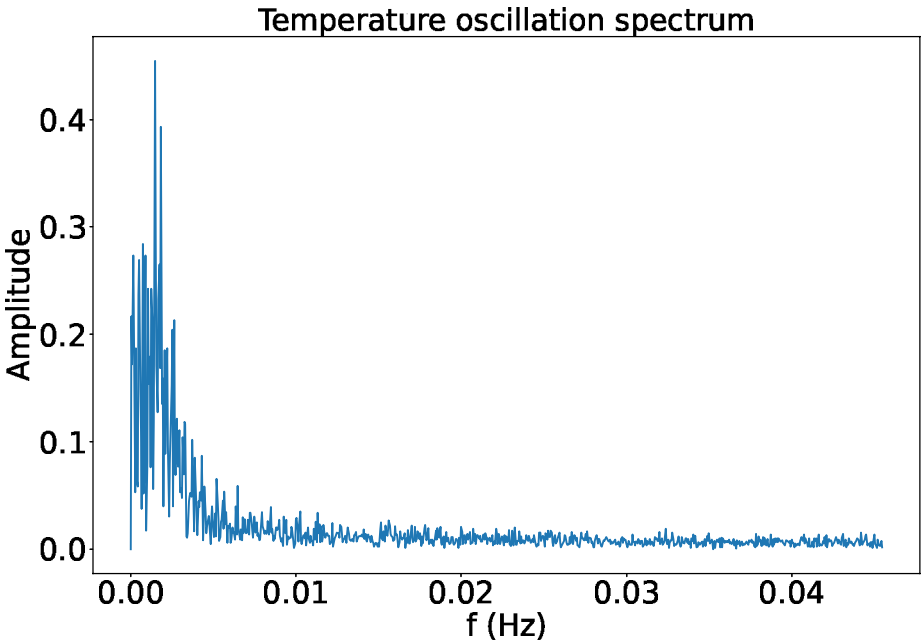


Figure 6: Spectrum from Pipe 3 malfunctioning.

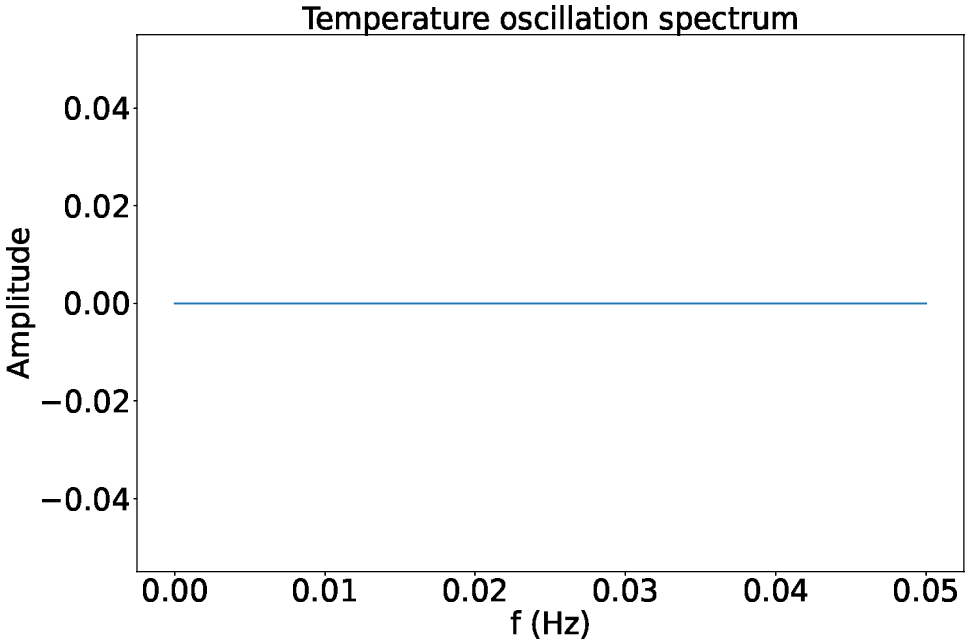


Figure 7: Spectrum from Pipe 1.

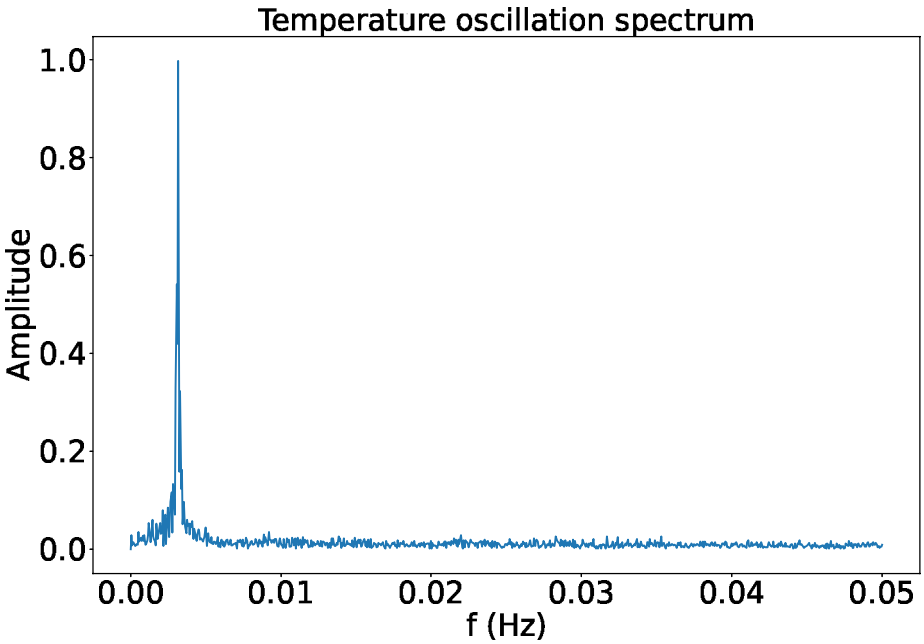


Figure 8: Spectrum from Pipe 2.

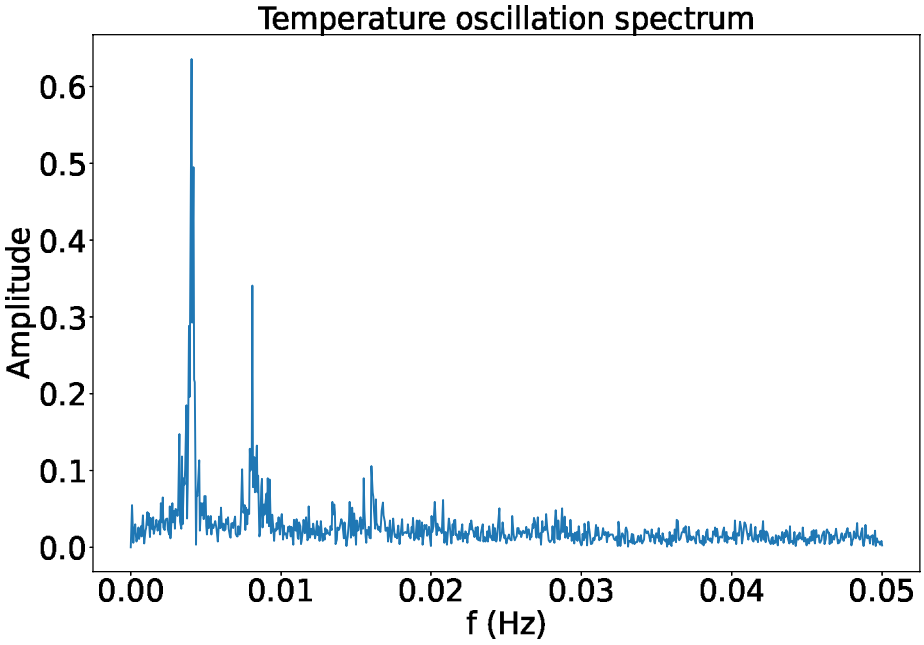


Figure 9: Spectrum from Pipe 7.

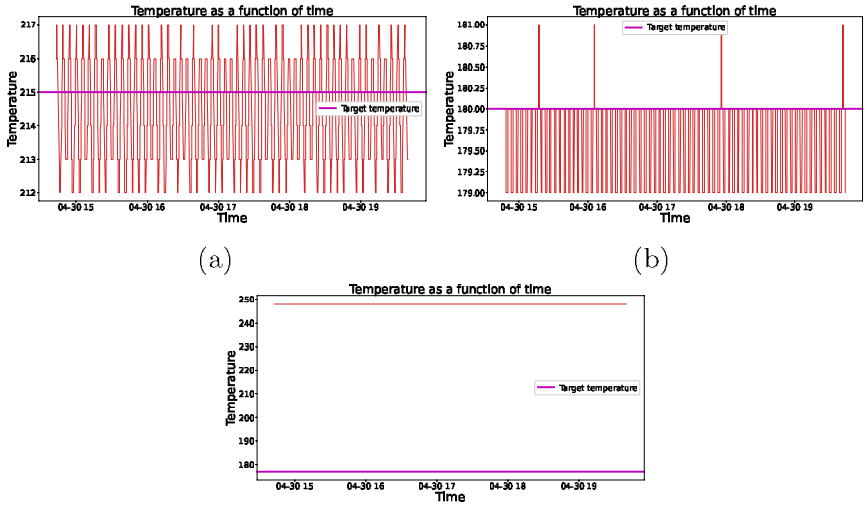


Figure 10: Time series of (a) Pipe 2 , (b) Pipe 7 and (c) Pipe 1.

A frequency spectrum may have different representations and may be sorted by the magnitude of amplitude values or by the magnitude of frequency values. Spectrum sorted by the magnitude of frequency values is the more common representation form shown in Figures 4, 5, 6, 7, 8 and 9. The selected representation affects the comparison of two spectra and therefore has an effect on novelty detection. For example, if two spectra are being compared by their distance, the distance computed between two spectra has likely a different value when calculated between 100 largest amplitude values and their corresponding frequencies compared to 100 largest frequency values and their corresponding amplitudes.

Mathematical interpretation of a spectrum is a a multidimensional array with frequency and amplitude coordinates. A frequency spectrum has ℓ frequency values $F_1, F_2, F_3, ..., F_\ell$ and ℓ amplitude values $A_1, A_2, A_3, \ldots, A_\ell$. Each frequency value $F_i, i \in [1, \ell]$ has a dependent amplitude value $A_i, i \in [1, \ell]$. Spectrum may be considered as a multidimensional array with 2ℓ coordinates in which each $F_i, i \in [1, \ell]$ and $A_i, i \in [1, \ell]$ exist in their own subspaces. However, due to the dependency of amplitude and frequency pairs, $A_i, i \in [1, \ell]$ must always correspond to the amplitude value of frequency $F_i, i \in [1, \ell]$.

A spectrum representation based on the magnitude of frequency is formed by sorting the frequency components from smallest to largest and ordering the amplitude values to the same order as their corresponding frequencies. ℓ frequency components and ℓ amplitudes form a 2ℓ dimensional array. Let this representation be referred to as conventional spectrum representation since this is the common way to represent a frequency spectrum. Figure 11 visualizes the conventional spectrum representation

Definition 2.1. Conventional spectrum representation Let $F_1, F_2, F_3, ..., F_\ell$ represent the frequency values ordered from smallest to largest and $A_1, A_2, A_3, \ldots, A_\ell$ their corresponding frequencies. conventional spectrum representation of a spectrum is defined as a one-dimensional array $(F_1, F_2, F_3, ..., F_\ell,\ A_1, A_2, A_3, \ldots, A_\ell)$ with 2ℓ dimensions.

Conventional spectrum representation sorts spectrum based on the magnitude of frequency values but spectra may also be sorted based on the magnitude of amplitude values. Amplitudes may be sorted such that $A_1, A_2, A_3, .., A_n$ represent the n largest amplitudes in decreasing order and $F_1, F_2, F_3, ..., F_n$ their corresponding frequencies. The dimensions of the spectrum can be reduced by selecting $n < \ell$. Let this representation based on the magnitude of the amplitude be referred to as amplitude sorted representation. Figure 12 visualizes amplitude sorted representation with dimensions reduced to 20. As dimensions are reduced to 20, 10 largest amplitudes and 10 of their corresponding frequencies are left. The largest amplitudes and their corresponding frequencies are referred to as component pairs.

Definition 2.2. Amplitude sorted representation Let $A_1, A_2, A_3, .., A_n$ represent the n largest amplitudes in decreasing order and $F_1, F_2, F_3, ..., F_n$ their corresponding frequencies. Biggest amplitude representation is defined as a one-dimensional array $(\ A_1, A_2, A_3, .., A_n\ ,\ F_1, F_2, F_3, ..., F_n\ )$ with $2n$ dimensions, $n \leq \ell$.

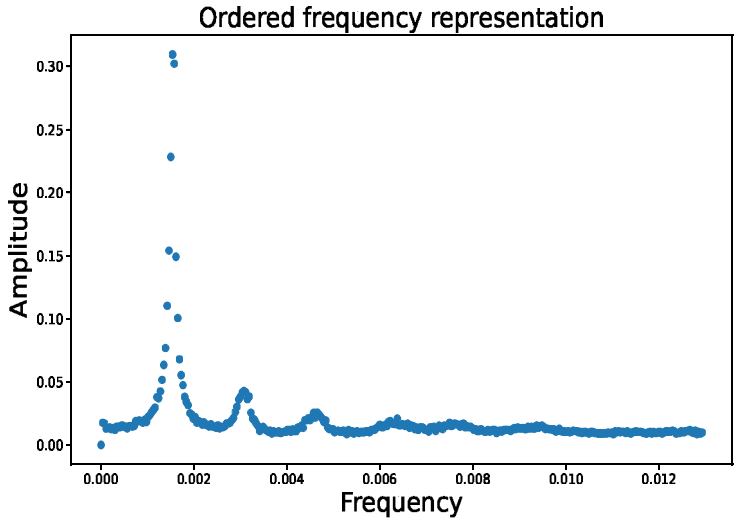


Figure 11: Conventional spectrum representation of a spectrum. Frequencies and their corresponding amplitudes are sorted based on the magnitude of the frequency.

The effect of spectrum representations, 2.1 and 2.2, on novelty detection will be studied by selecting the spectrum representation to be a parameter in the novelty detection pipeline. The next chapter, Chapter 3, introduces the novelty detection pipeline and it will be applied in Chapters 4 and 5.

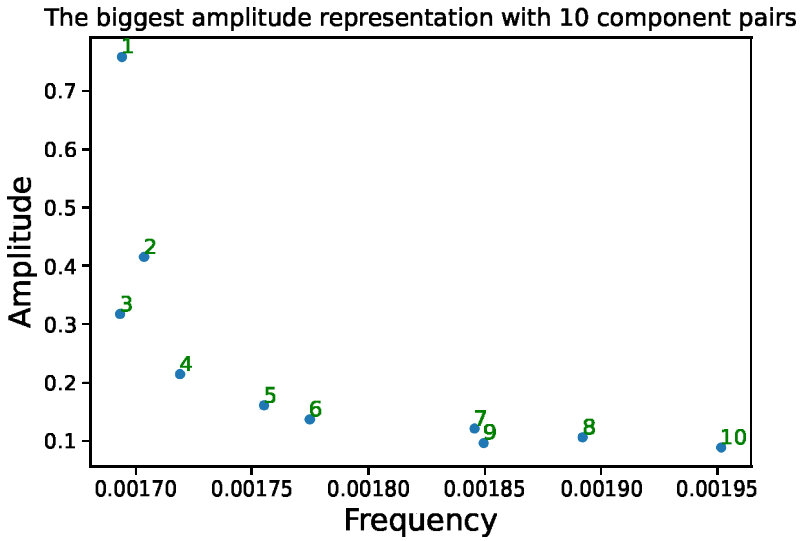


Figure 12: A spectrum represented with amplitude sorted representation with 10 largest component pairs. Amplitudes and their corresponding frequencies are sorted based on the magnitude of the amplitude value and numbers above the datapoints indicate the sorting order from largest to smallest.

3 Novelty detection

This chapter first defines novelty detection and explains its theory. Then, the data analysis process for applying novelty detection is introduced. This process is referred to as the novelty detection pipeline which consists of several data representation and machine-learning steps.

Novelty detection, the task of classifying unseen data as normal or abnormal, can be seen as a one-class classification task [1]. The normal, positive, class is formed by most training data samples. Test samples are classified either as being part of the positive class or not. Test data receives a score called the novelty score which is based on the number of samples predicted to be outside the positive class [1]. Data with the novelty score above the defined threshold is classified as abnormal.

As no abnormal data is required for training the model, novelty detection methods are useful in situations in which data on normal behaviour is available, but the data on abnormal behaviour cannot be acquired [1]. The abnormal behaviour of a system may be too expensive or dangerous to acquire, and the simulation of the abnormal behaviour might be impossible due to this behaviour being unknown, such as in medical diagnostics, industrial failure detection, credit card fraud detection, and astronomy catalogue mapping [1].

It is important to note that sometimes the terms novelty detection and outlier detection are used interchangeably [1]. Both novelty detection and outlier detection are considered as anomaly detection but have slightly different goals [8]. Outlier detection is used in situations in which the goal is to remove outlier data points from training data [8]. An outlier is defined as a single point or a small group of points that appear to be inconsistent with the rest of the data. For example, in small training datasets, outliers can lead to overfitting, thus weakening the performance of a machine-learning model. Removing some of these inconsistent data points may be justified in order to enhance the model performance. In novelty detection, some training data points may be left outside the normal class as outliers, but the focus is on recognizing abnormalities from test data, not removing the outliers from training data [8]. Briefly, novelty detection searches for points inconsistent with the training data while outlier detection searches for inconsistent points inside the training data [8]. Although the focus of this thesis is novelty detection, outlier detection is applied in order to increase the accuracy of the novelty detection model.

A review of novelty detection by Marco A.F. Pimentel, David A. Clifton, Lei Clifton and Lionel Tarassenko classify novelty detection techniques into five categories: (i) probabilistic, (ii) distance-based, (iii) reconstruction-based, (iv) domain-based, and (v) information-theoretic techniques [1]. This thesis applies novelty detection techniques from distance-based category. The motivation for first testing distance-based methods rises from their simplicity. It could be seen that there are clear differences in spectra, for example, as shown in Figure 4, that could possibly be captured with distance calculations. Distance -based novelty detection includes clustering and the nearest neighbour methods [1].

he novelty detection techniques are one-class classification techniques. To achieve a successful classification, several pre-steps, such as representing and cleaning data, must

be carefully performed [2]. The pre-steps combined with the one-class classification step form a novelty detection pipeline. This thesis has developed a distance-based novelty detection pipeline for temperature oscillation data which will be presented in the next section.

3.1 Novelty detection pipeline

The novelty detection pipeline consists of data representation steps and a classification step. The pipeline is applied to recognize the malfunctioning of industrial pipes. The first step of the novelty detection pipeline is to represent the data. This thesis examines two representations, the more common conventional spectrum representation and the less used amplitude sorted representation. The effect of dimension reduction in amplitude sorted representation is also tested. The justification for performing dimension reduction arose by inspecting the images of spectra (Figure 4): the differences in the frequencies of the largest amplitudes are more apparent than with the smallest components.

After selecting the representations of spectra, they will be normalized to vectors of unit length. Normalized vectors have a useful property, the dot product of normalized vectors directly tells the cosine of the angle between these vectors. This property is used in this thesis.

After normalizing the data, a median spectrum will be extracted. The median for every amplitude and frequency value is computed from spectra from time periods in which the behaviour of the machine is considered normal. This median spectrum is machine-specific and considered as the template to which other future spectra will be compared. Figures 11 and 12 in Section 2.2 show the median templates for both conventional spectrum representation and amplitude sorted representation, respectively. The median spectrum was chosen as it is robust to outliers and outliers are known to occur in the data (Figure 3).

All spectra, normal and abnormal, will likely differ from the median template but the hypothesis is that malfunctioning spectra differ more from the template than normal spectra as visualized in Figure 13. The difference estimation of a spectrum and the median template is performed by calculating the distance between every Ai and Fi of the spectrum and the corresponding Ai and Fi values in the template. Three distance functions are tested in this thesis: Euclidean, Manhattan and Cosine angle distance. Euclidean distance, also known as L2 norm, between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\sqrt{\sum_i (x_i - y_i)^2}. \tag{1}$$

Manhattan distance, also known as L1 norm, between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\sum_i |x_i - y_i|. \tag{2}$$

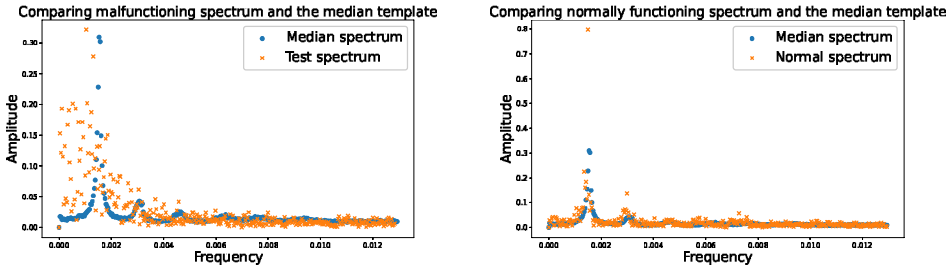
Cosine of the angle a of two-unit length vectors $\mathbf{x}$ and $\mathbf{y}$ is fully determined by their dot product:

$$\cos(a) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||||\mathbf{y}||} \tag{3}$$

Where

$$||\mathbf{x}|| = \sqrt{x_1^2 + ... + x_n^2} \tag{4}$$

represents the Euclidean norm of vector $\mathbf{x}$.



(a) (b)

Figure 13: The hypothesis is that malfunctioning spectra differ more from the median template spectra than normal spectra.

As mentioned, all spectra, normal and abnormal, will likely differ from the median template but the hypothesis is that malfunctioning spectra differ more from the template than normal spectra. In other words, some difference between spectra and the template is accepted but an error threshold separating accepted and unaccepted differences is determined. The next step of the model after extracting the median template is to learn the error threshold for the distances between spectra and the template. The error threshold is defined to be the maximum value of the distances between training spectra and the template (the minimum value in case of cosine angle distance) after outlier removal. Minimum value instead of maximum value is used with cosine angle distance since smaller values indicate longer distances.

Outlier removal before error threshold determination is performed in order to prevent inconsistent points for defining the error threshold. Two machine-learning methods for outlier removal are tested: Local Outlier Factor and DBSCAN. The theory of these two methods will be discussed in Subsections 3.1.1 and 3.1.2.

After learning the error threshold for the difference between spectra and the template, testing may be conducted. Test data will be classified to be normal (1) or abnormal (-1) by calculating the distance of a test data point to the median template, and comparing this test distance to the error threshold. Points outside the error threshold will be classified as abnormal. Table 1 summarizes the steps of the novelty detection pipeline.

The pipeline applies two machine-learning methods, Local outlier factor and DBSCAN. The theories of these methods are introduced in Subsections 3.1.1 and 3.1.2. After the theory, the novelty detection pipeline is applied in Chapters 4 and 5 with two interests of

1. learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction and

2. differentiating the temperature behaviour of unsimilar pipes.

Chapter 4 applies Local Outlier Factor for determining the error threshold in part 5 of the pipeline while Chapter 5 uses DBSCAN for this task. Chapter 4 also studies the effect of different data representations on the accuracy of the model. Both conventional spectrum representation and amplitude sorted representation. are tested. Chapter 5 uses only the most accurate representation found in Chapter 4 and focuses on improving the accuracy by tuning the parameters of DBSCAN.

Pipeline for novelty detection

1. **Data collection**

   - Training data for extracting the median template
   - Training data (not used in creating the template) for learning the error threshold for distances between spectra and the median template
   - Test data (normal and malfunctioning, if available) to test the performance of the model.

2. **Data representation**

   Represent frequency spectrum data with conventional spectrum representation or amplitude sorted representation

3. **Data normalization**

   Normalize data such that every spectrum has unit length

4. **Extracting the median template**

   Use part of the training data to calculate median value for every amplitude and frequency value. Let this be referred to as median template.

5. **Determining the error threshold**

   Use rest of the training data to define the error threshold for the distances between spectra and the median template.

   The error threshold is defined to be the maximum distance value of the training distances between spectra and the template (the minimum value in case of cosine angle distance) after outlier removal.

   This thesis tests three functions to compute the distance between a spectrum and the median template: Euclidean 1, Manhattan 2 and cosine angle 3. Outlier removal is tested with two methods: Local Outlier Factor and DBSCAN.

6. **Testing**

   Calculate the distances between the test data and the median template. Test data points outside the error threshold will be classified as abnormal.

Table 1: Summary of the novelty detection pipeline of this thesis.

3.1.1 The theory of Local outlier factor

This subsection explains the theory behind Local Outlier Factor (LOF) applied in the novelty detection pipeline. In the pipeline, LOF is used to remove outlier data points from the distances between training spectra and the template. Outlier removal is an important step of the determination of the error threshold of the novelty detection pipeline since it prevents inconsistent points for determining the threshold.

LOF compares the density of a point to densities of its k-nearest neighbours [3]. If a point has considerably lower density to its neighbours, it will be classified as an outlier. LOF assigns a Local outlier factor score for every point and points belonging to a local cluster usually have a score around 1 [3].

In order to define the Local outlier factor score of an object, four other concepts are needed. A word object in the definitions refers to a multidimensional vector in space D in the context of this thesis.

First, k-distance of an object p is defined.

Definition 3.1. K-distance of an object p [3]

K-distance of an object p, *k-distance(p)*, refers to the distance $d(p, o)$ between objects p and $o \in D$ that satisfies

1. at least k objects $o' \in D \setminus \{p\}$ have a property of $d(p, o') \leq d(p, o)$ and
2. at most k-1 objects $o' \in D \setminus \{p\}$ have a property of $d(p, o') < d(p, o)$

, in which k is a positive integer.

In other words, k-distance of an object p is the longest distance p has to its k-nearest neighbours.

With the help of k-distance, k-distance neighbourhood of an object p may be defined.

Definition 3.2. K-distance neighbourhood of an object p [3]

$$N_k(p) = \{q \in D \setminus \{p\} | d(p, q) \leq k\text{-}distance(p)\}. \tag{5}$$

K-distance neighbourhood of an object p contains all objects not further from p than k-distance and may include more than k object if there are multiple objects k-distance away from p.

Third definition is referred to as the reachability distance.

Definition 3.3. Reachability distance [3]

Reachability distance of an object p with respect to an object o refers to $reach\text{-}dist_k(p, o) = max\{k\text{-}distance(o), d(p, o)\}$.

If p and o are further from each other than o is from its k-nearest neighbour, reachability distance of p with respect to o is the real distance between the objects. If p and o are closer to each other than o is from its k-nearest neighbour, reachability distance of p with respect to o is the distance o has to its k-nearest neighbour.

Fourth formula required to define Local outlier factor is called local reachability density.

Definition 3.4. Local reachability density of an object p [3]

$$lrd_k(p) = 1/(\frac{\sum_{o \in N_k(p)} reach\text{-}dist_k(p,o)}{|N_k(p)|}) \tag{6}$$

Local reachability density is the inverse of the average reachability distance of an object p.

Finally, the local outlier factor is defined as in 3.5.

Definition 3.5. Local outlier factor of an object p [3]

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$$

Local outlier factor describes the average ratio of the local reachability densities of the k-nearest neighbours of a point p and the point p itself [3]. A point considered as normal have reachability density similar to its neighbours. Points with similar reachability density to their neighbours score around 1 [3]. In other words, local outlier factor compares the density of an object to the densities of its k-nearest neighbours and classifies objects with lower density compared to their neighbours as outliers.

Breuning, Kriegel, Ng and Sander suggest having k, the number of neighbours, more than 10 to remove statistical fluctuations in the results [3]. Statistical fluctuations refer to the differences in LOF scores within points in the same neighbourhood. The higher the k the similar the LOF scores of points within the same neighbourhood [3].

Next subsection introduces DBSCAN which is the other alternative for outlier removal applied in the the novelty detection pipeline of this thesis.

3.1.2 The theory of DBSCAN

DBSCAN is a density-based clustering algorithm that recognises clusters with varying shapes [5]. DBSCAN classifies points to belong into a cluster if they locate in an area of defined minimum density [5]. Other points will be classified as outliers.

Four concepts are needed to define clusters and recognise outliers with DBSCAN. First, the ε-neighbourhood, $N_\varepsilon(p)$, of an object p in object space D is defined in 3.6.

Definition 3.6. ε-neighbourhood [5]

$$N_\varepsilon(p) = \{q \in D | dist(p,q) \leq \varepsilon\}. \tag{7}$$

ε-neighbourhood contains all objects not further than ε away from p, including p itself.

ε-neighbourhood is not enough to determine a cluster because border points of a cluster would not be found with ε-neighbourhood [5]. Border points have smaller ε-neighbourhood than points in the middle of the cluster referred to as core points.

However, border points are in the neighbourhood of core points, meaning that border points locate closer to core points than outliers [5]. A point being in a neighbourhood of a core point is said to be directly density-reachable from core point.

Definition 3.7. directly density-reachable [5]
Object p is directly density-reachable from object q with regards to parameters ε and k if

1. $p \in N_\varepsilon(q)$ and
2. $|N_\varepsilon(q)| \geq k$

A border point can be directly density-reachable from a core point but two border points of a same cluster cannot be directly density-reachable [5]. The definitions of density-reachability and density-connectivity are required to describe the relation of two border points in a cluster.

Definition 3.8. density-reachable [5]
Object p is density-reachable from object q with regards to parameters ε and k if there is a chain of points $p_1, ..., p_n$, $p_1 = q, p_n = p$ in which p_{i+1} is directly density-reachable from p_i.

Definition 3.9. density-connected [5]
Object p is density-connected to an object q with regards to parameters ε and k if
there is a point o from which both p and q are density-reachable.

All points in a cluster, border points and core points, are density-connected [5]. A cluster can be determined with density-reachability and density-connectivity.

Definition 3.10. cluster [5]
A cluster C with regards to parameters ε and k includes objects from space D that satisfy

1. $p \in C$ and q is density-reachable from p, then $q \in C$
2. $p, q \in C$ are density connected

All points not belonging to any cluster are considered outliers [5]. Briefly, a cluster includes core points whose neighbourhood with radius ε contains at least k points and border points directly or indirectly in the neighbourhood of core points.

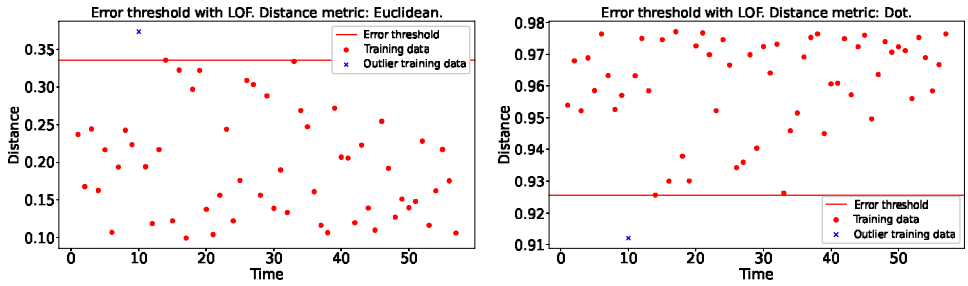
This thesis uses DBSCAN to define the error threshold for distances between spectra and the median template spectrum. The determination of the error threshold is step 5 of the novelty detection pipeline. DBSCAN learns the cluster of normal training distances and removes outliers. After outliers have been removed, the error threshold is determined by the maximum value (minimum value with cosine angle distance) in training data. Test points outside the threshold will be classified as abnormal.

4 Novelty detection with Local Outlier Factor

This chapter applies the novelty detection pipeline with Local Outlier Factor (LOF) as the method for determining the error threshold. LOF removes outlier data points from the distances between training spectra which prevents inconsistent points for determining the threshold. An unseen spectrum will be classified as abnormal if its distance is outside the error threshold.

LOF analysis of this thesis was performed with Scikit-Learn `LocalOutlierFactor` class [4]. `LocalOutlierFactor` classifies points with the negative of the Local outlier factor score smaller than -1.5 as outliers [9]. $N_{neighbour}$ parameter of `LocalOutlierFactor` was set to default 20. Breuning, Kriegel, Ng and Sander suggest having number of neighbours more than 10 to remove statistical fluctuations in the results [3].

Figure 14 visualizes the error threshold determined from the training data by `LocalOutlierFactor` with distances between spectra and the template being (a) Euclidean distances and (b) cosine angle distances.



(a) (b)

Figure 14: Error threshold for the distances between the median template and training spectra determined with LOF. Distances are (a) Euclidean and (b) Cosine angle distances. Datapoints are training points classified as normal (1) or as outlier (-1) and the threshold is set to the value of (a) the maximum point and (b) the minimum point after outliers have been removed. Any points (a) above and (b) below the threshold will be classified as abnormal.

Sections 4.1 and 4.2 apply the novelty detection pipeline with LOF to the two interests of this thesis:

1. Learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction.

2. Differentiating the temperature behaviour of unsimilar pipes.

The performance of the pipeline in the first task is measured with the ability of the model to differentiate abnormal spectra of Pipe 3 (Figure 6) from normal spectra of Pipe 3 (Figure 5) and classify normal spectra of Pipe 3 as normal. The performance in the second task is measured with the ability of the model to differentiate normal spectra of Pipe 3 (Figure 5) from spectra of Pipe 1,2, and 7 (Figures 7, 8 and 9 respectively). The analysis is performed with both conventional spectrum representation and amplitude sorted representation of the spectra.

4.1 Novelty detection with conventional spectrum representation and Local Outlier Factor

This section presents the results of applying the novelty detection pipeline with data represented with conventional spectrum representation in step 2 of the pipeline and LOF used to determine the error threshold for distances between spectra and the template in step 5 of the pipeline.

As a summary, the novelty detection pipeline includes following steps

1. **Data collection**
2. **Data representation**
3. **Data normalization**
4. **Extracting the median template**
5. **Determining the error threshold**
6. **Testing**

In data representation step, data is represented with conventional spectrum representation.

In the median template extraction step, median template was computed from 94 normal frequency spectra from Pipe 3. The spectra are subsequent, overlapping and from five-hour long time windows. Figure 11 in Section 2.2 shows the median template of the 94 training spectra.

After extracting the median template, 57 new normal spectra of Pipe 3 not used in creating the template were used to determine the error threshold for distances between spectra and the template. Three distance functions for computing the distance between the mediam template and spectra were tested: Euclidean, Manhattan and Cosine angle distance.

After training the model, testing was conducted with 24 malfunctioning spectra from Pipe 3, 27 normal spectra of Pipe 3, 62 spectra of Pipe 1, 62 spectra of Pipe 2 and 62 spectra of Pipe 7. All spectra with distance value outside the error threshold were classified as abnormal.

Results show that only Manhattan distance function applied in distances between the template and spectra is able to recognise malfunctioning spectra of Pipe 3 from normal spectra of Pipe 3 (Table 2) Manhattan distance function is also the most accurate in recognizing normal data of Pipe 3 with 77.8 % accuracy (Table 2). However, Manhattan distance fails in the second task of differentiating pipes 1,2 and 7 from Pipe 3. While Euclidean distance and and cosine angle distance reach to 100% accuracies in differentiating Pipe 2 from Pipe 3, Manhattan distance has an accuracy below 2% (Table 3). The overall accuracy for differentiating pipes 1,2 and 7 from Pipe 3 reaches its maximum of 85.1% with cosine angle distance (Table 4).

Results indicate that novelty detection model with conventional spectrum representation and LOF classification has a poor overall performance with a single distance function. Not a single distance function performs well in both two tasks of

Pipe 3 data type	Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle
Malfunctioning	0.0	**1.0**	0.167
Normal	0.593	**0.778**	0.444

Table 2: Accuracy for recognizing malfunctioning and normal data of Pipe 3.

Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle
1.0	0.016	**1.0**

Table 3: Accuracy for differentiating spectra from Pipe 2 from spectra of Pipe 3.

1. learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction and
2. differentiating the temperature behaviour of unsimilar pipes.

Manhattan distance performs best in the first task but fails in the second while Euclidian and cosine angle exhibit the opposite behaviour.

The next section studies whether the novelty detection pipeline with LOF performs better with amplitude sorted representation of the spectra.

Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle
0.797	0.588	**0.851**

Table 4: Overall accuracy for differentiating spectra from Pipe 2,1 and 3 from spectra of Pipe 3.

4.2 Novelty detection with Amplitude sorted representation and Local Outlier Factor

This section discusses the results of applying the novelty detection pipeline with data represented with amplitude sorted representation in part 2 of the pipeline and LOF applied to determine the error threshold in part 5.

The effect of dimension reduction on the Amplitude sorted spectra is also studied. The dimensions are reduced by selecting less amplitudes and their corresponding frequencies than in the original representation. The number of dimensions of a spectrum is represented as the number of component pairs, which denotes the number of frequency and amplitude pairs. The number of component pairs is computed by dividing the number of dimensions by two. The original number of component pairs is 346.

The novelty detection pipeline includes steps of

1. **Data collection**
2. **Data representation**
3. **Data normalization**
4. **Extracting the median template**
5. **Determining the error threshold**
6. **Testing**

In data representation step, data is represented with amplitude sorted representation.

The same training and test data were used as in the Section 4.1. The median template in step 4 was computed from 94 normal, subsequent and overlapping frequency spectra from Pipe 3.

After extracting the median template, 57 new normal spectra of Pipe 3 not used in creating the template were used to determine the error threshold for distances between spectra and the template. Three distance functions for computing the distance between the mediam template and spectra were tested: Euclidean, Manhattan and Cosine angle distance.

After training the model, testing was conducted with 24 malfunctioning spectra from Pipe 3, 27 normal spectra of Pipe 3, 62 spectra of Pipe 1, 62 spectra of Pipe 2 and 62 spectra of Pipe 7. All spectra with the distance value outside the error threshold were classified as abnormal.

Results show that the pipeline recognises malfunctioning data of Pipe 3 with 100% accuracy with at least 20 component pairs and with all three distance functions (Table 5).

100 % accuracy is reached in recognizing normal data of Pipe 3 with 2 component pairs (Table 6). However, such a small number of component pairs fails to recognise malfunctioning data. In higher dimensions, normal data of Pipe 3 is recognised with

accuracies between 81% and 96% with Euclidean and Manhattan distance while cosine angle distance performs slightly worse (Table 6).

Regarding the second task of differentiating pipes 1,2 and 7 from Pipe 3, Manhattan distance and high dimensions perform the best. Pipe 2 that has the most similar behaviour to Pipe 3 is differentiated with more than 98% accuracy with Manhattan distance and at least 250 component pairs (Table 7). The overall accuracy for differentiating spectra from pipes 1,2 and 7 from spectra of Pipe 3 reaches 99% with Manhattan distance and at least 250 component pairs while other distance functions fail in this second task.

Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle	Number of component pairs
0.0	0.0	0.0	2
1.0	1.0	1.0	20
1.0	1.0	1.0	100
1.0	1.0	1.0	250
1.0	1.0	1.0	346

Table 5: Accuracy for recognizing malfunctioning data of Pipe 3.

Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle	Number of component pairs
1.000	**1.000**	0.963	2
0.963	0.852	0.778	20
0.926	0.889	0.815	100
0.926	0.815	0.815	250
0.852	0.852	0.741	346

Table 6: Accuracy for recognizing normal data of Pipe 3.

Accuracy, Euclidean	Accuracy, Manhattan	Accuracy, Cosine angle	Number of component pairs
0.065	0.065	0.081	2
0.081	0.194	0.145	20
0.081	0.403	0.145	100
0.081	**0.984**	0.161	250
0.145	**0.984**	0.177	346

Table 7: Accuracy for differentiating spectra from Pipe 2 from spectra of Pipe 3.

Pipeline with amplitude sorted representation, LOF, Manhattan distance between the template and spectra and 346 component pairs has the best overall performance.

Accuracy	N largest component pairs	Distance function
0.027	2	Euclidean
0.027	2	Manhattan
0.034	2	Dot
0.378	20	Euclidean
0.642	20	Manhattan
0.507	20	Dot
0.527	100	Euclidean
0.750	100	Manhattan
0.588	100	Dot
0.534	250	Euclidean
0.993	250	Manhattan
0.628	250	Dot
0.588	346	Euclidean
0.993	346	Manhattan
0.635	346	Dot

Table 8: Overall accuracy for differentiating spectra from pipes 2,1 and 3 from spectra of Pipe 3.

Pipeline with LOF, Manhattan and 250 component pairs peforms also reasonably well. Pipelines with other parameters perform well only in recognizing Pipe 3 data or differentiating unsimilar pipes but do not succeed in both tasks. The first task of differentiating malfunctioning data of Pipe 3 is more important but a model that performs well also in differentiating unsimilar pipes is preferred because this type of model indicates better ability to recognise varying abnormal data.

Figure 15 visualizes the predictions for (a) malfunctioning and (b) normal test data from Pipe 3 with Manhattan distance and 346 component pairs. Figure 15 (b) shows that most normal test points classified as abnormal locate less than one distance unit away from the threshold while malfunctioning datapoints in Figure 15 (a) are more than two distance units away from the threshold. There is a clear separation of the malfunctioning data. It should be noted that the test datasets are small, such as 27 normal test spectra from Pipe 3, which causes accuracy to vary considerably even with a few classification mistakes.

Results show that the novelty detection pipeline with amplitude sorted representation and LOF has a better overall performance than the novelty detection pipeline with the conventional spectrum representation and LOF. Regarding the two interests of this thesis,

1. learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction and

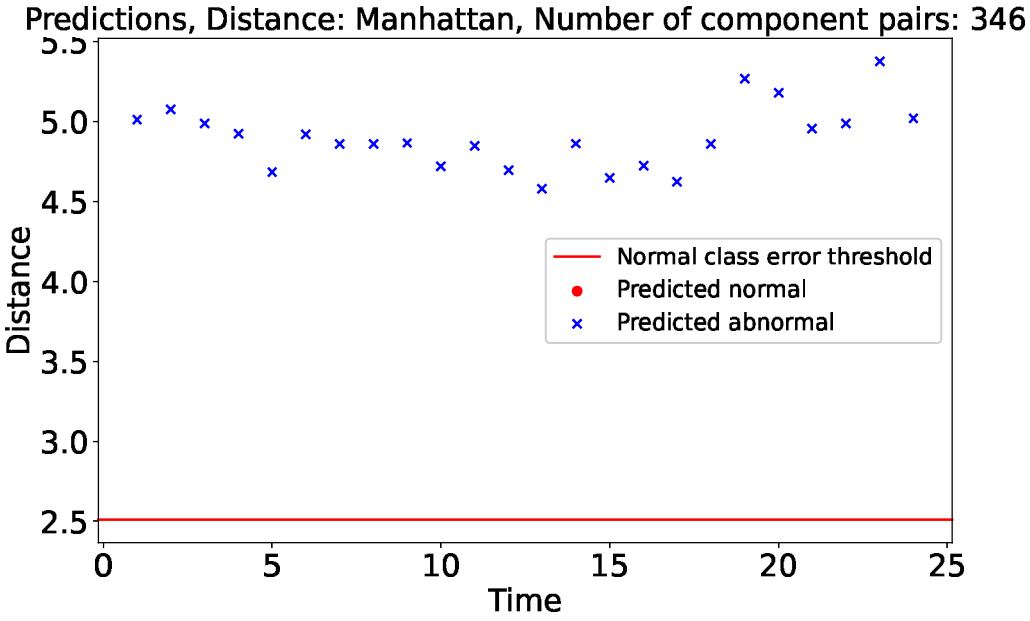
2. differentiating the temperature behaviour of unsimilar pipes,

there is no single distance function to use with the conventional spectrum representation model that would perform well in both tasks. The pipeline with amplitude sorted representation, Manhattan distance function and LOF threshold determination, performs reasonably well in both tasks.

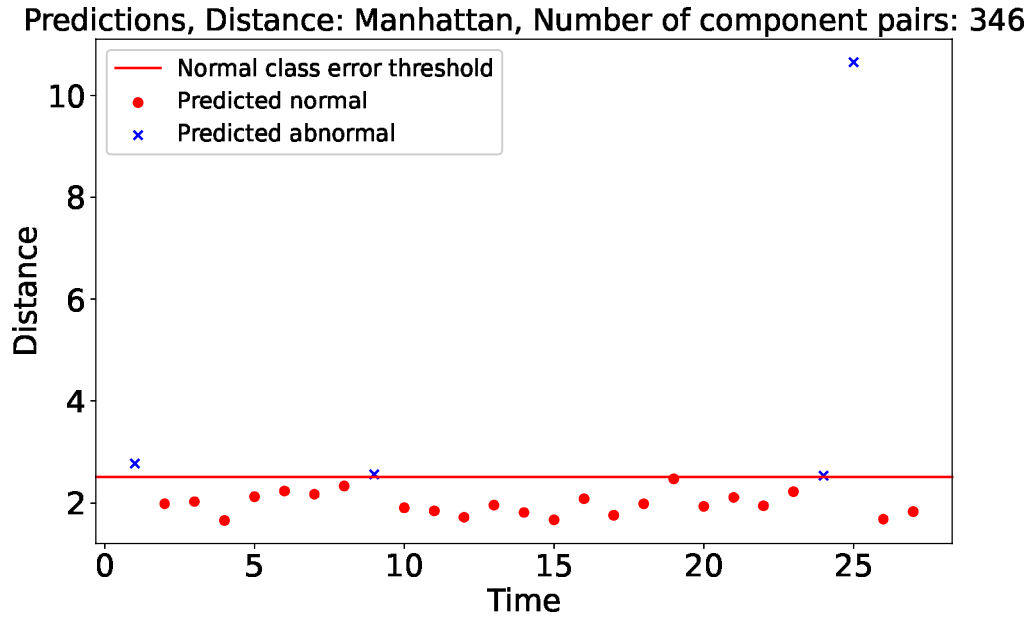
The problem with the best performing model with amplitude sorted representation, Manhattan distance function and LOF threshold determination is that it gives false negatives with normal test data. Most false negatives are training points just above the threshold (Figure 15 b) and a slight increase of the threshold could reduce the number of false negatives. However, Figure 16 (b) shows that even a 0.5 distance unit increase of the threshold leads to many false positives with Pipe 2.

In industrial settings, to reduce the number of false negatives without increasing the number of false positives, the classification of the industrial pipe could be done based on several subsequent classifications of spectra. For example, a pipe could be classified to malfunction only if more than 50% of subsequent spectra from 12-hour time period is classified as abnormal. This approach would reduce the number of false negatives shown in Figure 15 (b) but would not increase the number of false positives shown in Figure 16 (b).

In Chapter 5, novelty detection pipeline will be applied with amplitude sorted representation, Manhattan distance function, 346 component pairs and DBSCAN as the classification method. The interest is to test whether DBSCAN improves the accuracy compared to LOF when other parameters are selected to be the most optimal by far.

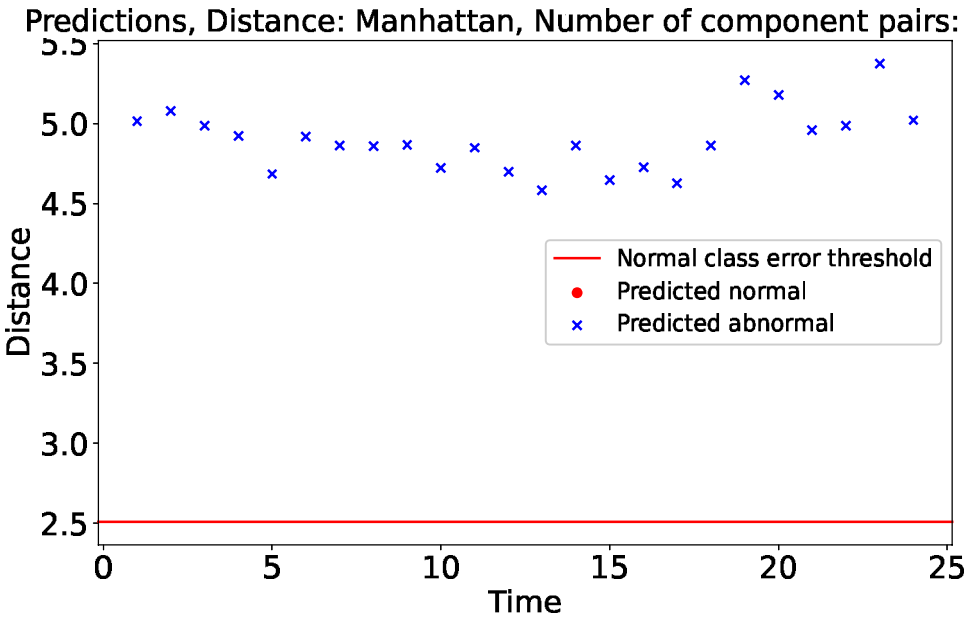


(a)

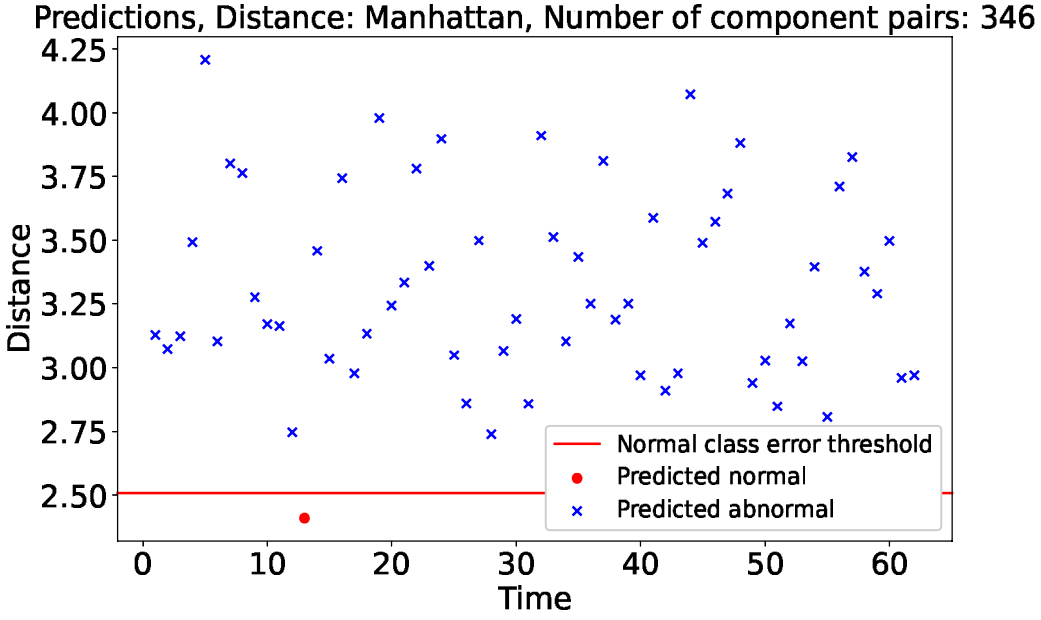


(b)

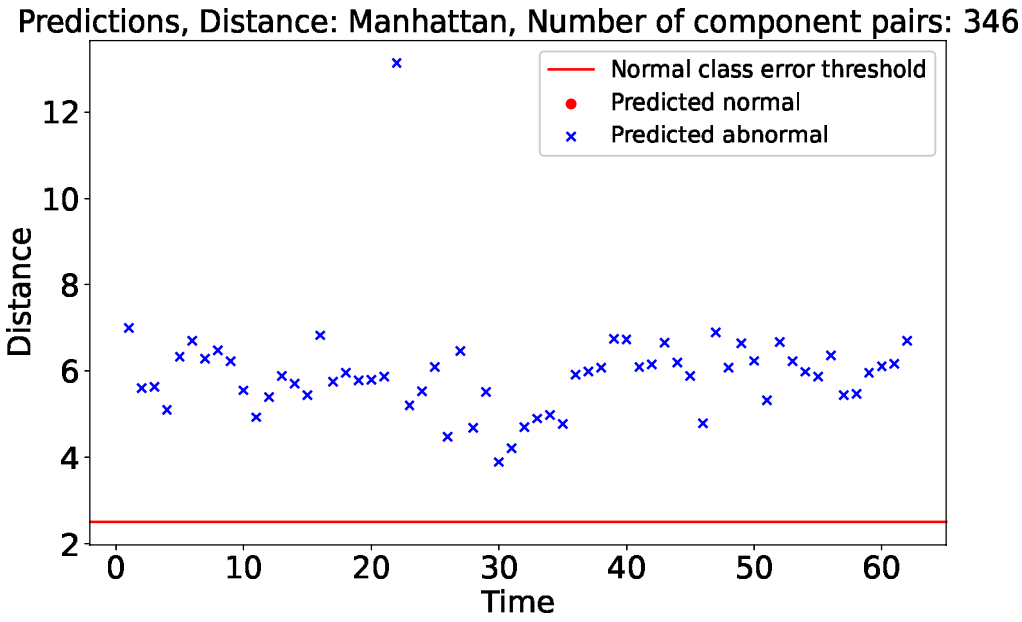
Figure 15: Classification of (a) malfunctioning and (b) normal test data from Pipe 3 with the novelty detection pipeline with amplitude sorted representation, LOF, Manhattan distance function and 346 component pairs.


Predictions, Distance: Manhattan, Number of component pairs: 346
Distance
Time
Normal class error threshold
Predicted normal
Predicted abnormal

(a)


Predictions, Distance: Manhattan, Number of component pairs: 346
Distance
Time
Normal class error threshold
Predicted normal
Predicted abnormal

(b)



(c)

Figure 16: Differentiation of (a) Pipe 1, (b) Pipe 2 and (c) Pipe 7 data from Pipe 3 data with the novelty detection pipeline with amplitude sorted representation, LOF, Manhattan distance and 346 component pairs.

5 Novelty detection with DBSCAN

This section applies the novelty detection pipeline with DBSCAN as the method for as the method for determining the error threshold. The interest is to test whether DBSCAN can outperform LOF studied in Chapter 4 when other parameters are selected to be the most optimal by far: amplitude sorted representation without dimension reduction and Manhattan distance between spectra and the template.

DBSCAN class applied in this thesis is Scikit-Learn `DBSCAN` [4]. Parameters required to adjust for `DBSCAN` are ε and k [10]. The parameter k has a smaller effect on the performance and should take the smallest value possible to reduce computational power [7]. Sander, Ester, Kriegel and Xu suggested k to be set to 2∗dimensions -1 [7]. However, with one dimensional data, parameter k would have value 1. Having k as 1 means one isolated point could form a cluster. In order to increase the amount of points required to form a cluster, first value of k tested will be 3. In other words, at least three points are required to form a cluster.

Ester, Kriegel, Sander, and Xu suggest a heuristic on determining ε based on the k-nearest neighbours of training data points [5]. The value of ε is suggested to be determined from a figure of sorted k-distance values of training data points. K-distance refers to a distance between a point and its kth nearest neighbour. As the value for k is set to 3, the value for ε may be determined from 3-distance plot. Ester, Kriegel, Sander, and Xu suggest searching for valleys from the k-distance plot [5]. Valleys refer to sharp angles in the figure. The k-distance value of an apparent valley should be selected for ε and the value should be small as possible [6]. Figure 17 visualizes the determination of ε with $k = 3$. There is an apparent valley in 0.05 and this is set as the initial value of ε.

`DBSCAN` with parameters $k = 3$ and $\varepsilon = 0.05$ was used to define the normal class from training data. Figure 18 (a) shows the threshold of normal class determined by DBSCAN while (b) shows the threshold determined by LOF (Subsection 4.2) for comparison purpose. The reference LOF model in Figure 18 (b) has had the best overall performance by far. Both thresholds have been determined for one dimensional training data of Manhattan distances. Manhattan distances between spectra and the median template have been computed from spectra data represented by amplitude sorted representation and 346 component pairs. The threshold determined with DBSCAN (Figure 18 a) is equal to the threshold determined with LOF (Figure 18 b) and therefore DBSCAN will not perform better than the best performing model of section 4.

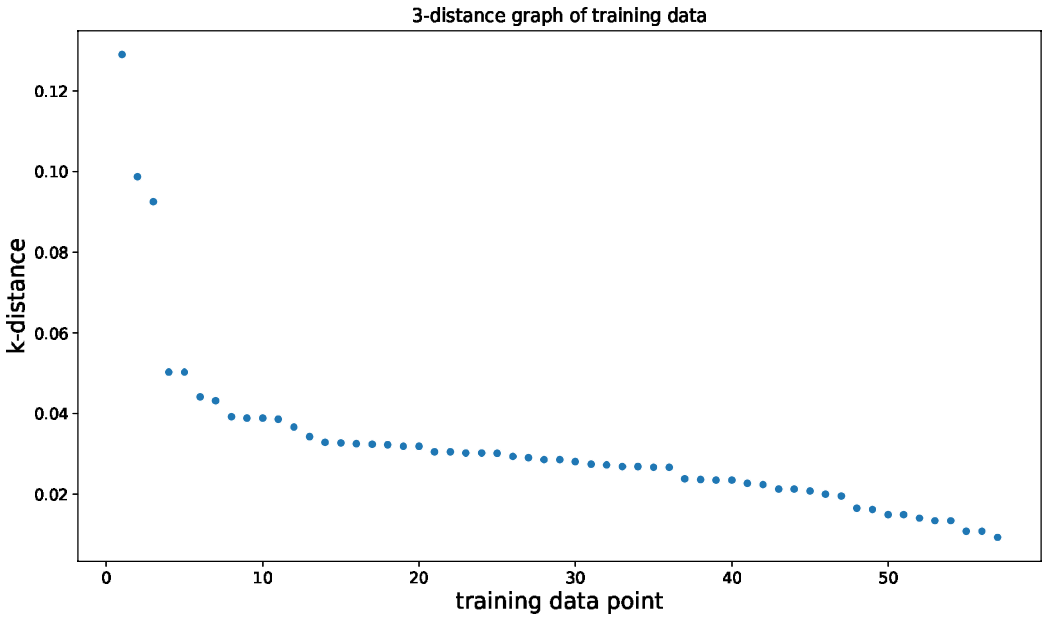
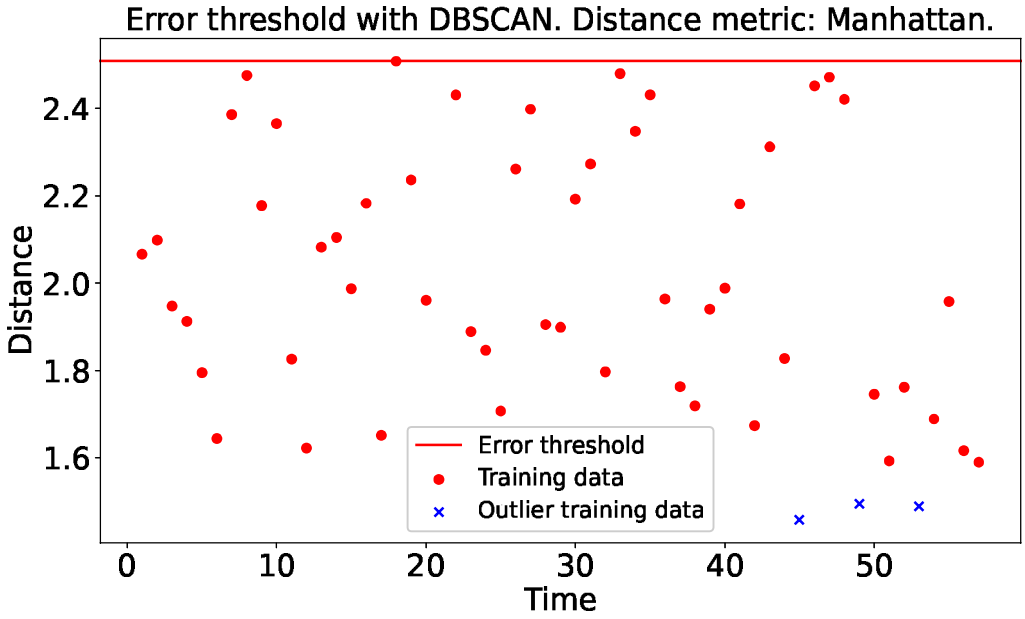
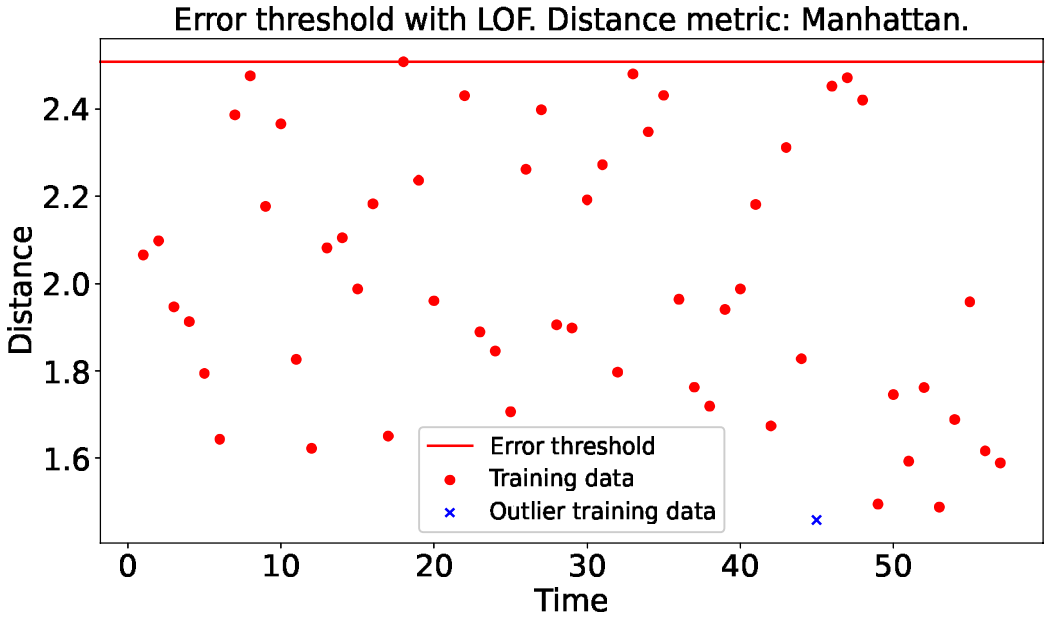


Figure 17: Parameter ε of DBSCAN is suggested to have value of the k-distance in a sharp angle point [5]. Based on the figure, ε could have a value of 0.05.



(a)



(b)

Figure 18: Threshold for the normal class of distances between the median template and training spectra determined with (a) DBSCAN and (b) LOF. Datapoints are training points classified as normal (1) or as outlier (-1) and the threshold is set to the value of the maximum point after outlier removal. Data points with value above the threshold will be classified as abnormal.

6 Conclusions

This thesis analysed temperature data from industrial pipes with two interests of

1. learning a normal temperature behaviour of a pipe and recognizing when this pipe starts to malfunction and

2. differentiating the temperature behaviour of unsimilar pipes.

For solving these two tasks, novelty detection was chosen as the method. Novelty detection models do not require malfunctioning data for training and therefore suit for industrial purposes in which acquiring malfunctioning data is expensive.

Developing a model to recognise malfunctioning is the main goal of this thesis. The second interest of differentiating unsimilar pipes supports the more important first task in situation in which small amount of malfunctioning test data is available. The ability of the model to recognise unsimilar pipes describes the ability of the model to detect abnormal data and therefore suggests the ability of the model to detect malfunctioning.

After brief introduction to the temperature data in time domain (Section 2.1) a conclusion was made that the most reliable novelty detector will be gained by combining time domain analysis with frequency domain analysis. However, due to the simplicity of time domain analysis, the rest of the thesis focused on frequency domain.

In frequency domain, feature engineering was emphasized and two different representations, conventional spectrum representation and amplitude sorted representation, for the data were introduced (Chapter 2). In Chapter 3, a data analysis process for temperature oscillation data, referred to as novelty detection pipeline, was designed. The accuracy of the pipeline with the two data representations was studied in Chapter 4. The pipeline with the amplitude sorted representation was observed to have better overall performance compared to conventional spectrum representation.

The novelty detection pipeline with the best overall performance includes data represented with amplitude sorted representation without dimension reduction in part 2 and LOF or DBSCAN as the method for removing outliers in part 5 of the pipeline. The distance function used to compute distances between spectra and the median template in part 5 of the pipeline is Manhattan distance in the model with the best overall accuracy. The best performing model reaches 100 % accuracy in differentiating malfunctioning spectra of Pipe 3 from normal spectra of Pipe 3 (Table 5), 85% accuracy in recognizing normal spectra of Pipe 3 (Table 6) and 99% accuracy in differentiating spectra from unsimilar pipes (Table 8).

The model with the best overall accuracy gives more false negatives compared to false positives (Subsection 4.2). A solution for this was proposed: the classification of a pipe could be done based on several subsequent classifications of spectra. For example, a pipe could be classified abnormal only if more than 50% of subsequent spectra from 12-hour time period is classified as abnormal. This approach would reduce the number of false negatives shown in Figure 15 (b) but would not increase the number of false positives shown in Figure 16 (b).

Chapter 5 studied whether the accuracy of the best performing model could be improved by changing Local outlier factor classification to DBSCAN. However, it was observed that the performance could not be improved by changing the method that searches the outliers from the normal class due to the definition of the normal class. The normal class is defined as follows: The normal class includes all points with value equal or lower than a threshold. The threshold is the training point with maximum value after removing the outliers. Figure 18 shows that the threshold with the most accurate LOF model and DBSCAN model is the same outermost training data point and cannot be changed by removing outliers.

As a summary, this thesis focused on developing novelty detection method for industrial temperature oscillation data in frequency domain. The most promising model found is the novelty detection pipeline with amplitude sorted representation without dimension reduction, LOF or DBSCAN as the method for removing outliers, and Manhattan distance function used in the distances between spectra and the median template. Further improvement on the accuracy of the model can be achieved by classifying the pipe based on several subsequent spectra. In industrial usage, the frequency domain novelty detection pipeline developed in this thesis should be combined with time domain analysis to detect possible malfunctioning situations not observed in frequency domain. The results are promising but the test dataset is relatively small and therefore, further testing is required to validate the results.

This thesis focused on distance-based novelty detection. If further testing reveals serious flaws in the developed distance-based novelty detection pipeline, reconstruction-based novelty detection methods, such as neural networks, could be tested. However, neural networks can be difficult to train in high dimensions and dimension reduction of the frequency data should be considered [1].

References

[1] Pimentel, A.F.M. & Clifton, A.David & Clifton, Lei & Tarassenko, Lionel. *A review of novelty detection.* Signal Processing, 2014. Vol.99. pp.215–2499.

[2] Zheng, A. & Casari, A. *Feature Engineering for Machine Learning.* O'Reilly Media, Inc, 2018. Chapter 1.

[3] Breunig, MM. & Kriegel, HP. & Ng, RT. & Sander, J. *LOF: identifying density-based local outliers.* Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 16.May 2000. pp. 93-104.

[4] Pedregosa, F. & Varoquaux, G. & Gramfort, A. & Michel, V. & Thirion, B. & Grisel, O. & Blondel, M. & Prettenhofer, P. & Weiss, R. & Dubourg, V. & Vanderplas, J. & Passos, A. & Cournapeau, D. & Brucher, M. & Perrot, M. & Duchesnay, E. *Scikit-learn: Machine Learning in Python.* JMLR, 2011. pp. 2825-2830.

[5] Ester, M. & Kriegel, HP. & Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise.* Kdd, Aug 2 1996. Vol. 96, No. 34, pp. 226-231.

[6] Schubert, E. & Sander, J & Ester, M & Kriegel, HP & Xu, X. *DBSCAN revisited, revisited: why and how you should (still) use DBSCAN.* ACM Transactions on Database Systems (TODS), Jul 31 2017. Vol 42, No. 3, pp. 1-21.

[7] Sander, J & Ester, M & Kriegel, HP & Xu, X. *Density-based clustering in spatial databases: The algorithm gdbscan and its applications.* Data mining and knowledge discovery, Jun 1 1998. Vol 2 No.2, pp. 169-94.

[8] Scikit Learn documentation 2.7. Novelty and Outlier Detection. [Cited 29 June 2020]. Available at: https://scikit-learn.org/stable/modules/outlier_detection.html

[9] Scikit learn LocalOutlierFactor documentation [Cited 7 July 2020]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbours.LocalOutlierFactor.html

[10] Scikit learn DBSCAN documentation [Cited 16 July 2020]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN