Østfold University College

MASTER'S THESIS

LEAK DETECTION OF WATER PIPELINE NETWORKS

WITH ACOUSTIC DATASET ANALYSIS

Mohammad Askari Jirhandeh

Master's degree in applied computer science

School of Computer Sciences
Østfold University College
Halden, Norway
Spring 2021

MASTER'S THESIS

LEAK DETECTION OF WATER PIPELINE NETWORKS

WITH ACOUSTIC DATASET ANALYSIS

Mohammad Askari Jirhandeh

Master's degree in applied computer science

School of Computer Sciences
Østfold University College
Halden, Norway
Spring 2021

Abstract

The detection of the water leakages is significantly important in different underground pipeline networks due to the lack of fresh water in today's life. Fast detection and accurate recognition of the leakages, in monitoring systems, became one of the top researches in the field. Aim of this master thesis is to test and evaluate a set of non-invasive sensors suitable for detection of the leakages to the buried urban water pipelines as well as acoustically comparing the essential data features of this project with other related features in some other datasets in this field.

Our contribution in this study is the methodological approach where we examine some of the machine learning techniques for leak detection, in which decision tree classification methods, apart from neural network approaches, that used for the task shows satisfactory predictive results. On the other hand, the result comparison of different feature selection of the classification methods along with data preprocessing strategies used in different learning approaches for acoustic noise datasets is the second contribution in this project. We compared the attributes of other similar datasets with each other from the acoustic aspect and reported the most important attributes applicable for our studies.

The algorithm results evaluated at the end and the best possible machine learning techniques and attribute collection is discussed in detail for acoustic leak detection of water pipelines.

Keywords: noninvasive sensors, water pipelines, internet of things, LoraWAN, machine learning, acoustic leak detection

Acknowledgements

I would like to thank my patient supervisor, Dr. Maben Rabi for his endless support and continuous positive advices regarding the thesis and towards my feature career.
I would like to thank all my teachers, faculty of Computer science in Høgskole i Østfold for their positive attitude and helpful knowledge to students.
Special thanks to IFE staff, all members of the project and who have been involved in this thesis.
I like to thank my friends who always been by my side, being supportive for a long time.
And at the end, I would like to thank my family and specially Mom, who's been always my angel.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

This chapter outlines the thesis overview. It describes the motivation and the general concept behind the research and the questions which are going to be covered by the work.
It explains the theoretical background and methodology alternatives for evaluation and performing the thesis experiments. It will briefly inform the thesis structure at the end.

1.1 Background and related work

1.1.1 Motivation

There is a large amount of water loss in the world each year due to leaky pipelines. This volume of wastage is one of the main causes of the water crises while our planet is running out of the fresh water therefore, the number of people struggling for water resources is getting increased.
There is a tremendous population of the world have no proper access to safe drinking water and it's increasing by every passing year and by 2025, 2/3 of this population will be living in lack of water condition[1].
Lack of water resources could threaten agricultural crops, infrastructure and even humans' life. Risking human's life, from one hand, and the rupture of the huge pipes and incidence of catastrophic wars due to the water crises from the other hand, results devastation.
Recent researches show that, each average person consumes around 135 liters of fresh domestic water per day which implies that the consumption is rising with the population growth [2]. Therefore, these mentioned statistics highlight the requirement of the water pipeline monitoring systems.
Although transportation of the water in all cities around the world is through the tubes and pipelines, there is a possibility of damage, failure, or destruction of the pipes over

time due to many circumstances surrounding us. Natural hazards or human faults in different situations could lead to the leakages in the water pipelines.

Apart from all other researches, our planet is suffering from lack of water resources due to climate changes and this problem is becoming irrecoverable catastrophic issues in different continents from many aspects. There are different errors in water supply in some other continents with enough water resources which may go undetected for a long time until they turn into huge rupture and cost cities millions of dollars[3].

Lack of water resources becomes one of the significant issues in different countries around the world and turns out to be the challenging topic for the crisis management, hence the detection of the water leakages with precise scale is significantly important.

Many questions come to the picture at the time of facing the failure. where are the exact points of the leakages? How's detection could be developed in terms of seriousness of the failures? To solve the problem by maintenance staff, how long is it take to verify the failure?

In 2016, Statistics implies that, around 99.5 percent[1] of the residents in Norway are connected to the municipal water supply system which is the safe supplied drinking water from many aspects. It's obvious that the massive infrastructure behind it to supply the water for such a huge demand, needs smart maintenance.

In 2016, about 3800 leak repairs in corresponding infrastructure were reported. So, the first solution comes to mind is that how to minimize the failure? How does pipeline replacement work? So, the quality of the pipes must be maintained with the minimum requirement before applying any leak detection strategy.

Its estimated to take 145 years[2] to replace and reach the satisfactory quality level for all the available pipelines, said statistics Norway.

Although the pipeline standardization process is a long-time plan and currently it's on progress, but there is an obvious need for leak detection strategy to be developed parallelly.

---

[1] Statistics Norway 2016

[2] https://www.fhi.no/en/op/hin/infectious-diseases/drinking-water-in-Norway/

The topic has become one of the priority projects in most of the municipalities in Norway to search deeper about the best possible solutions according to their corresponding underground piping infrastructure. So, Viken county and related different municipalities with respect to their water piping infrastructure are involved in this investigation as well.
Still drinking water in some cities in the county is transferred under old pipes from many years back, and the leak detection of the pipelines is discovered manually through the human experienced operators for each specific case, reported by the municipality.
Investigation on using different sensors to detect the leakages was the first proposed solution by the research institute and approved by all involved parties while acoustic sensors nominated to be the priority candidate according to that research.
Acoustic sensors seem to be promising model in different leak detection scenarios[4]. On the other hand, the combination of acoustic sensors and machine learning algorithms with the help of different patterns on captured sound signals, proved to be reliable alternative to identify and locate leakages in different pipelines[5][6].

1.2 Exploring research

In order to get a better insight of the scenario we started a study to explore the topic with same simulation projects, since we didn't have any information to start with. Being aware of different simulation scenarios about the acoustic sensor performance in detection of the water leakages could give us better intuition of the project.
Information gathering about the project which we had to start, we overviewed some interesting projects with similar methods from the sensor installation perspective. In [7] the sound propagation strategy is used throughout the pipeline with the predefined measured distances between the acoustic sensors, that the likely leakages can be detected through the distance calculation by the acoustic sound speed intensity.
Another acoustic sound propagation simulation technique is operated [8] inside the laboratory to locate the pipeline leakages with respect to the fluid as the noise spreading material. It stated that, the changes of acoustic noise equal to 10dB at the leak location can

be detected almost from 6 inches (15,24 cm) distance from the leak point itself. The experiment is conducted with simulation tools.

1.3 Problem statement

In mentioned water leakage detection project, the whole scenario divided into three parts from the general perspective. First part, Implementation and installation of the proper sensor for leak detection, then data collection for a long time with a help of wireless antennas on LoRaWAN technology[3] is the second part of the project and third part is the analysis of the received data in order to detect the leakages.

The mentioned classification shows that, from the research aspect these parts are almost three separated sections, joint together in the form of a project. In our contribution in this thesis we have gained required knowledge by conducting a research study in the first section and then we have completed an investigation in last section by analysis the data received from the project and comparing the result with few similar acoustic cases relevant to our research.

So, in data analysis part we first apply different machine learning algorithms in acoustic water leak detection environment. Secondly, we try to identify the best possible package solution in terms of the acoustic sensor and relevant employed algorithms in detection of the leakages in urban water pipelines.

In defined project, we try to propose best possible method for leak detection and consequently efficient attribute collection with respect to different datasets.

[3] https://lora-alliance.org/about-lorawan/

1.4 Research questions

Basically, freshwater is crucial for human health. So, maintaining the efficient water distribution system is essential for our survival. Detection of the leakages in an accurate way using any known method is a difficult task hence it should be measured precisely for better efficient output.

The experiments and algorithms applied in this study aim to answer following questions:

1) What are the appropriate machine learning methods in acoustic leak detection? (How to study?)
2) Which attributes are playing important role in detection of the leakages in urban water pipeline data analysis? (what to study?)

1.5 Required background study

This section is a briefly explanation of the main technical concepts of the study. Before start, in order to have a better understanding of the scenario and to have a better view of the related technical concepts into our project, we need to overview some machine learning terms.

We introduce the preprocessing method in analysis of acoustic sound data and then we focus on different machine learning concepts and general overview in anomaly detections.

1.5.1 Acoustic sound classification of scenes and events

Basically, there was a growing demand in machinery fault diagnosis with different approaches reported by [9]. The promising approaches are claimed to be pressure and temperature sensor-based [10] [11], vibration sensor-based [12] and sound detection, acoustic sound detection is one of the methods in monitoring fault diagnoses.

Applied feature extraction approaches in machinery fault detection through vibration data, is another traditional method of fault diagnostic scenario. The performance is

improved by applying deep learning to learn features from vibration data and modified diagnosis performance through classification [12].

1.5.2 Acoustic sound detection and deep learning

After machine learning development in last two decades, deep learning approaches played an important role in acoustic sound detection and classification. To build an appropriate model from the training dataset and applying different strategies to decrease the errors which led on with reliable results in this field.

Applying Fourier transform algorithm to convert the signals into frequency representation of continuous time signals is a common method of researchers and commercial products [13]. The article shows that using ANN, artificial neural network to build a model to find the location of the leakages with a high number of accuracies is another efficient instance using machine learning techniques on acoustic sound datasets.

1.5.3 Norway water distribution system, Viken-Halden municipality

There are 700 million[4] cubic meters of drinking water delivered throughout a year in Norway water services. The water consumption for each person estimated as 200 litres daily [14].

In Norway, the open water is controlled and managed at the state level and then the distribution is handled by the water workers often at municipality level, and at the end of the line drinking water is managed by the end users.

Almost 1600 waterworks which supply water, covers 90% of the population of the country and the other 10% of the population use private wells. Since the land has the enough water resources, still the majority of the water supply in Norway is based on surface water unless there is a need to use the ground water in case of geographical boundaries.

Out of all supplied water, households use 41% of the water production, 2% for the cabins for holiday periods and 25% goes to the industry. Researches show that approximately,

[4] https://www.norskvann.no/images/torilh/The_water_services_in_Norway_endelig.pdf

32-34% of the remaining drinking water produced is lost due to the leakages and line disruptions of the water distribution system which is a highly considerable amount of water.

For a safer water distribution in the state, Halden municipality is also involved with several projects for developing, testing, and implementing different solutions, as artificial intelligence and IT solutions shows promising output in monitoring and maintenance strategies.

Nordic innovation[5] is divided the system solution to the challenges in into two different parts in Halden city. The first part is advanced monitoring at reservoir with a water quality verification unit and second part, is the smart fiscal flow units to measure, water flow rate, temperature, absolute pressure, turbidity, return stop valve user and acoustic sensor for leakage positioning.

1.5.4 Machine learning concepts

As most of the concepts and techniques have been used all over this thesis is based on machine learning concepts, we briefly explain some of the important phrases and technical concepts of the workflow.

Basically, every machine learning approach with its corresponding dataset requires some specific steps as follows:

- Pre-processing of the dataset
- Suitable division of dataset into training and testing sections
- Training the model from divided training dataset
- Predicting the target values from the build model
- Techniques to calculate & evaluate the target value
- Optimization techniques for model improvement
- Model comparisons among different algorithms

[5] https://www.nordicinnovation.org/

- Visualization of the desire result

Each of the mentioned steps above, includes many techniques and method strategies which must be chosen corresponding to our dataset type and the goal of the project in that specific boundary. So, for better understanding of the following concepts, we will explain the techniques which has been applied in this thesis.

1.5.4.1 Pre-processing of dataset

The very first step to start the machine learning approaches is processing the dataset in most desired form correspond to our problem.

Clearly, preprocessing steps impacts the accuracy of the machine leaning algorithms and significantly improves the accuracy. An the experiment which is conducted on big data before and after the preprocessing techniques in [15] can positively approve the claim.

1.5.4.2 Handling null values

There are always some null values available in majority of datasets. The datasets retrieved from a real-world scenario usually comes with some null values which is not understandable for the machines.
One of the solutions to handle this situation is to remove the rows and columns included with the null value. Usually this will happen when we combine two different parts of the available datasets to make a larger meaningful dataset for our project.
There are some other methods available to handle the situation like imputation of missing values, but it depends to our datasets if the number of the null values are not negligible in that dataset.

1.5.4.3 Encoding

When we have some categorical values in our dataset, we must encode it to numbers and the numerical values which is understandable for the machines before we fit and evaluate the model. There are several techniques available like integer encoding and one hot

encoding. In what follows we explain them briefly as we have used them alternatively with respect to our datasets.

Integer encoding is when each label mapped to an integer, so the number of integers could be as many as required for that specific labels. Usually it happens when we deal with ordinal categorical data with ordered additional information.

One hot encoding is when each label mapped to a binary digit. It usually happens when we deal with limited group of labels in our categorical dataset.

However these are not the only techniques available for encoding as there are some experiments evaluated the comparison and accuracy of different encoding techniques [16] but it depends to the nature of corresponding dataset.

1.5.4.4 Normalization

The scaling techniques that the values are shifted between 0 and 1 is called normalization. One of the normalization subsets is also called as Min-Max scaling[6] which changes the values of numeric columns in the dataset without distorting differences in the ranges of values.

However, the investigation on several normalization methods from different data preprocessing research areas on normalization impact to improve the classification performance shows that, normalized data supports the outcome in terms of better predictions on classification problems[17].

It also believes that; the mean and standard deviation measures are more important and suitable for normalization in compare with Min-Max and median measures. It's obvious that we can add some features to our dataset by calculating the mean, median, Min-Max, and standard deviation from the data points of our dataset. These mentioned important points regarding data normalization will be discussed explicitly later in coming chapters.

On the other hand, some studies implies that, this type of scaling does not necessarily have impact on the outputs all the time and reliable changes on accuracy and precision

[6] www.analyticsvidhya.com

depends to the dataset itself and it happens when the features have different ranges. In other worlds, applying normalization impacts the prediction if the data points are not distributed well[7].

After all, due to the different features properties that our datasets may have, data normalization has more subjective nature rather than having objective nature and applying normalization goes back to the nature of the data which we deal with, plus having a better insight about the whole dataset.

1.5.4.5 Standardization

When we deal with some attributes with numerical values in our dataset, and they are far away from each other, as two different attributes from numerical point of view, we may apply Scaler to transfer them into an acceptable range.

Although, we can write our own function to do so, as there is formula available from statistical science but there is a readymade standardization function available from Sklearn library which makes it more comfortable to apply the concept.

It calculates the mean[8] and standard deviation of that column and then for each data point it subtracts the mean and divides the result by standard deviation to transform all the values into the suitable scale.

The difference between normalization and standardization comes to the picture when both concepts rearrange and make the range of the points meaningful for the machine.

There are different aspects available to deal with it as both concepts has more subjective, rather than objective nature as mentioned earlier, but it's found when the distribution of the data follows the Gaussian distribution, the standardization technique is reported to be more meaningful[9] data application.

---

[7] www.medium.com

[8] www.towardsdatascience.com

[9] www.analyticsvidhya.com

So, as we mentioned earlier, if the distribution of the data in our dataset does not follow the Gaussian distribution, then we better apply the normalization to rescale them between 0 and 1.

1.5.4.6 Training the model with machine learning algorithms

Next step after preprocessing of the data, is to apply training algorithms on training dataset. This is to make the model from split training dataset. After the dataset is prepared to feed into the algorithms, it must be split into two training and testing parts in a suitable way.

Usually the division is in 75%-25% or 80%-20% of the corresponding dataset in such a way that the larger amount of the data goes for training subset and the rest of that kept for the testing subset. So, the result of the algorithm after applying on training section is tested and compared with the testing subset which was not included in the training data subset.

But more precisely, if there is insufficient data, then it's better to use some techniques for data splitting [18] like cross validation in case of supervised learning.

1.5.4.7 Supervised, Unsupervised and Semi supervised learning

Labelled dataset availability is the difference between supervised and unsupervised types of learning. In supervised learning, our dataset is labelled, and we can make a correction of our prediction from the training subset with mentioned labels. It's called supervised hence the process of learning can be thought of like a teacher supervising the learning process.

If there is no label data given, then there is no correction anymore on the training subset and the algorithms decide on their own to discover some interesting pattern or structure from that dataset, so it's called unsupervised learning.

If the large amount of dataset is partially labelled, then we probably could have the teaching and correction method but not for all our dataset, so the mixture of supervised and unsupervised techniques can be used as progress method (semi supervised) of this type of learning.

1.5.4.8 K-Nearest neighbors' algorithm (KNN)

A supervised machine learning algorithm that can be used both for regression and classification problems. The concept of KNN algorithm is based on this assumption that, similar things exist in close proximity. In this algorithm, k is the number of the neighbors chosen at the beginning of the procedure, then for each instance in dataset, it calculates the distance (also called Euclidean distance) of that point to those neighbor points and adds the instance to the closest category. This procedure is repeated for all the instances of that dataset.

1.5.4.9 Cross validation

Cross validation is another statistical method that can be referred as an evaluation method for machine learning models. A resampling procedure to measure machine learning performances on a limited data sample is another better definition of cross validation technique.

Perhaps when we are dealing with enough amount of data, applying cross validation may not impact a lot, but for limited data samples it's one of the key factors [19].

In this technique, the given dataset is going to be split into groups, and the number of groups are represented by a single parameter called k. So, in 3-fold cross validation the corresponding dataset is split into 3 groups.

1.5.4.10 Data sampling techniques, stratified sampling

There are many sampling methods are available, but we explain the stratified sampling as we applied it on our dataset.

It's a test set of the population, which represents the best entire population being studied[10]. The random sampling in stratified techniques is different and involves the ransom

[10] www.medium.com

selection of data from entire population. This method avoids bias sampling as there is a sample data selected from all different verities of the population.

1.5.4.11 Ensemble learning

In order to create another optimal predictive model with most accurate predictor, many base models combined in a new form of a new united optimized model which is called ensemble learning. Ensemble technique utilization is with the decision trees even though they are not the most popular one used for ensemble learning technique. Bagging and Random Forest models are different types of ensemble learning.

1.5.4.12 XGBoost

Extreme Gradient Boosting is a decision tree-based ensemble learning machine learning algorithm which uses the gradient boosting environment. This algorithm is generated in a development process from a decision tree base model. It covers a wide range of applications to solve different problems in regression and classification prediction problems. It's much faster than the other algorithms in the same class and adjustable with different environments.

1.5.4.13 AdaBoost

It stands[11] for Adaptive Boosting, so another boosting technique that is used in ensemble method of machine learning. The weights are reassigned with higher weights to each instance, which is incorrectly classified. So, the learners are grown exponentially. In other word, weak learners are turned into strong learners.

1.5.4.14 Random Forest

Unlike AdaBoost, we can have unlimited depth of the trees in Random Forest algorithm. In previous ensemble learning algorithm, the learner can have two children in first stage,

[11] www.mygreatlearning.com

but in random forest algorithm, the tree can have much more width in the beginning stages. It creates decision trees on data samples under supervised learning approaches, and gets the prediction from each, and finally takes the best solution by voting concept.
It overcomes the overfitting problem in datasets and maintains high accuracy. Normalization usually doesn't impact the performance in RF as usually there is not any significant changes in accuracy after prediction without scaling[12].

1.5.4.15 Bagging classifier

Bootstrap aggregating or Bagging algorithm is another powerful ensemble learning. It's an application of the bootstrap method for high variance machine learning so, it can be used to reduce the variance usually in decision tree algorithms, so overfitting can be avoided.
In other world, bagging has the primitive effect of random forest algorithms since RF is the improvement of the bagging algorithms.

1.5.4.16 K-means clustering

It is one of the simple, accurate and popular unsupervised machine learning algorithms. The value of the K as the target division digit, defines the number of the clusters must be looking for by the algorithm in the corresponding dataset.
It works in such a way that the beginning points for cluster centroids are selected randomly and then with iterative calculation the position of the centroids are stabilized and the related points in each boundary will find their place by the distance calculations.

1.5.4.17 Model tuning

Hyperparameter optimization is to increase the model accuracy by customization of the model to the corresponding dataset. Random search and grid search are two different approaches of hyperparameter tuning.

---

[12] www.tutorialspoint.com

1.6 Thesis short structural setup

The thesis is structured as follows:

- Chapter 2 is the review of the literature. It provides background information on available researches and previous works which is done in related field. In the section concerning search range, several methods and algorithms are presented. This section describes some of the different techniques developed for performing different sensors and their connectivity issues. Finally, another proposed approach to solve the mentioned problem partially and it's the topic of this thesis itself.

- Chapter 3 is the methodology of the thesis. It gives a general description of the design and planning in order to solve the research questions. It reflects the master topic itself and describes the progress of the implementation part.

- Chapter 4 represents the implementation process. It explains from the first to last step of taking action to solve the problems for answering the research questions in detail. It also handles result and evaluation of the thesis. It gives a detailed description of the new method and the research methods that have been developed. This chapter also explains how the performance is evaluated.

- Chapter 5 presents discussion. It describes whether the problem can be solved using the approach presented in all the parts. It compares the results and discusses the outcomes from the comparison.

- Chapter 6 is conclusion which provides the summery of the work carried out in the thesis. This chapter briefly explains the goal of the thesis and how they gained outcome satisfies the corresponding goal and the future work.

Chapter 2 Related works

The thesis project is carried out at the Institute from the Institute For Energy Technology Halden (IFE) in Norway for the urban water pipeline network of the city, joint with the municipality of the Halden as co-project leadership.
Picture shown below is the control unit of Halden municipality water distribution system.

Figure 2-1 Controlling unit of Halden municipality water distribution system

2.1 Research topic area

As the keyword selection for our research is essential at this point, we go through the mentioned research questions once again. According to our research questions:

- What are the appropriate machine learning methods in acoustic leak detection?

- Which attributes are playing important role in detection of the leakages in urban water pipeline data analysis?

The first research question seeks all the approaches available when two combinations of "machine learning" and "acoustic water leak detection" meet each other.
Since our project is about the detection of the water, we consider the word "water" as separate keyword in our searching keywords as leak detection of other liquids might have some other scientific reasons to deal with, especially when we are investigating in acoustic sound and sensor field.
We used high citation databases and the google links listed in table for our research along with combination of "water" AND "leakage" AND "machine learning" AND "acoustic sensor" as nominated keyboards.
We have adopted the methodology of Kitchenham & Bacca which categorize the process into three sub-categories as, planning, conducting the research and reporting the result [20] [21].
In planning sub-category, we must select suitable journals and define the criteria of our study. There are three criteria as general, exclusion and specific defined in planning section. According to our literature review methodology, selection of journals is the first step of the planning section of our systematic review.
The list of the nominated high citation databases along with found papers in 1st research iteration, with corresponding keyboards are listed as shown in Table2-1.

Nominated databases	1st research iteration based on selected keyworks	2nd research iteration based on title and abstract reading	3rd research iteration after reading the article
ACM	14	9	5
IEEEXplore	21	18	14
ScienceDirect	44	15	13
SpringerLink	21	5	3
Wiley Online Library	27	2	1

Google scholar	354	28	15
Total	**481**	**77**	**51**

Table 2-1 Nominated databases and search results

We must mention that, duplicated papers from google scholar search are discarded from the table 2-1. After selection of the journals, we must have a clear understanding of inclusion and exclusion criteria of studies.

2.1.1 General criteria

By having the proper criteria, we can categorize our findings for further analysis. We collected the papers published between 2008 and 2021.

Studies that describe the leak detection framework with wireless sensor networks from the 2nd research iteration demonstrated different categories. For example:

- Water quality monitoring with wireless sensor networks
- Water leak detection in different environments like, soil, underwater etc.
- Real time water leak detection with acoustic sensors as well as other sensors under machine learning approaches
- Leak detection in oil & gas industry by wireless sensor networks under machine learning approaches
- Leak detection using inner spherical detector (dynamic) approach in water and oil
- Leak detection according to the leak size

In general, we can classify gained information from different perspectives like types of the pipelines and the techniques used in leak detection methods. In most of the researches the pipeline systems are restricted to water, oil and gas, wastewater, and industrial pipelines. But the techniques fall into two large groups as **direct** and **indirect** methods.

The direct method of leak detection is when we realize directly that some pipe burst or explosion or even the leakage has occurred in our pipeline system. Visual inspection and soil sampling are the examples of direct method. Currently, some municipalities in Norway are using this method for leak detection.

Another direct method is hardware-based approach. This method itself categorized into two large classes namely 1) In-pipe devices and 2) Out-pipe devices, as mentioned in [22].

With respect to the rules and regulations in most of the municipalities in Norway, by using In-pipe devices, we need to go through many circumstances to get the required permissions from the authorities. So, the acoustic sensor method of leak detection is one of the "out-of-pipe device" approaches in this classification which seemed to be suitable for that purpose.

On the other hand, for indirect methods, software-based approaches in different status like, static, dynamic and combination of both, is another promising leak detection strategy. In what follows in review of the literature, we are going to investigate pipeline leak detection approaches in combination of two hardware & software-based methods from direct and indirect classes, and we will focus more on data driven part of the software-based methods.

2.1.2 Exclusion criteria

Another section of eliminated studies from the 2nd research iteration are the studies that not identified as articles in selected journals along with studies with "no open access" label. Studies included with the target keyword but are about some other topics or the term only appears in the references, placed in this criterion.

2.1.3 Specific criteria

The papers which gathered in the 3rd research iteration column, are the related researches to the project and they are reliable to be referred as verifiable resources. They have come out from some specific related criteria as follow:

- Signal processing & supervised machine learning
- Feature selection
- Leak detection with neural network with Mel frequency coefficient of acoustic sound
- Ensemble learning approaches for acoustic scenarios

2.2 Review of the literature

In our literature review, apart from conducting the research according to our keywords and title related projects, we tried to lead the investigation towards answering the mentioned research questions.

Our first research question clearly demands all available researches in the field. Although machine learning approaches in analysis of the acoustic sound data's for leak detection techniques is our primary research topic, but what we found, is the combination of all these terms and phrases which could be helpful to understand the whole project step by step. So, in what follows we overview in detail the specific criteria as 3rd research iteration.

2.2.1 Finding related literatures

In this part we try to answer the mentioned research questions based on the literature survey.

2.2.2 All available researches in field

RQ1: What are the appropriate machine learning methods in acoustic leak detection?

Reviewing all the researches in the field help us to find almost what should be studied?

As the project scenario is based on the occurrences of the real world and it is in touch with our daily life, the collection of the required data could be under different conditions which impact the entire performance evaluation like different techniques and methods for water leak detection in [23].

As mentioned earlier, direct method of leak detection classifies into two large in-pipe and out-pipe device classes.

Apart from applying machine learning algorithms, sometimes the detection of usual behaviour of the signals found by finger printing method [24] as some other methods are available as well.

Although acoustic methods is often used for a direct leak detection in some specific situations like, background leak, when the pressure caused by the leak is very low or when

the soil is already waterlogged at the time of leakage, it's not trustworthy to detect the leak by acoustic devices said [25]. The reason why, the acoustic technique is useful for more small leakages is that, the frequency of vibration goes down as leak size increases. Stealing the water is one of the sub-category challenges of the smart water IoT monitoring system. So, the real time leak detection monitoring in long range with the help of the internet of things can be solution for both problems at the same time. Implementation of the smart water system with the Lora technology with ultrasonic sensor is another approach which holds many similarities with the current project from the wireless sensor network perspective [26].

In [27] author claims to develop a system for the user with easy installation and self-calibration system, to show when, where and how much water they are using. the method implies that having the vibration sensor is worth and feasible to exploit the correlation among vibration on each pipe and reading meter to estimate the water flow rate in each pipe. the disadvantage of the method is that each pipe requires a separate vibration sensor which is a tedious task, and not feasible from financial aspect.

the most prevalent technology used in oil and gas industry for leak detection is the wireless sensor network [28]. The study is the comprehensive review detailed comparison of the most recent systems investigated for monitoring various anomalous events in oil and gas industry. The important requirement for WSN deployments in the related industry is discussed.

In this study [29] some other pipeline leakage detection framework for district heating systems DHS using multisource data is proposed, which the remotely sensed thermal infrared imagery, visible imagery, and GIS data are utilized.

Leak detection techniques with microelectromechanical approach is discussed in [30].

From the qualitative analysis approach, which is done on the research topic, they found 3 main categories, 1) MEMS WSNs 2) MEMS accelerometers 3) MEMS hydrophones. among them MEMS accelerometer is based on machine learning models. Data from pressure and flow sensors were used for detecting large leaks whereas smaller leaks were detected using data from acoustic/vibration sensors. For large leaks, a relatively lower number of sensors required since large leaks generate pressure pulses which could be detected over a long distance. Pressure sensors identified large leakages based on

transient methods while acoustic sensors used to complement pressure sensors in identifying small leaks. The study shows that, acoustic sensors play crucial role in wireless sensor network leak detection, even if we use the acoustic sensors included in some other techniques.

The detection of the leakages in different pipelines is investigated from different perspectives. One of the classified categories is the size of the leak. The leak localization in pipelines with small leakages, takes different strategies than detection of the pipe bursts. SELS TENG or single electrode liquid-solid triboelectric nanogenerator [31] is another method for identifying and detecting the liquids leakages. High classification accuracy is achieved, combining the application of TENG with big data and machine learning approaches.

Water pipeline burst detection with the help of the sudden changes in water flow/pressure is another method of finding the leakages which is classified as abnormal changes in anomaly detection strategies [32].

2.2.3 State of the art

RQ2: Which attributes are playing important role in detection of the leakages in urban water pipeline data analysis?

Using tethered robot with acoustic sensor is another approach for detecting the leakages in distributed water systems [13]. One of the drawbacks in mentioned system is the continuous maintenance of the sank robot in the drinking distributed system to avoid water pollution. Another issue which was our consideration at the beginning of the project is to convince the authorities to use such device inside the drinking water system, in touch with the water itself, which takes lot of efforts to report and deal with different organizations.

Apart from ensemble learning, another mixed up approach which is been popular for acoustic sound scenarios in the field is to employ some techniques improving the main applied machine learning algorithm performance. The impact of local binary pattern (LBP), an efficient texture operator, with different machine learning approaches as well as neural network algorithms is tested and reported in [33].

Another interesting study about the detection of an event from acoustic signals shows efficient result from applying machine learning algorithms on corresponding dataset which can be extended to our project scenario. An acoustic signal recognition technique is tested to detect the obstructed pipes in water circulation system with the help of support vector machine (SVM) algorithm [34].

The methodology shows the mixture of machine learning and a single acoustic sensor, "a viable option to predict pipe obstructions and the type of obstruction", said by the author.

By going through most of the studies, and analyzing their performance evaluations, it can be stated that using combined techniques and strategies usually gives better efficiency in applying different class approaches. The methods developed by combination of two or more techniques, shows more successful performance in leak localization [22] and can be used as promising approaches in future development.

PipeNet system which is made by combination of both "pipeline" and "network" is another interesting method that shows promising leak detection with acoustic metrics under data analysis approaches [35].

Few false alarms to a range of pipe material applications, cost effective to produce, install and maintain, ability to distinguish between sensor fault and system fault, as well as having flexibility in data-flow based programming environment, makes the PipeNet system a promising approach, said by the author [35].

An acoustic leak detection approach based on CNN with Mel frequency cepstrum coefficients is proposed in [36]. Acoustic approaches can be categorized into two classes. first class is only the detection of the leakages, but second class is not only leak detection but structural condition inspection with some techniques like tethered and acoustic emission. There is a feature extraction methodology for acoustic sound conversion handled by Mel frequency as the auditory feature covered by the convolutional neural network approaches to detect the leakages is another interesting source related to the second research question.

The effectiveness of SVM algorithm over RVM is verified in the [37]. It shows that SVM can give much better accuracy in case of multiclass classification rather than binary classifications. It also stated that, the acoustic emission features are used to identify and localize the leakages in pipeline with applying suitable instruments.

Analysis of applying K-means clustering algorithm for the vibration data collected from the PVC pipe surface due to the water flow in order to classify the abnormal detection of the flow inside the pipe can lead into finding sudden detection of the leakages [38].
Another use of acoustic emission for detection of the leakages in different pipe material is discussed in [39]. The investigation is done on the water-filled plastic pipes using tuned wavelet for clustering and localization of acoustic emission signals as well as detection of the leakages.
Acoustic emission can be coupled with accelerometers to detect incidental events such as break or crack growing. This strategy shows the changes from small leak pipe to a big pipe crack event which is when the acoustic signals exceeds the predefined baseline [40].
The collected data is analyzed with different algorithms like SVM, decision tree and Naïve bayes with very high-level accuracy in distinguishing the leak states from non-leak states.
Testing different machine learning algorithms like SVM, KNN and neural network in classification of the acoustic sounds on a customized dataset from the ambient events, implies the better performance of the customized LSTM-CNN algorithm in compare with other classification algorithms in different sound environments [33].
Another research proposes a novel method in high pressure steam leak diagnosis [4]. The method is to find out the distinguishable features from the acoustic signals which are captured by remotely microphone sensors and evaluated by the RF, XGB and KNN algorithms. The outperformed results of the decision tree algorithms on this pressure case is important for our study.
One of the common methods to find out sudden water leakage in smart water systems is the use of regression machine learning approaches on the water consumptions [41]. The model can be made by hourly intervals provided information on cumulative water consumption.

2.2.4 Summary of the literature review

The table shown below, briefly represents the literature review summary:

Nominated databases	Leak detection with ML approach	Leak detection with ML & acoustic approach
ACM	5	3
IEEEXplore	14	2
ScienceDirect	13	9
SpringerLink	3	1
Wiley Online Library	1	0
Google scholar	15	0
Total	**51**	**16**

Table 2-2 Literature review summary

The detail of the related found papers is briefly described and classified with research questions.

Chapter 3 Methodology

3.1 Design model

In this chapter will explain in detail how we designed our project in order to answer the research questions. After the review of the literatures in previous chapter, we came up with that conclusion that, some more experiments and investigations are required in software-based data-driven part of the leak detection techniques.

The method used in this study is based on the process model structured from design science research methodology in information science offered in [42] which is originated from behavioral and design science in [43]. Behavioral science tries to find out "what is true?" while, design science paradigm searches for "what is effective?" so both are inseparable requirement factors of a suitable research methodology.

The important part of the design science is that the research should produce an "artifact", which addresses the problem correctly and its utility, quality and efficacy must be evaluated rigorously.

3.2 Exploring of the methodology

Exploring the design model of our study according to the design science methodology we have taken lead us thorough 6 important activities in nominal sequence as follows:

- Problem identification and motivation
- Define the objectives for a solution
- Design and development
- Demonstration
- Evaluation
- Communication

In problem identification and motivation section, we define the research problem and we try to justify the solution, as problem definition is the reason to develop an artifact which is the solution to our problem.

This section interpreted in several ways by different researchers like important and relevant problems, analysis [44], identify a need and important and relevant problems by some other researchers [42] . This identification is tried to be explained in previous chapters as the only resource required for this section is the knowledge about the problem along with the importance of the solution.

In defining the objectives for a solution section, we must drive what is possible and feasible from the problem definition, as the objective of the solution. The resources required for this section is again the knowledge of the state of the art about the problem and the current available solutions.

Moving from objectives to design and development is the 3rd iteration process of the general procedure. After this section in project, the artifact is created and can solve one or more instances of the problem. The testing is the operation of the demonstration activity, which involves the use of artifact in experimentation, simulation, case study and other appropriate activities.

Figure 3-1 illustrated the general design science research method which can used for IT projects

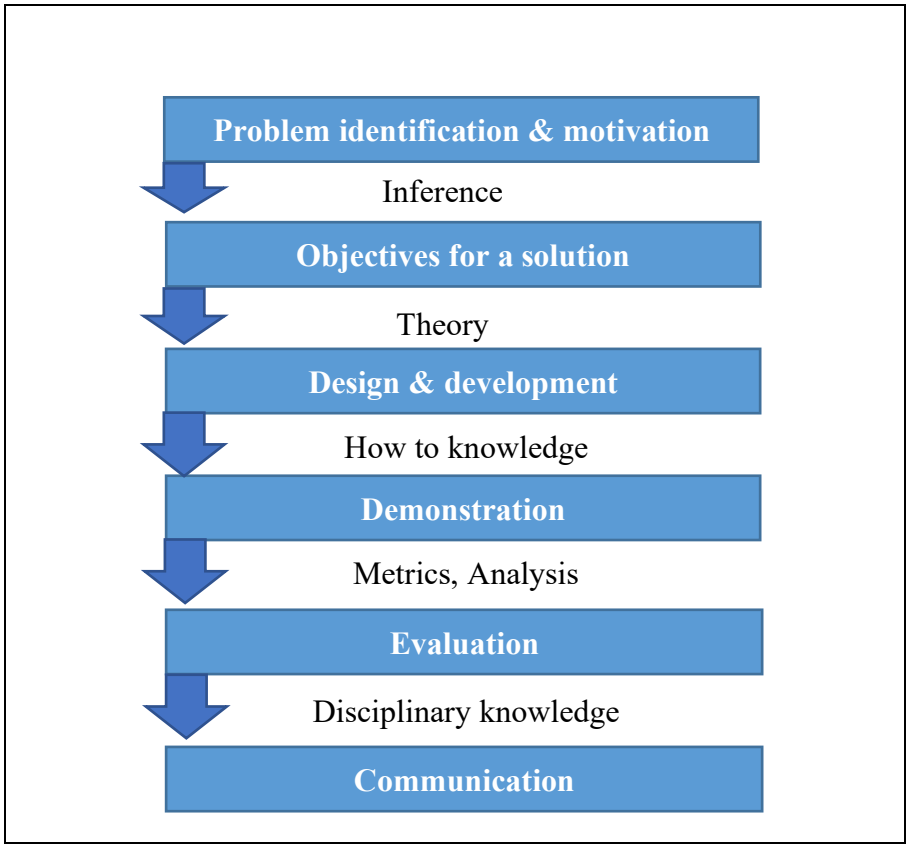


Figure 3-1 Design science research method process model in information science

Evaluation, measures how well the artifact can support the claimed solution to the problem. This activity consists of different comparison among the required satisfactory level of the solution to the problem and actual performance of the artifact in demonstration part. The final activity of the mentioned process model is the communication activity which testifies the utility of the artifact. In this section after utilization of the model, it reveals whether the model or artifact is designed rigorously.

3.3 Planning and design

Although this study consists of two different approaches, but the focus of this research is towards deploying data science techniques for the analysis of the received data, as well as available data in leak detection applications.

The study is involved with different steps which carries different concepts in each stage like, sensor installation, acoustic sound investigation, sensor evaluation, wireless sensor network, data storage, and analysis of the stored data at the end of scenario. So, the project is divided into three main sections as follows:

1) Sensor installation and configuration
2) Data transfer from the sensor to a local gate and transferred to the storage
3) Analysis of the stored data

As we mentioned earlier, our intention in this study is the 3rd part of this classification, and what we discuss further from other parts of this classification is the required information to support the hypothesis and the mentioned research questions.

Figure shown below illustrates the schematic representation of the project scenario from the general perspective:

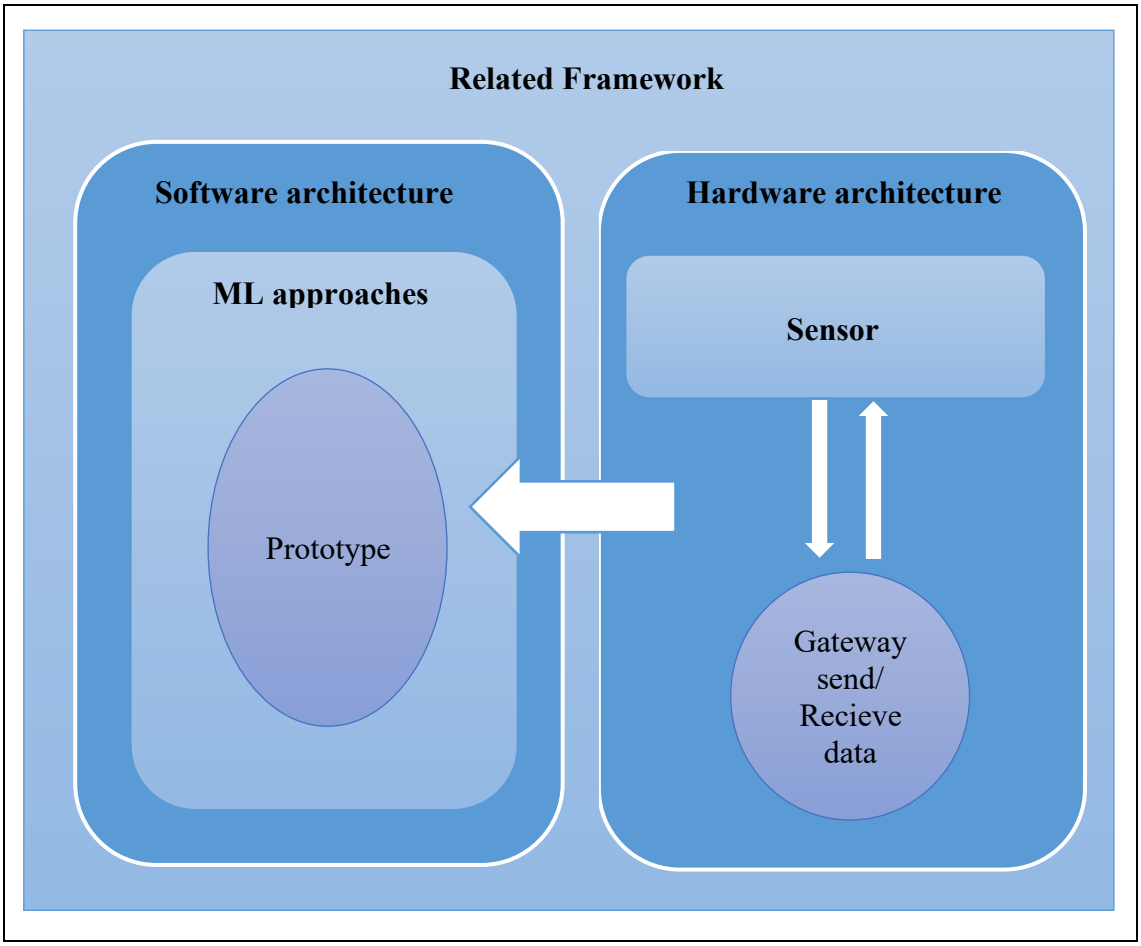


Figure 3-2 General perspective of the project included two different approaches communicating together

3.4 Hardware architecture

After problem identification, our first primary idea in defining the objective for the solution of the problem is represented as shown in the figure 3-3.

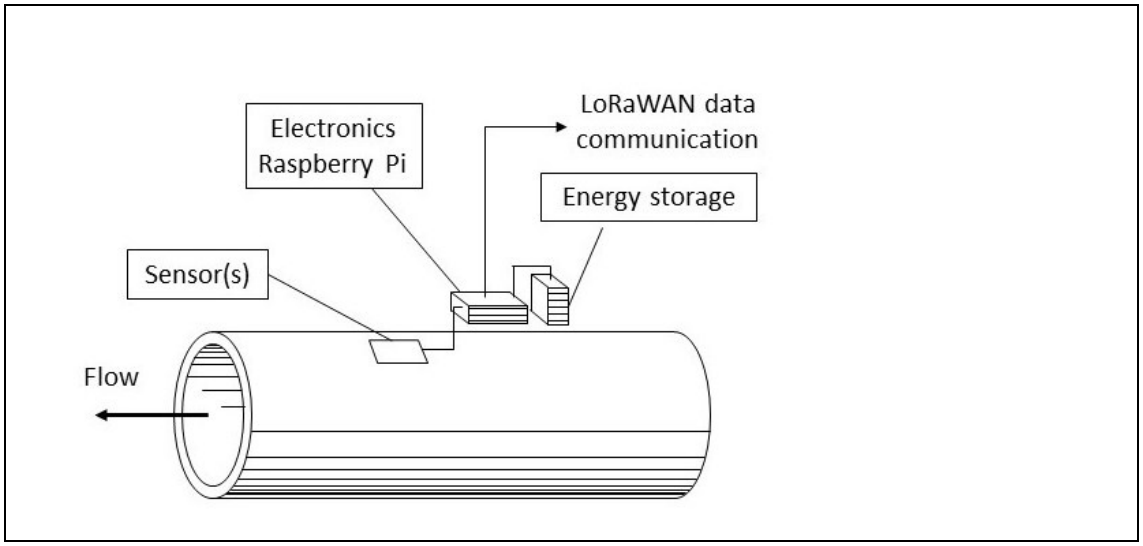


Figure 3-3 Primary idea[13] of the project in defining the objective stage of the project

3.4.1 Sensor device

The nominated acoustic sensor in our project is consists of a contact microphone CM-01B[14] along with 3D digital accelerometer with high performance mode ability and enabling always-on low-power feature for an optimal motion experience.

The contact microphone built with sensitive and robust piezoelectric material combined with low noise electronic preamplifier to provide a sound or vibration pick up with buffered output. The microphone can minimize the external acoustic noise, while being highly sensitive to vibrations.

The microphone is located inside of the holder with metal spring behind it to induce microphone sensitivity as shown in the figure 3-3. The spring force is increased by screwing the plunger. The microphone and accelerometer installed on the top position of chosen water pipe located inside of the manhole. Then the response of the sensors verified as shown in the figure 3-3.

[13] From internal meetings IFE

[14] www.metrolog.net/cm01b.php?lang=en

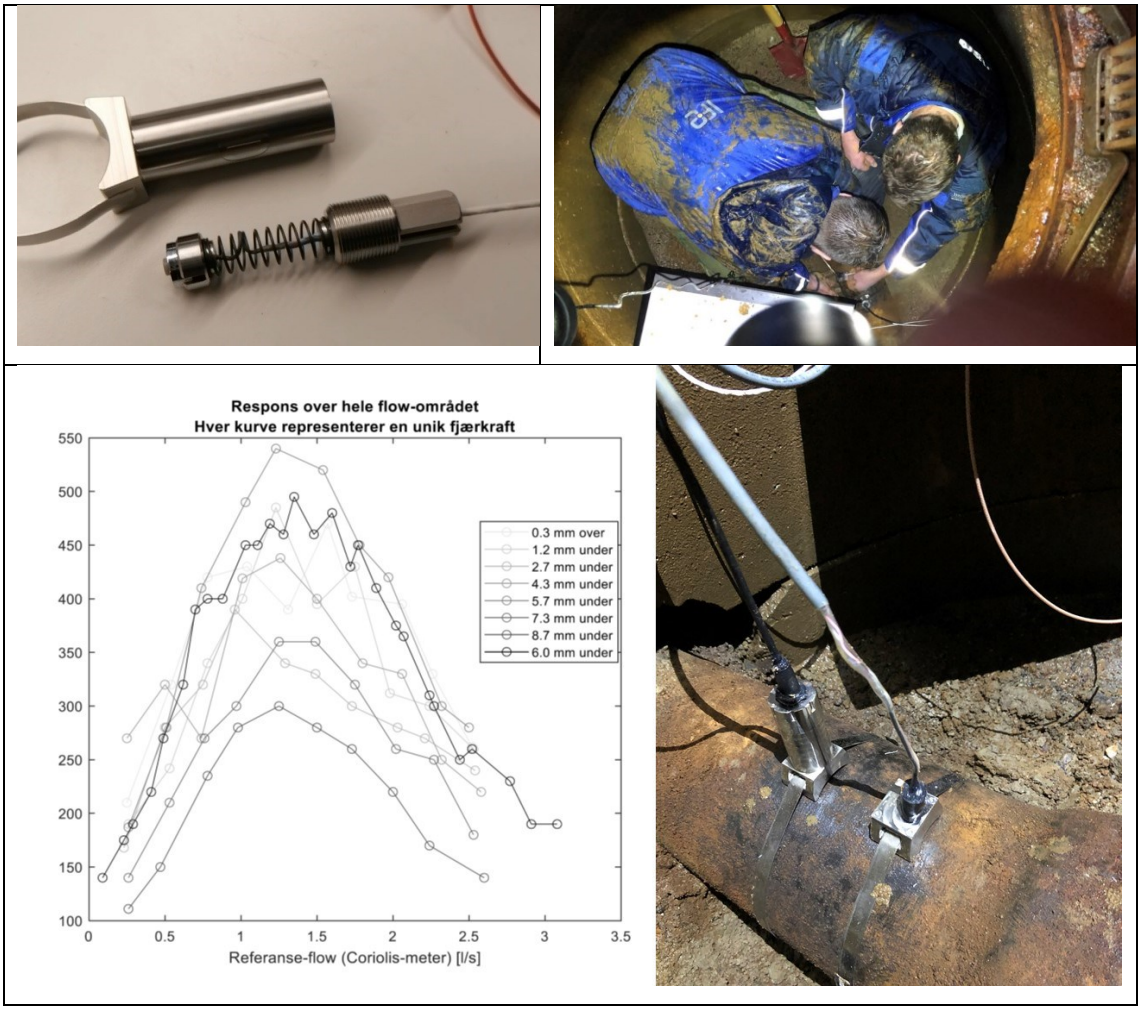


Figure 3-4 Installation of microphone & accelerometer from top position of the water pipe located inside of the manhole ~ 2 meters underground level - Response diagram of the sensors after installation (bottom left corner)

The location of the manhole is fixed after verification and confirmation among all the parties.

3.4.2 Send and receive gateway

Another important hardware architecture in our project is the transmission of the captured data to the storage, like wireless sensor network performance.

This part of the project is handled by LoraWan technology[15]. The concept of this technology is on star-of-stars topology which gateways trust the communication messages between the end devices and central network server.

A raspberry Pi connected to the sensors along with the antenna to transmit all the captured data of the sensors into the closest gateway station which is fixed on network-based infrastructure. The radio transmission handles the wireless section of the data transmission. Clearly there is a need for the energy consumption inside the manhole for the sensors and the raspberry Pi. In our study, we used a lithium battery to cover all the required current. The battery is fixed in order to complete and test our prototype as shown below:

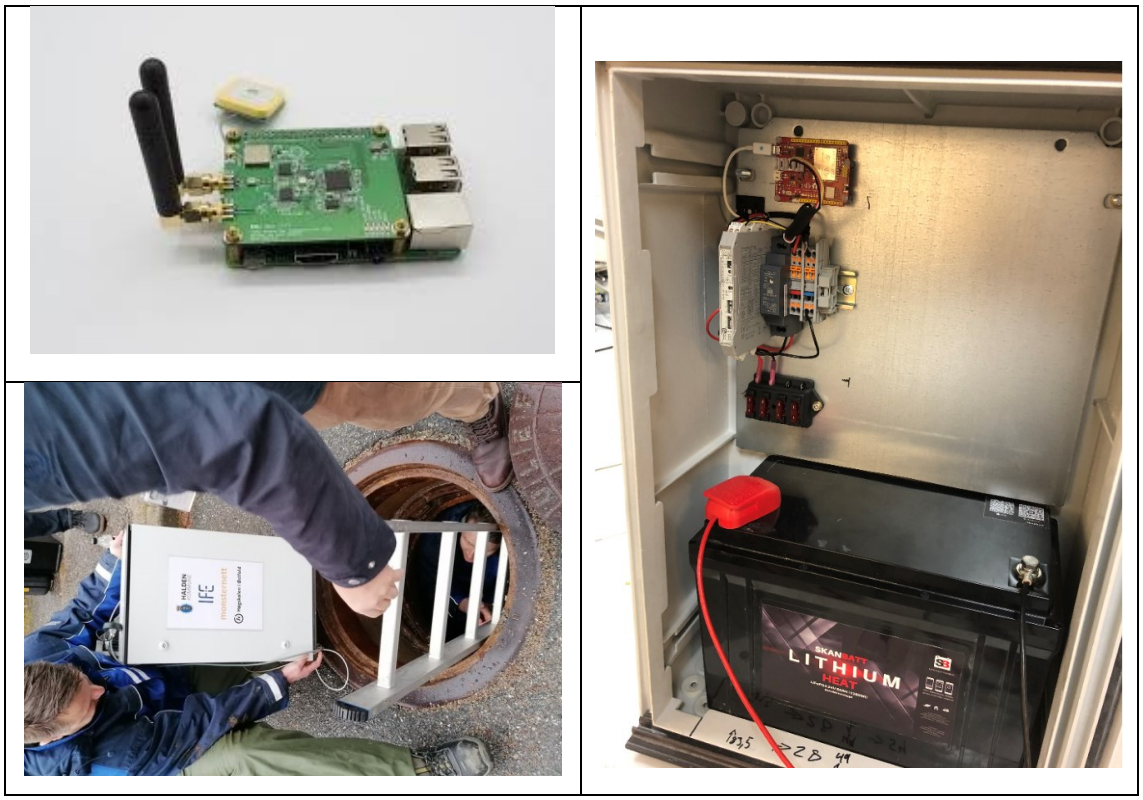


Figure 3-5 raspberry Pi of the LoraWan infrastructure and the energy consumption strategy

[15] www.lora-alliance.org/about-lorawan/

3.5 Software engineering architecture

After setting up the instruments in the field, we must find out the base software architecture of the study. In what follows, we focus on ML for software architecture, which targets on developing ML techniques for better application programs.
Although there is no difference between software components and machine learning components at the architectural level but, they will be considered as another components such as model generator or model consumer as well as being event generator and event consumer [45].
In here we will follow a normal data-driven process to realize the solution to the problem. The promised solution of the system uses IoT along with standard data pipeline architecture for data ingestion, data monitoring, statistical optimization, and data analytics to fulfill the demand for leak detection techniques.

3.6 Machine learning approaches

The machine learning approach in this study, must be included with real time prediction workflow, as well as historically prediction workflow. The real time prediction workflow is to make online predictions from the streaming dataset which is keeping updated by a time variable. Our project intrinsically demands the need of real time prediction in detection of the leakages at the time, but we skip that in this study, since we must have rest API calls enabled. The project is not reached the final level in storing the received dataset in real time way.
Figure 3-6 is the representation of the historical prediction workflow:

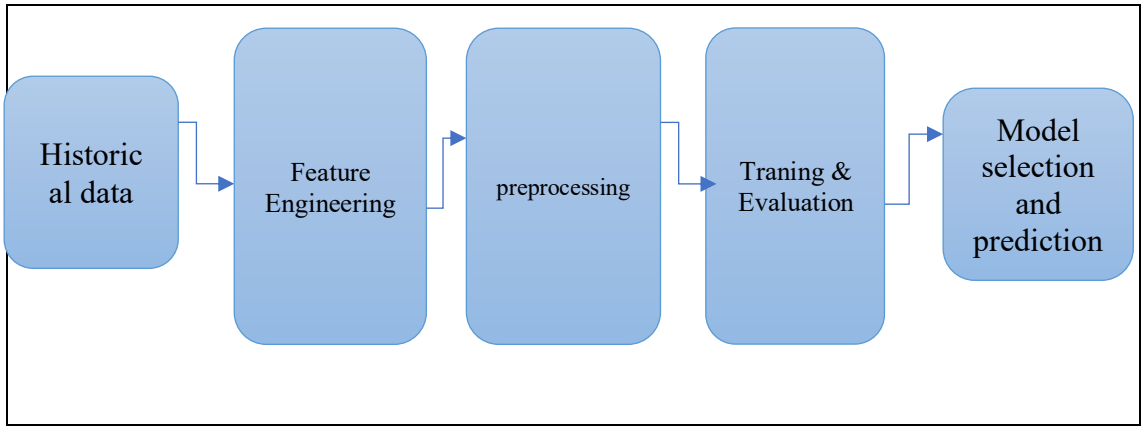


Figure 3-6 Historical prediction workflow

3.6.1 Data collection and consolidation

Data collection is the first step of every machine learning application. In this section, the required data is collected in different methods and will be changed into the suitable format for the learning application of the project.

When we receive datasets to work with, at the starting point of the project, usually we must deal with unstructured type of data. Unstructured datasets take lots of efforts to be prepared and suitable to the context of the project and in most of the cases the most challenging part of the machine learning process, is the data preparation part.

The more clean and structured data we prepare when the training section of the project begins, the more accurate model we will have after building the model and finally it leads us for better prediction.

Structured data is made by numbers, dates, string and usually takes less memory, but unstructured data types could be media files, text files and emails with larger capacity. More than 80% of the enterprise data will be unstructured [46].

In this study, we deal with 3 different datasets. In what follows we will discuss about the category and type of them in detail.

3.6.2 Data preparation

The datasets that investigated in this study are as follow:

1) Received dataset from the installed sensors on water pipeline in Halden city

2) Tosshullet water data flow from Halden municipality guard system
3) Yorkshire[16] water daily acoustic logger data
4) MIMII[17] dataset baseline

- First dataset is the transmitted data by LoraWAN technology from the sensors which already installed on the water pipe inside the manhole.
  The recorded file in the CSV format with two attributes as sensitivity and frequency from the microphone. There is no time set for the data, but we can estimate the dates, from the starting capturing point.

- Second dataset is the flow data of the water from the guard system of the Halden city belong to one of the water pipes in Tosshullet area.
  This dataset is also in the CSV format, consist of three attributes as date, time, and the flow of the water on that specific date.

- Third dataset is the Yorkshire acoustic logger data, belong to the water system of the Yorkshire county in northern England. The data consist of two files logically connected to each other. The first file is the acoustic logger data, consist of logger IDs, the acoustic average level, and spread of the noise measured over a few hours. The next included available file is the leak alarm file which consists of the logger IDs, leak alarm and leak found attributes noticed by the inspection visited date after receiving the alarm leak.

- Last dataset is the sound dataset for malfunctioning industrial machine investigation and inspection, known as MIMII dataset. "It contains the sounds generated from four types of industrial machines, i.e. valves, pumps, fans, and slide rails. Each type of machine includes multiple individual product models, and the data

[16] https://datamillnorth.org/dataset/yorkshire-water-daily-acoustic-logger-data
[17] https://github.com/MIMII-hitachi/mimii_baseline/

> for each model contains normal and anomalous sounds. To resemble a real-life scenario, various anomalous sounds were recorded. Also, the background noise recorded in multiple real factories was mixed with the machine sounds."

Data preparation must be done for each of these datasets in order to rearrange the entire rows and columns. Usually the given data must be rearranged with respect to the date and time of the event to be more understandable for further development.
The null values plus incomplete data columns must be handled for each dataset. Usually there are some methods to deal with those values in each dataset, but if the amount of that is not large, its recommended to be discarded. In this project we have removed the null values along with the uncompleted data points after data preparation of CSV files.
To work with the dataset and access the data in the script the CSV file is loaded into the Pandas data frame library in Python 3,7. This library tool handles most of the preparation process at the beginning of each project.

3.6.3 First dataset-First transmitted data from installed sensors

Two different files transmitted from the sensors by LoraWAN technology, are belong to the installed accelerometer and microphone. The microphone is consisting of piezo film inside to capture the event movement and convert into the volt to show the recorded vibration in volt per mm as described in the manual. Therefore, we must have the proper timestamp to link the accelerometer files side by side to realize the direction of the vibration in case of any event.
The timestamp is not available in the recording configuration of the sensor, but we can estimate the experiment time from the sensor installation date.
After receiving the data from two sensors, we have three coordination direction in accelerometer file and sensitivity along with frequency in received file from microphone. We converted the frequency of that noise from time domain into the frequency domain by Fourier transform equation. We realized that there is no important event in that period. On the other hand, there is no leakage occurrence reported in that period, said by the guard set of Halden municipality monitoring system.

3.6.4 Second dataset- Tosshullet water flow dataset

During the project period, the water monitoring guard system of the Halden municipality reported a leak detection in Tosshullet, Halden.

When we requested the data from the municipality guard system, we received the water flow data from 17th December of 2019 into 3rd October of the year 2020 periodically for the targeted location. The data consists of date, time, and the water flow in every second throughout the day.

After pre-processing unit, we assumed two different conditions for the captured dataset. First, we assumed that the data is unlabelled, and we tested the unsupervised K-means clustering algorithm to find out any anomaly behaviour and in second assumption, we made the dataset labelled with respect to the municipality manual leak report. The labelled leak found column is added and we performed a set of supervised learning algorithms to find out classification accuracy. There are two scenarios defined inside the supervised learnings to test the result with/without date included in the dataset. Two classification algorithms (XGB and Random forest) tested and the result is reported with grid search hyperparameter tuning.

3.6.5 Third dataset – Yorkshire acoustic logger data

This dataset is chosen and tested as one of the case studies in this thesis. The positive point about this project is the scenario which is acoustically analysis of the leak detection with help of the loggers. It almost holds most of the aspects of our project.

As explained earlier, Yorkshire water system is deployed a set of acoustic loggers for leakage detection. The loggers are fixed and constantly listening to the water flow network. They will gather more data in case of hearing a leak noise from the system. If the leak alarm raises, there will be a site visit to check the area. Some points near the alarmed point is also visited to see if a leak can be identified.

We can have two different assumptions on this dataset, since only few instances of the combined dataset are holding "yes" label for leak alarm and leak found columns and for the large number of rows there is no clear label addressing.

- On our first assumption is, finding the leak instances of the data and set them as leak label then we can label the rest of the dataset for “no leak” as there is no reported leak. Considering whole of the dataset as labelled dataset and proceed with supervised learning.
- The second assumption is, finding the leak instances of the data and set them as leak label and leaving the rest of the dataset as it is and consider it as semi dataset for semi supervised machine learning.

Except two attributes for leakage data in the dataset, we are dealing with four main attributes like ID, Date, Average level and spread level of the noises. We will also verify the algorithm performance with considering only two features of the average level and spread level of the noise from the feature importance perspective.

3.6.5.1 Using Label propagation concept and self-training approach

Using labelled part of the dataset for training part to make the classification model and then predicting the unlabelled part of the dataset with respect to the created model is called self-training semi supervised strategy. In our second assumption, we used self-training approach. the technique [47] [48] works as follow:

- Step 1: starting with the labelled part of the dataset and split into the training and testing sets. We used 70% for train and 30% for the test
- Step 2: trained classifier is used to predict labels for all the unlabelled data instances. In this part, the probability of the label being correct is verified and the highest ones classify as ‘pseudo-labels’
- Step 3: concatenation of the ‘pseudo-labels’ with the first labelled training data and retraining the model with new training dataset
- Step 4: predicting the left unlabelled instances and evaluation of the algorithm performance. The steps can be repeated until no more unlabelled instances left.

3.6.6 Fourth dataset – MIMII

From this dataset, we can find out the importance features of the acoustic data in some other environments. The dataset belongs to the malfunctioning industrial machine

investigation and inspection. The main intention of using such dataset is what are the main features of the industrial acoustic noises and how it helps us in our project.
So, the first step to analyze the audio is the file conversion. After the files are transferred into the CSV file, data preparation is started.
We built some other features like, mean, median and standard deviation apart from minimum and maximum features to help our training model to achieve a better result.
comparing autoencoder and LSTM neural network with some other algorithms on this dataset with respect to our intention and data preprocessing is another experiment of this part.
We conducted hyperparameter tuning optimization to see the changes in our algorithm's performances in different combinations of the Min, Max, Mean, Median and standard deviation attributes.

3.6.7 Feature engineering

3.6.7.1 Fast Fourier transform equation and audio to CSV file conversion

Using Fast Fourier Transform is the common method to convert sound from time domain to frequency domain. "Fourier analysis converts the signal from its original domain to a representation in the frequency domain and vice versa". The Fourier transform formula [49] of the function f(x) is the function F(w) where:

$$\mathrm{F}(w) = \int_{-\infty}^{\infty} f(x) e^{-iwx} dx$$

And the inverse Fourier transform is

$$\mathrm{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(w) e^{-iwx} dw$$

With the corresponding algorithms, we can see the data points of the sound files and vice versa.

3.6.8 Data pre-processing

3.6.8.1 Feature selection

It's a very important concept of machine learning that can change the entire performance of the algorithm in each machine learning approach. To do so, we must have a clear understanding of the problem and the dataset we are dealing with.
This core concept in machine learning, identifies the important related attributes of corresponding dataset and removes the irrelevant and less important features from the dataset. We can make some supportive attributes and add to them to the working dataset to impact the prediction.

3.6.9 Train – Test data split

Splitting the data into train and test set is the next job after the pre-processing section.
The nature of the problem of our project and its target values, implies that the classification machine learning approach must be chosen. So, the data splitting is much easier when need to apply the classification approach. Usually we don't need to worry about the timestamp, and we have already shuffled our dataset before loading for training session.
In this study, splitting operation is handled by using the Scikit-learn library packages.

3.6.10 Model training and evaluation

Our dataset is ready for training now. We have the desired dataset prepared for the training to create the best predictive model. In this project, we have set our pre-processing in such an integrated way that we received a pickle file for each algorithm separately, as it is required sometimes to have different set of datasets rearranged for our new algorithm. For instance, sometimes the type of the data we load for the neural network is different than others. Apart from that, we need to apply different approaches on our dataset, so we need to be prepared and plan for such sudden changes from the beginning of our development process.
there are many techniques available, for the evaluation of our model. Confusion matrix is the first evaluation method applied in our project as we need to know the ration between

the true positive and false negative. Some evaluation metrices which give us the better insight of the model performance throughout the project are described below.

3.6.10.1 Evaluation metrices

- Classification accuracy is what we usually looking for in evaluation of the classification algorithms. The formula[18] shown below describes the term accuracy:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TruePositive + TrueNegative}{TotalSample}$$

This definition is the also the confusion matrix accuracy, where we try to identify the number of the correct and incorrect predictions in positive and negative predictions.

- Area under the curve is one of the usual metrices for the evaluation section. "AUC is the area under the curve of plot false positive rate vs true positive rate at different points in [0,1]".
- Testing the accuracy is measured with F1Score. It says, how well the model classifies the instances correctly. The corresponding formula is shown below:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- The precision is the number of correct positive results divide by all positive results as shown below:

$$Precision = 2 * \frac{TruePositive}{TruePositive + FalseNegative}$$

All the metrics are tested, and we took Area Under Curve (AUC) and Receiver Operating Characteristics for all the supervised machine learning algorithms.

[18] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

3.6.11 Feature importance

Feature importance is the set of techniques to examine the attributes and scoring them to a predictive model that shows the importance level of each, while making the prediction. Applying feature importance is essential in our case studies and the result of different experiments in importance of the features is one of our main intentions of the case studies to help our project.

3.6.12 Optimization techniques

After the model created, it needs to be modified in its best possible outcome combination as there are many values are available for the algorithm parameters, which can be tested. So, the tuning process is started.

3.6.12.1 Grid search for model tuning

The process of model tuning goes through the hyperparameter tuning optimization process. It's the technique, that makes sure that, all the possible combinations with different parameter values defined for the algorithm are tested with the available model and the best possible output is already chosen.
A parameter is an internal characteristic of the model and the value of that depends on the data being applied. Grid search tests all possible combinations of different appropriate parameters and report the best results along with the best parameters.
For all the applied algorithms in our study we have examined the hyperparameter tuning optimization process and desired outputs are noted.

3.6.13 Imbalanced dataset in machine learning

3.6.13.1 Tactic to handle imbalanced datasets

We know that, one of the main challenges of our datasets in this study is the imbalanced data. We don't have the complete labelled or unlabelled datasets and most of them are semi labelled.

One of the important techniques to handle the imbalanced data is the sampling strategies and stratified sampling is one of the effective sampling among them. "It ensures each subgroup within the population receives proper representation within the sample".
We see the effect of stratified sampling in our case studies.

3.7 Thesis's tools and equipment

The thesis algorithms are, XGB, RF, KNN, AdaBoost, Bagging and some preprocessing concepts like feature extraction uses python 3,7. The mentioned algorithms along with the convolutional neural network CNN, use the following third-party framework and libraries:

- Keras modules[19] – A high level deep learning tool for TensorFlow backend
- Scikit-learn – A very useful machine learning data analysis tool [50]
- Numpy – Fundamental package for scientific computing in python [51]
- Matplotlib[20] – A library for creating static, animated, and interactive visualizations in Python
- Yaml[21] – A parser and emitter for Python

The following third-party frameworks used for audio feature extraction:

- Librosa[22] – A python package for sound analysis
- Audioread[23] – "Decode audio files using whichever backend is available"
- SciPy[24] – An open source library for mathematics, science, and engineering

---

19 https://faroit.com/keras-docs/1.2.0/
20 https://matplotlib.org/
21 https://pyyaml.org/wiki/PyYAMLDocumentation
22 https://librosa.org/doc/latest/index.html
23 https://pypi.org/project/audioread/
24 https://www.scipy.org/docs.html

Chapter 4 Tests, Results & Evaluation

In this chapter we show the details description of the methodology section of our study. The description of experimentational setup in different approaches like, purpose, explanation, and motivation of the test. Corresponding tests results are represented, and the evaluation of the gained results are shown at the end of this chapter.
We explained in previous chapter, that with respect to our datasets, somehow related to each other in different manners, we started with the pre-processing section and we will discuss the applied approaches on these datasets in detail in discussion chapter.
In what follows, to avoid confusion, each dataset is pre-processed, tested, and evaluated separately and the results are discussed at the end of the chapter.

4.1 First dataset

4.1.1 Data preparation and pre-processing results

The transmitted data by LoraWAN technology from the microphone and accelerometer, is included with two attributes as, frequency and sensitivity.
Corresponding file with the same period is transmitted from the accelerometer with 3 directions as X, Y and Z. The sensors are installed on 20th of December at 13:30 and they started to capture from the same time. It must be mentioned that both files don't have the timestamp. The captured frequency from the microphone, should be converted from time domain into the frequency spectrogram for further development.

4.1.1.1 Frequency spectrogram

The spectrogram is another visual method of representing the signal strength over the time. By converting the frequency column to the spectrogram, not only we can see

whether there is energy at the event from 0 vs (n) HZ but also, we can see how energy levels vary over time[25].

The representation of the frequencies of first and second received microphone file is shown in figure 4-1 right column. The corresponding X, Y and Z points from accelerometer is illustrated in the left side of the figure.

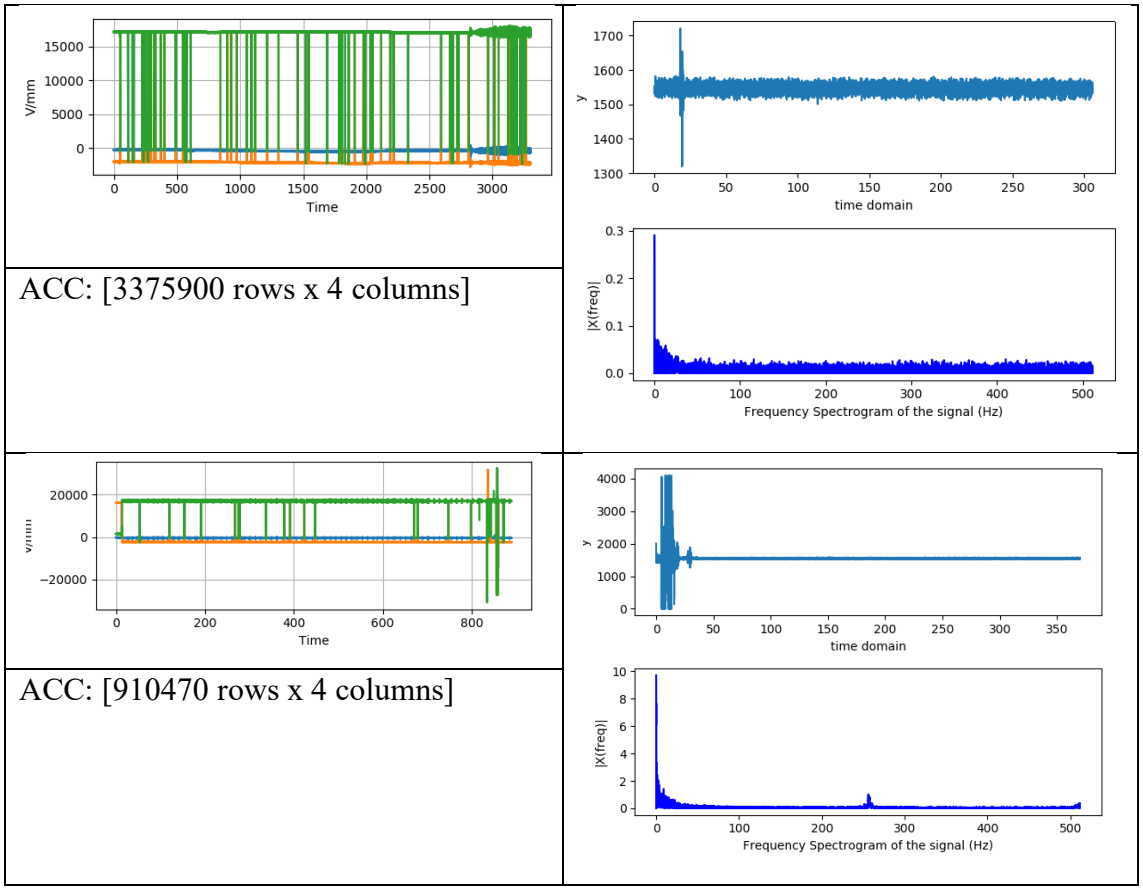


Figure 4-1 converted frequency spectrogram of first couple of received files

The whole capturing period is less than a month due to the limited capacity of the battery installed inside the manhole. Following figure shows the whole period of capturing

[25] https://pnsn.org/spectrograms/what-is-a-spectrogram#:~:text=A%20spectrogram%20is%20a%20visual,energy%20levels%20vary%20over%20time.

vibration and sound for the installed accelerometer and microphone along with the frequency spectrogram.

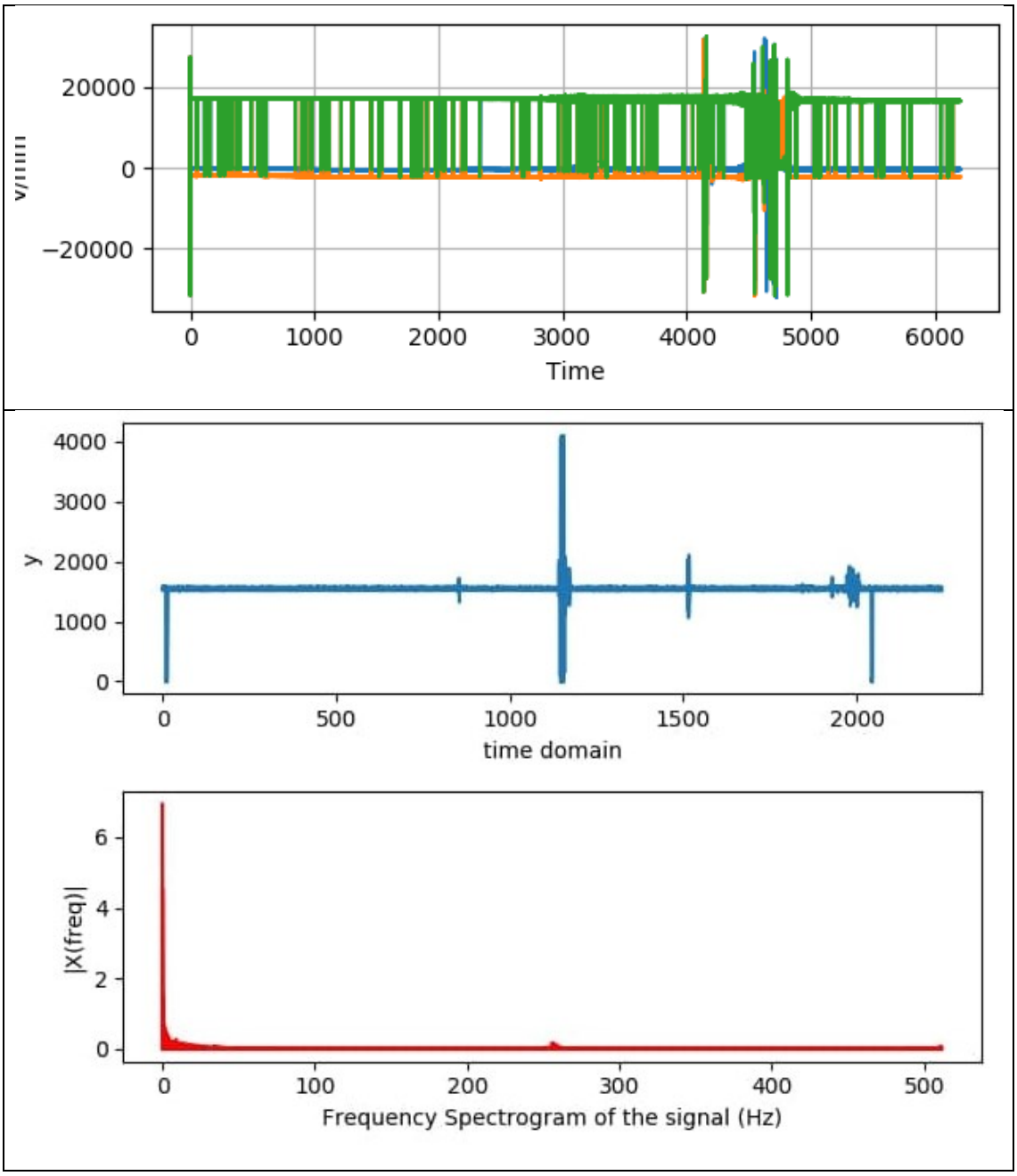


Figure 4-2 Entire of file representation of accelerometer & microphone in that period with frequency spectrogram

The more is the time domain period representation of the plot; the less file variation could be illustrated in the picture as the entire file must be shown in a small picture. So, in figure

4-1 we tried to show some detail example of the file in small scale period for both accelerometer and microphone, but in the figure 4-2 we can see the entire file represented over the time. We must mention that there is no leakage in that mentioned period, reported by the municipality.

For better analysis and better vote for any types of detections, we need to collect some more attributes, specially from the machine learning perspectives. In water flow distributions, the flow of the water is one of the important attributes and in many researches, flow of the water counted as one of the important features included in the studies.

4.2 Second dataset – Tosshullet water flow

The schematic representation of the received data from the guard system for occurred leakage in specific area is shown in figure 4-3.

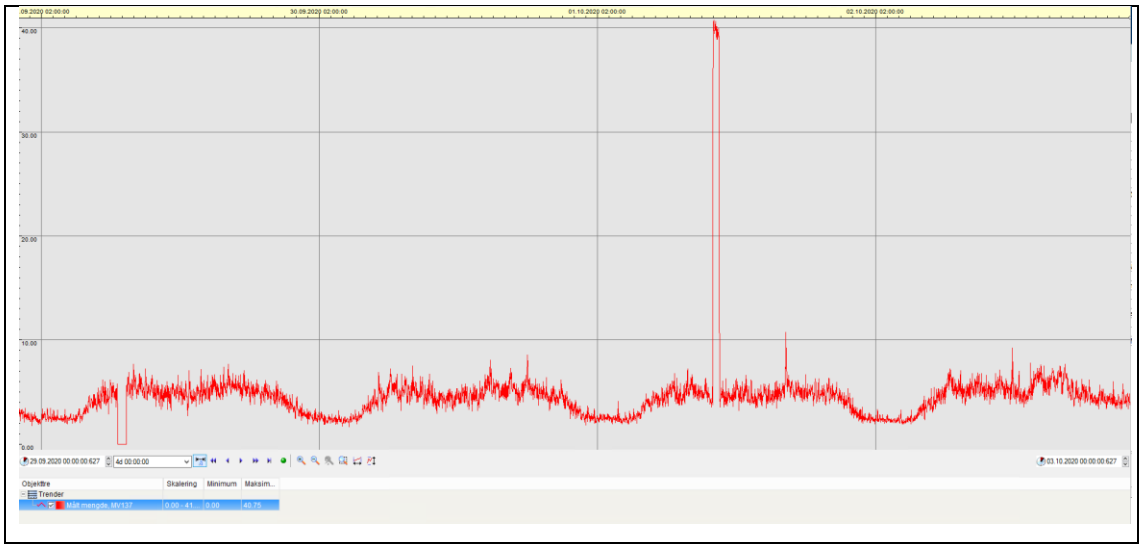


Figure 4-3 Representation of the data flow in guard system

The following table shows the details of the sudden flow changes illustrated in above picture.

Date & time	ms	flow m3	
10/1/2020 11:55	265	5.2710248	
10/1/2020 11:56	265	4.3837456	Sudden changes
10/1/2020 11:57	265	3.9280918	
10/1/2020 11:58	265	11.7644862	
10/1/2020 11:59	265	34.7010598	

Table 4-1 shows the exact date with sudden flow changes in the dataset

4.2.1 Data preparation

From the machine learning perspective, before the mentioned sudden changes to be confirmed as the leakage, and receive the labels, we only knew that, an abnormal occurrence was happening in our distribution system. The above table is pre-processed as follows and we can see the algorithm results voted for anomaly detection even though the answer was clear to be abnormal detection in this case-study from the sudden changes in that successive ordinal numbers.

In this dataset the water flow is recorded in every second, hence we have enough number of rows to rearrange the dataset with respect to minute or hours. Restructuring the table with respect to hours makes more sense in detection of the leakages according to the available water flow data. We calculated the hourly flow of the water by taking the hourly average flow of the rows, hence the number of the rows are not sufficient for more accumulations like daily water flow.

For better prediction the above table is rearranged as follows with some important feature selections like calculating the min, max, mean, median and standard deviation of the corresponding row. For this dataset, we considered water flows with two different conditions in every one and three hours. The final shape of the data frame is shown below in table 4-2.

Year	Month	Hour	Flow	min	max	mean	median	std
2019	12	0	147.2625	1.8625	3.2875	2.454375	2.41875	0.337973
2019	12	1	122.025	1.6	2.5125	2.03375	2.025	0.1852664
2019	12	2	29.925	1.575	2.1125	1.8703125	1.875	0.1504767
2019	12	3	65.2	1.55	2.125	1.8111111	1.78125	0.1487081
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Table 4-2 few rows of data preparation and feature selection of the dataset

We can have two different assumptions dealing with this dataset. If we consider the mentioned dataset as unlabelled set of data, then we must choose anomaly detection approach for the further processing.

4.2.2 k-means clustering in anomaly detection approach

If we consider the unlabelled form of the data, by applying k means clustering algorithm in order to recognize the abnormal behaviour of the mentioned dataset, we will get the

following classification of the Tosshullet data in that 3 days captured data from 29/09/2020 to 03/10/2020 :

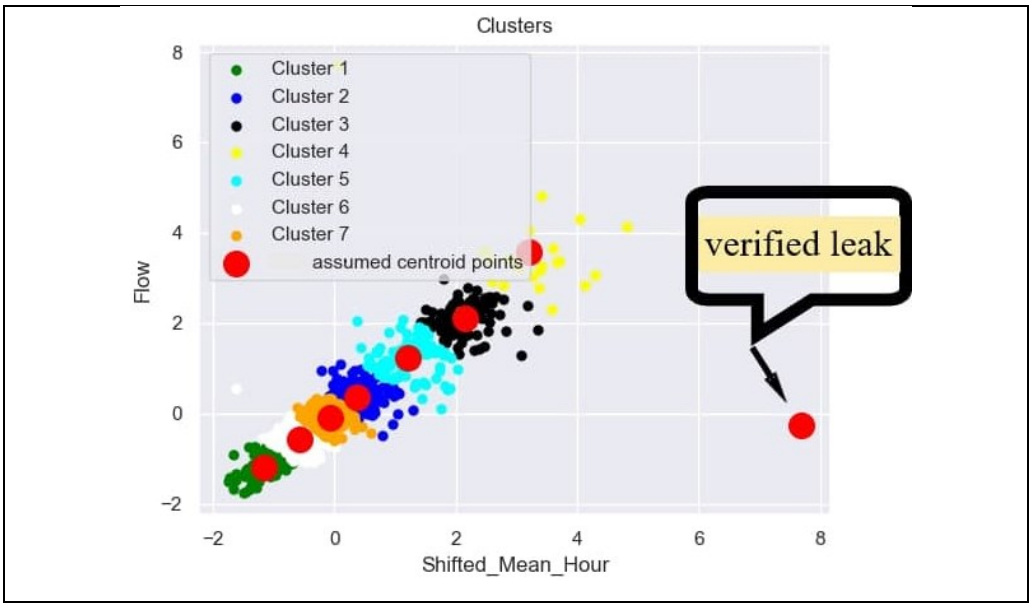


Figure 4-4 K-means clustering result for anomaly detection

According to the above picture different classes of water flow are recognized as different categories of water consumption hours in people's daily life. The classes are represented with different colours in the above picture and clearly the single class separated from the line with abnormal behaviour, is the occurred leakage of the dataset by applying K-means clustering algorithm.

Another assumption is when the mentioned dataset is labelled, as we have the confirmed leak hours from municipality. The leak column is created and added to the dataset with all the sudden flow changes labelled as leakage according to the guard system report while the rest of the rows are left as no leakage.

With this line, we also considered the dataset with respect to every 3 hours water flow for more result comparison. Testing the dataset with/without date column is another test condition to be verified.

In this scenario we considered 8 following possibilities to analyse our dataset with two RF and XGB supervised algorithms for better comparison. Standard scaler is applied to regulate the dataset and cross validation technique is done in order to test all possible

combinations. The following table describes the mentioned combinations and corresponding results:

Every 1-hour flow	Every 3-hours flow	No date	Date with OHE	RF	XGB	%Accuracy
*		*		*		99,43
*		*			*	100
*			*	*		99,28
*			*		*	99,96
	*	*		*		99,90
	*	*			*	99,90
	*		*	*		99,90
	*		*		*	99,90

Table 4-3 Model comparison of different method combinations tested with two supervised learning algorithms

The reason why the results turn out to be too good, is the imbalanced data problem. We have a few numbers of leak labelled occurrences in our dataset against large number of "no leak" labelled rows, and this makes the dataset imbalanced.

4.2.3 hyperparameter tuning optimization algorithms

Even though the results of our experiments are extremely high due to our dataset, but for sake of documentation, the performed result of our hyperparameter tunings with 3-fold cross validation are as follows:

Every 1-hour flow	Every 3-hours flow	No date	Date with OHE	RF	%Accuracy after hyperparameter tuning
*		*		*	99,96
*			*	*	100
	*	*		*	99,90
	*		*	*	99,90

Table 4-4 Hyperparameter tuning optimization results for RF

There are changes in every 1-hour flow with random forest algorithms, which is not considerable hence the primary operation of the algorithm is clear.
Two previous datasets tried to give emphasis to the acoustic sound attributes and flow of the water tubes. As mentioned earlier the more attributes and important features included in the machine learning approaches, the more is the algorithm accuracy and finally better precision. In the following section, we apply different machine learning algorithms on similar case studies to get a better insight view about our project. Different parts of dealing with the datasets in these case studies gives us a better clue about the water leakage detection and the important fact of acoustic sounds.

4.3 Case study I

4.3.1 Third dataset – Yorkshire acoustic logger data

4.3.1.1 Data preparation

After combining all gathered data and preparing a single rearranged file for further processing after removing null values, we have a data table with following attributes:

ID	Date	value_Lvl	value_Spr	Leak Alarm	Leak Found
0	134	25-04	45	6	Y
1	397	16-04	21	5	Y
2	1076	16-04	18	3	Y
3	20	3-Sep	26	8	Y

Table 4-5 few instances of data preparation data frame

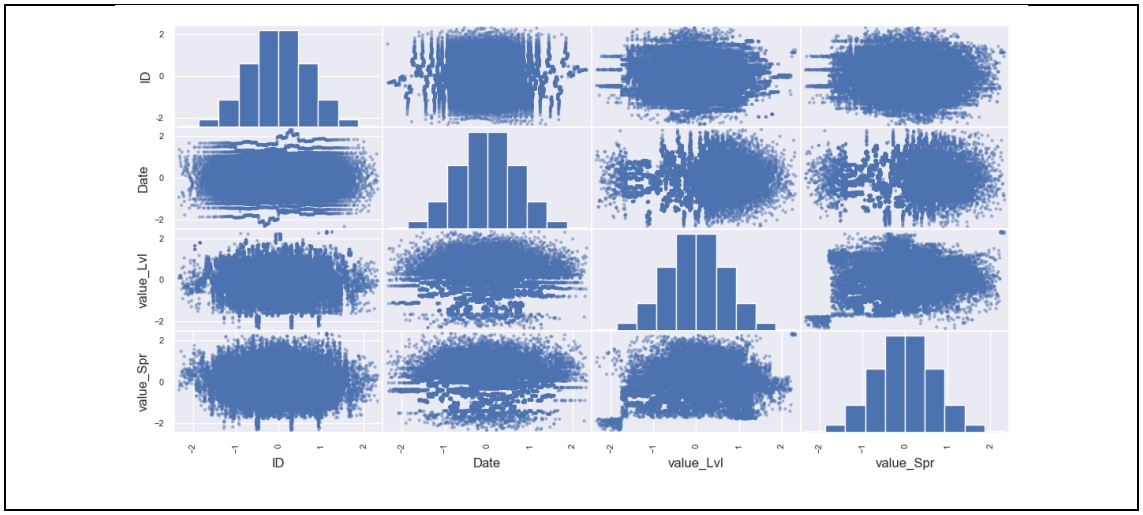
Since two last columns of the dataset is categorical values, we have applied integer encoding to make them understandable for the machine. Both one hot encoding and integer encoding is applicable for the date column to change it to numerical values.

Remember the K-means clustering for the previous dataset. We applied the k-means clustering algorithm on this dataset, to see the dispersion of the available data for “leak” or “no leak” categories. We received the result in figure after applying Gauss Rank scaler, hence applying normalization seemed to be ineffective (Table 4-6) for this dataset.

Roc Score Comparison table for normalization effect on dataset		
Algorithm name	**Before Normalization**	**After Normalization**
Random Forest	0.6750349545590731	0.6914936102236422
XGBoost	0.9883734200690462	0.8576974932587635
KNN	0.5899464172191445	0.6091800949971147
Adaboost Classifier	0.8920137848366796	0.8022279948046758
Bagging Classifier	0.5	0.5

Table 4-6 Normalization effect comparison with four features

By applying scaler with k-means clustering algorithm for the Yorkshire dataset, the dispersion of the available points with respect to the average level and spread level are as follows:



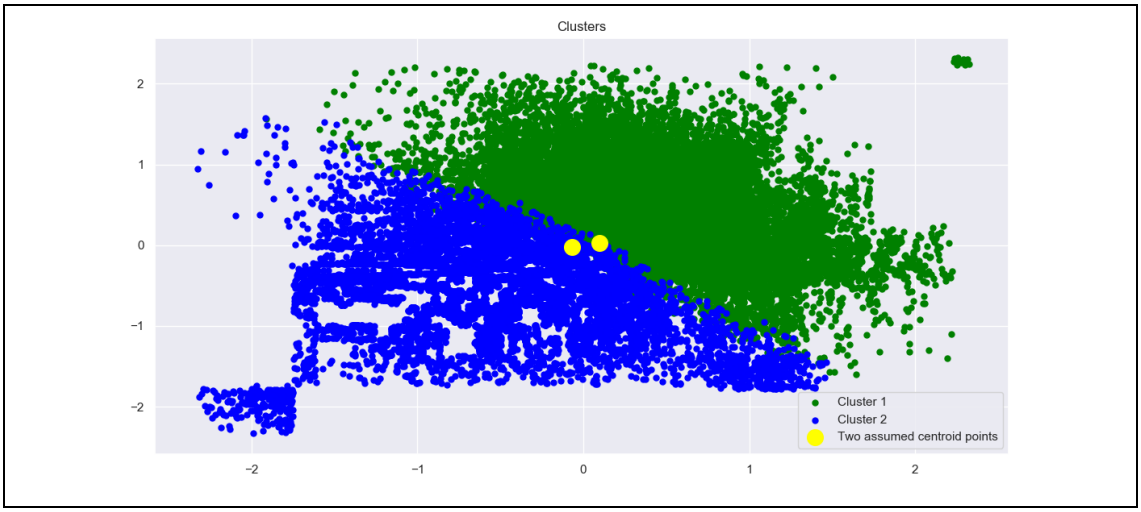


Table 4-7 Gauss Rank scaler with K-means clustering with respect to the average level and spread level of the noise

By applying correlation matrix, it's possible to find out the correlation of the attributes with each other. After elimination of target value and its highly correlated leak alarm attribute, the correlation matrix with 4 main attributes is shown below:

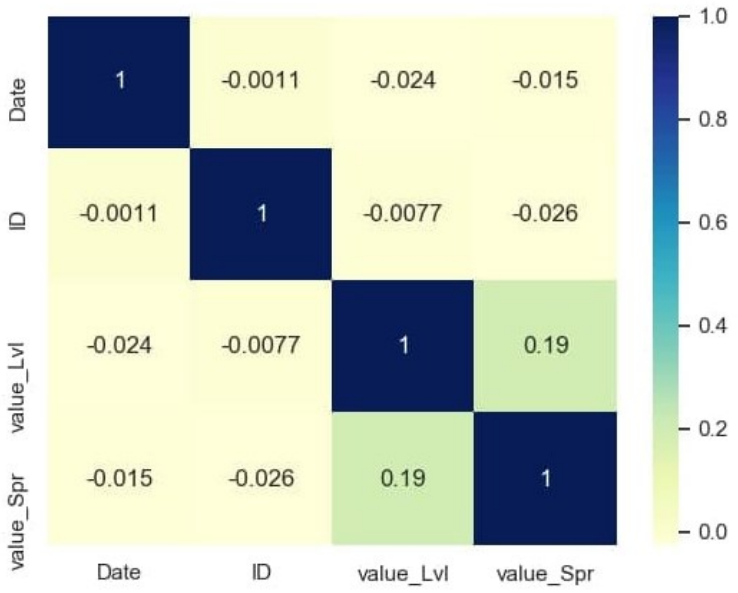


Figure 4-5 Correlation matrix for Yorkshire dataset attributes

In this dataset, after considering the leak found dataset as target value, we get four attributes of ID, Date; Value_Lvl and Value_Spr. we have considered two feature and four feature datasets by applying XGB, RF, KNN, Adaboost and Bagging algorithms.

4.3.1.2 Algorithm comparison

Before performing the algorithms, it's essential here to check the importance of the features again, so by applying the feature importance XGB decision tree algorithm, the importance of the features is shown below:

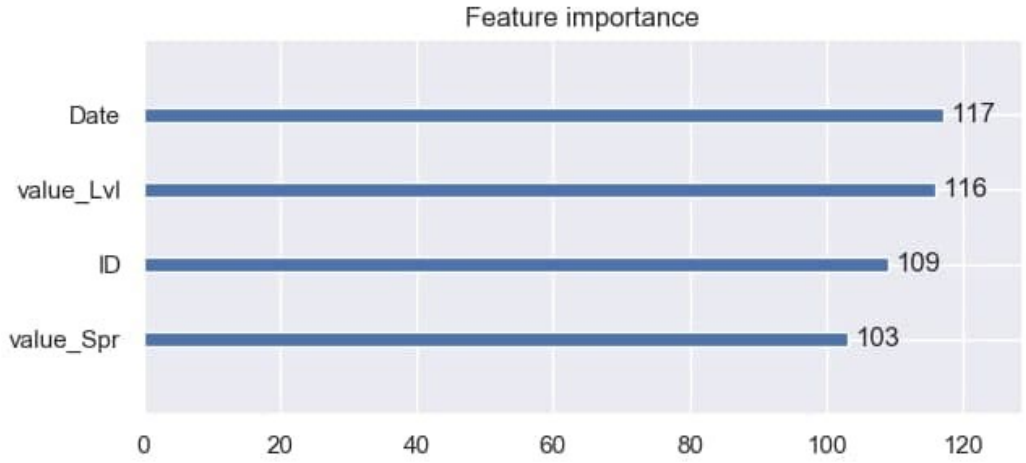


Figure 4-6 Feature importance XGB algorithm

In our first assumption, the result comparison table for the first assumption with corresponding Roc curves with/without considering date and ID column is shown below. The four features are date, ID, average level and spread level of noises.

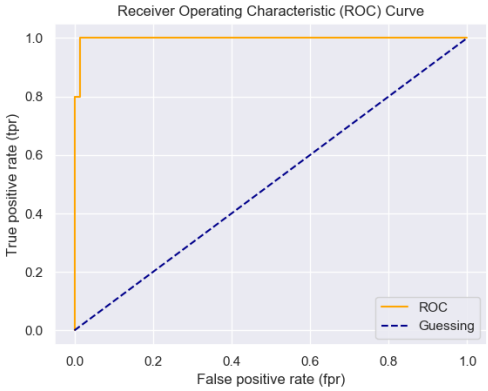
Table Comparison in XGB	
TWO FEATURES	FOUR FEATURES
Accuracy : 99.89% **roc_auc_score : 0.9964**	**Accuracy: 99.91%** **roc_auc_score: 0.9658**
	

Table 4-8 Result comparison of XGB algorithm

Table Comparison in Random forest	
TWO FEATURES	FOUR FEATURES
Accuracy : 99.94% **roc_auc_score : 0.7934**	**Accuracy : 99.88%** **roc_auc_score : 0.6373**
	

Table 4-9 Result comparison of RF algorithm

Table Comparison in KNN	
TWO FEATURES	FOUR FEATURES
Accuracy : 99.94% **roc_auc_score : 0.5986**	**Accuracy : 99.96%** **roc_auc_score : 0.7480**

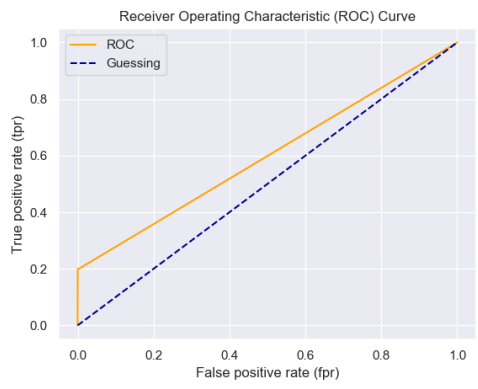
	

Table 4-10 Result comparison of KNN algorithm

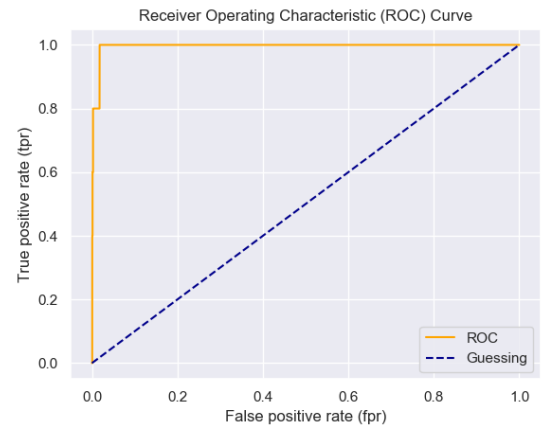
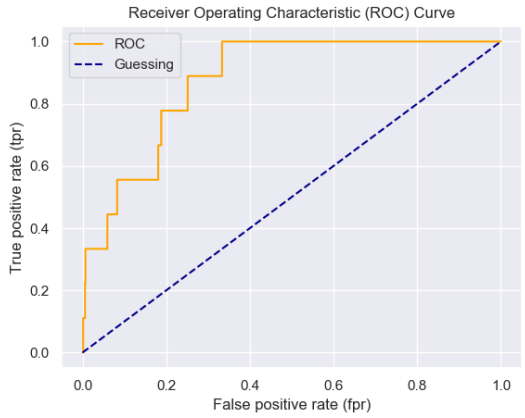
Table Comparison in AdaBoost	
TWO FEATURES	FOUR FEATURES
Accuracy : 99.93% **roc_auc_score : 0.9953**	**Accuracy : 99.89%** **roc_auc_score : 0.8766**
	

Table 4-11 Result comparison of Adaboost algorithm

Table Comparison in Bagging Classifier	
TWO FEATURES	FOUR FEATURES
Accuracy : 99.93% **roc_auc_score : 0.4981**	**Accuracy : 99.93%** **roc_auc_score : 0.4960**

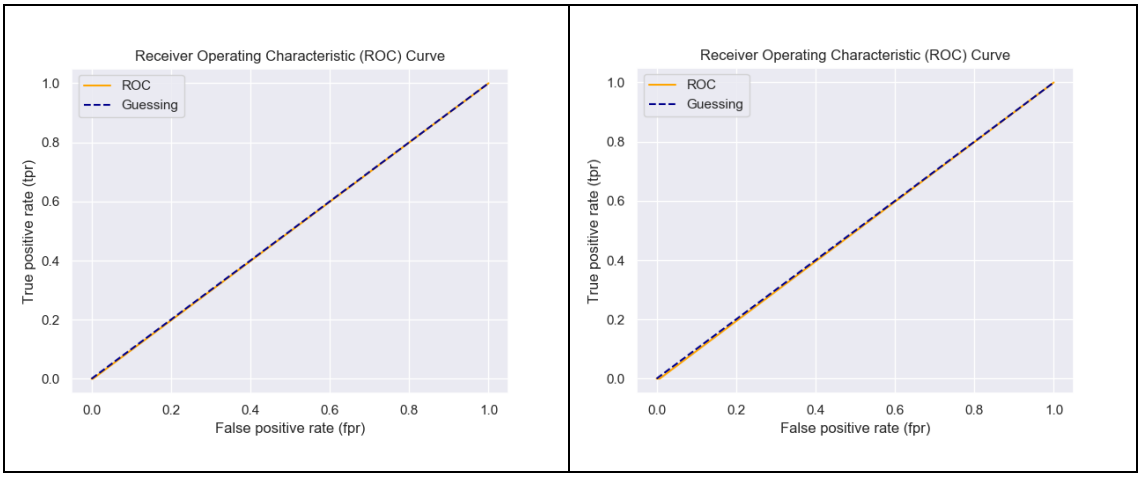


Table 4-12 Result comparison of Bagging algorithm

In this part, except KNN, almost all the algorithms performed better with 2 features dataset.

In our second assumption, if we consider the undefined leak found column as unlabelled then we deal with semi supervised learning and we have the following table:

Algorithm	**No of iterations**	**Test F1**
XGB	8	0,6842
RF	19	0,6315
KNN	8	0,7368
Adaboost	7	0,6315
Bagging	3	0,7105

Table 4-13 testing semi supervised learning with self-training approach on four features dataset

For the above results in above table, we tested the self-training approach in semi supervised learning. In order to utilize the mixed unlabeled and labeled data for classification, first we trained a classifier on small amount of labeled data, and then the classifier itself is used to make predictions on the unlabeled data.

The important part of this method is the evaluation of the algorithm since we don't have a unique labelled data part to refer to it. We have evaluated our algorithm performances with respect to the first labelled training dataset. The table above shows that the KNN algorithm performs better in compare with other algorithms after 8 iteration process.

4.3.1.3 LSTM - Autoencoder Neural networks

LSTM stands for Long Short-Term Memory model and is an artificial recurrent neural network which deals with sequential data. In this method the output of previous step is fed as input to the current step. Since we plan to test the large-scale dataset, not the household consumptions in small scales, we tried to utilize a technique which is the combination of LSTM and Autoencoder.

Basically, autoencoders are an unsupervised technique of learning and it doesn't need labeled data for training section. They generate their own label from the training data, hence called self-supervised learning. Encoder-decoder LSTM approach is for the sequence datasets and used for noise detection as well. The output performance of this algorithm along with standard scaler effect with 500 epochs and appropriate threshold on Yorkshire dataset is shown below:

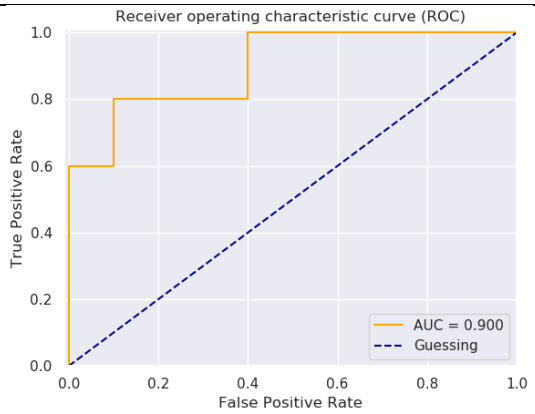
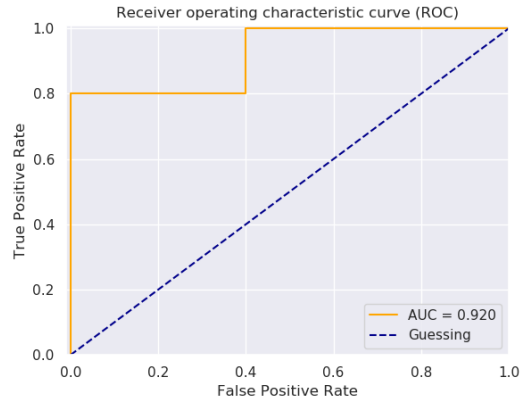
LSTM-Autoencoder algorithm	Area under the curve	Related diagram
Before Normalization With four attributes	AUC: 0,900	
After Normalization With four attributes	AUC: 0,920	

Figure 4-7 LSTM-Autoencoder result with normalization effect

4.4 Case study II

4.4.1 Fourth dataset – MIMII

Malfunctioning industrial machine investigation and inspection dataset is exactly dealing with acoustic noise detection. we have chosen this dataset to test different algorithms to find out important features which are helpful in our project from the acoustic noise detection perspective. We tried to find some similarities between the features of this dataset with average & spread level of the Yorkshire dataset.

4.4.1.1 Pre-processing dataset

After conversion of the pump labeled dataset into numerical values, we have extracted the minimum and maximum value of the amplitude as two columns of our dataset. It's clear that we retrieved mean, median and standard deviation consequently. The labeled column is set with 0 & 1 values as another separate column to discern normal from abnormal waves.

There are 4205 wave instances available in our dataset. All the waves are segregated into the minimum and maximum intensity, and the features like mean, median, and standard deviation is calculated for all the waves. We made the similar features as the previous dataset in Yorkshire project for better comparison.

A reliable method for choosing the best features is applying feature importance strategy with nominated algorithm. Importance of the features is the intensity of relative feature to improve the performance measurement. On the other hand, it verifies that how well each feature can improve the final performance of the algorithm. The result of applying feature importance method with XGB algorithm is shown in the figure below:

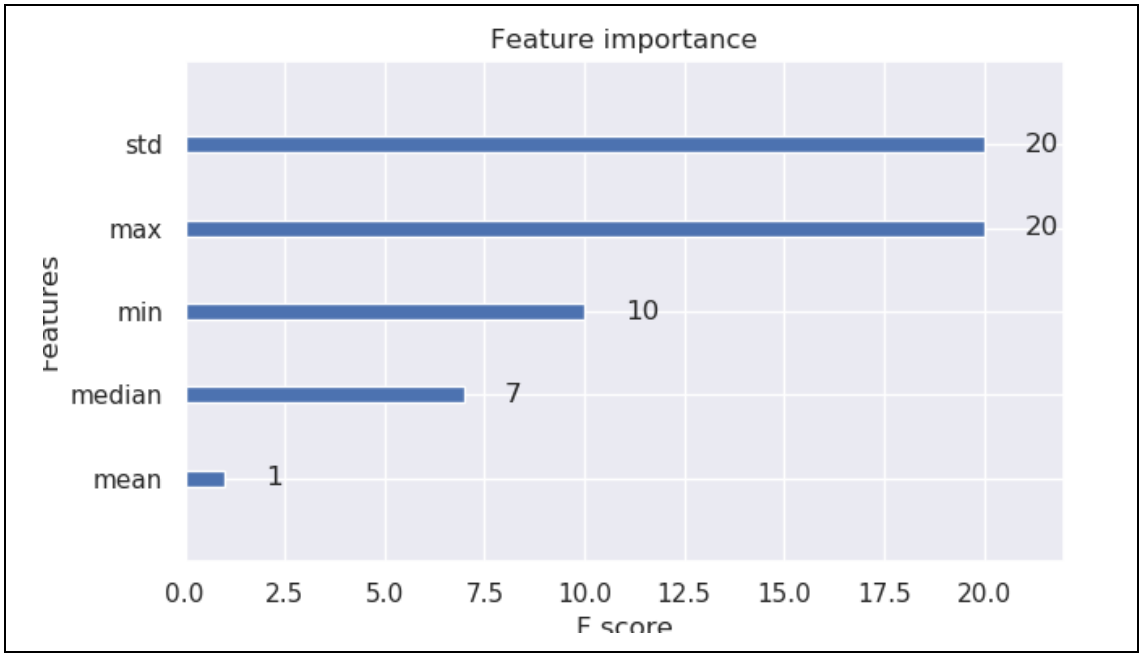


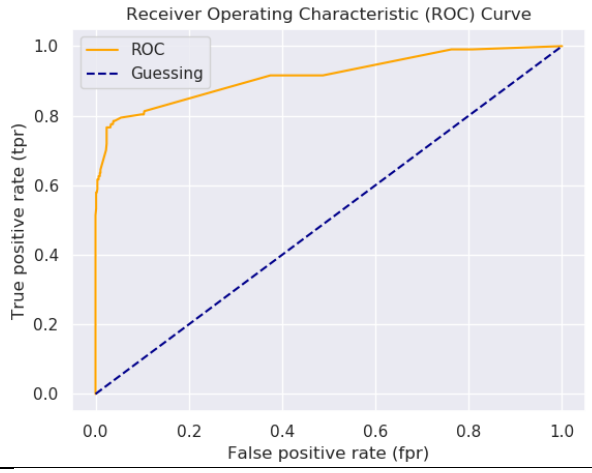
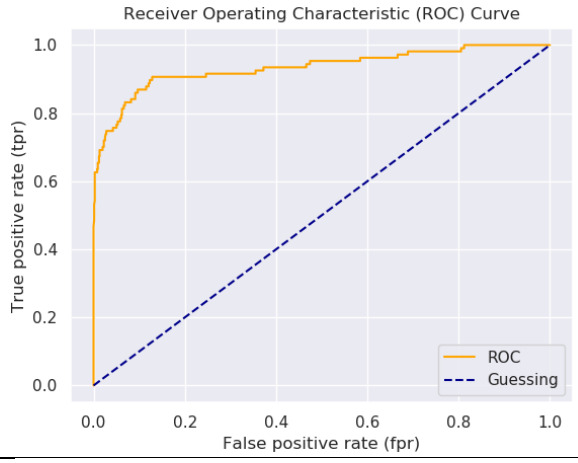
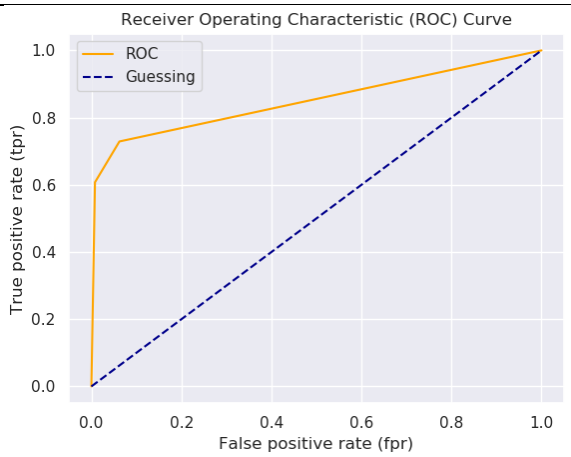
Figure 4-8 Applying feature importance strategy with XGB algorithm

The above figure implies that, features like, maximum, minimum, and standard deviation are likely to be more important from XGB algorithm calculation result.

4.4.1.2 Applying algorithms

In what follows we examine the result of the algorithms with respect to above feature importance. The following table is the result of the mentioned algorithms with Min, Max, Mean, Median and Standard deviation features of the dataset.

XGB	

Accuracy: 94.90% Precision: 0.96 Recall: 0.62 F1 Score: 0.75 AUC: 0.9144	
RF Accuracy: 94.66% Precision: 0.91 Recall: 0.64 F1 Score: 0.75 AUC: 0.93366	
KNN Accuracy: 94.43% Precision: 0.92 Recall: 0.61 F1 Score: 0.73 AUC: 0.8493	
AdaBoost	

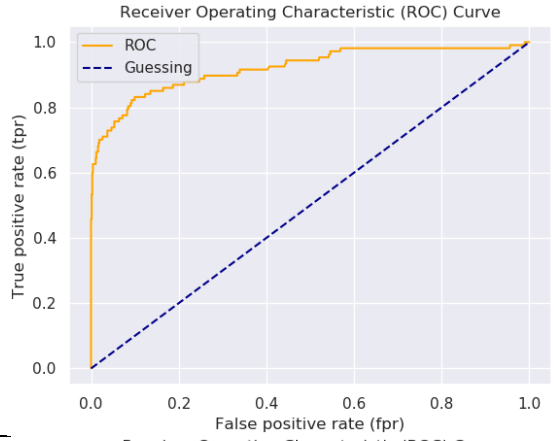
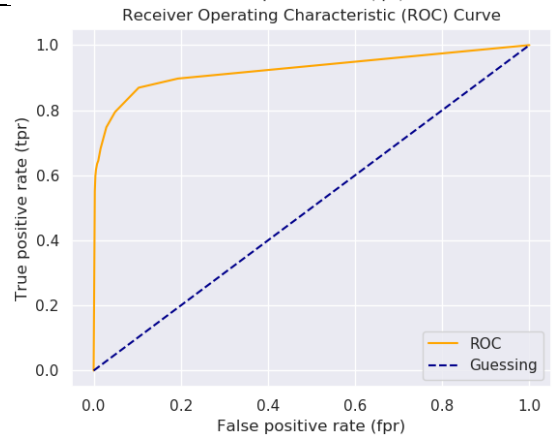
Accuracy: 94.55% Precision: 0.89 Recall: 0.64 F1 Score: 0.74 AUC: 0.9206	
Bagging	
Accuracy: 94.66% Precision: 0.90 Recall: 0.64 F1 Score: 0.75 AUC: 0.9238	

Figure 4-9 Result comparison of the algorithms

Although the received results from the algorithm performances are satisfactory for all the algorithms but Random forest shows the best results among all, with highest recall percentage along with Adaboost and Bagging techniques.

4.4.1.3 Hyperparameter tuning optimization- MIMII result

In applying XGB algorithm, we set different types of parameters as the power of this algorithm is constantly related into choosing proper tuning parameters.

The parameters like, n_estimators, max_depth, learning_rate, subsample, colsample_bytree, booster, eval_metric, verbosity, and n_jobs are selected after hyperparameter tuning with XGB algorithm. The corresponding result is shown in table below:

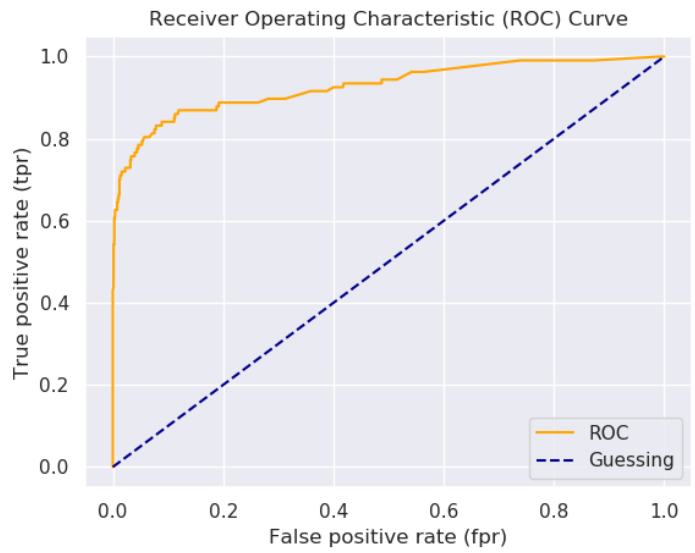
XGB After hyperparameter tuning	
Accuracy: 94.66% Precision: 0.92 Recall: 0.63 F1 Score: 0.74 AUC: 0.9277	

Figure 4-10 Optimization results

The value of the recall says that, how well the model can detect considered classes. In this type of projects, for detection of the unusual events in detection of the leakages, we need to have better value for recall, despite of having high precision and accuracy.

According to the feature importance result, the top 3 important features among the others are identified as, Std, Max and Min. considering these three as the attributes of our dataset, applying all algorithm once again to see the differences we have:

Algorithms	**Features**	**%Accuracy**	**Recall**	**AUC**
XGB	Std-Min-Max	95.01	0.64	0.9331
RF	Std-Min-Max	94.90	0.65	0.9359
KNN	Std-Min-Max	94.55	0.61	0.8673
AdaBoost	Std-Min-Max	94.78	0.64	0.9226
Bagging	Std-Min-Max	94.55	0.66	0.9181

Table 4-14 Algorithm results comparison with three important features

Now if we consider only two Max and Min features and applying the algorithms on them, then we have the following table:

Algorithms	Features	%Accuracy	Recall	AUC
XGB	Min-Max	94.78	0.61	0.8549
RF	Min-Max	94.90	0.63	0.8750
KNN	Min-Max	94.20	0.72	0.8345
AdaBoost	Min-Max	94.90	0.63	0.8111
Bagging	Min-Max	94.43	0.60	0.8442

Table 4-15 Algorithm results comparison with two important features

Except KNN, no further changes in most of the result parameters which shows that we are still dealing with most important attributes. The interesting point about KNN is that, there is a significant improvement in recall which is very important, and the higher recall result claims that our model is more reliable with better true predictions.
Result comparison of Mean and Std, another two available features we get:

Algorithms	Features	%Accuracy	Recall	AUC
XGB	Mean-Std	93.39	0.48	0.8433
RF	Mean-Std	92.23	0.48	0.8192
KNN	Mean-Std	92.23	0.45	0.7238
AdaBoost	Mean-Std	92.81	0.47	0.8388
Bagging	Mean-Std	91.88	0.49	0.8125

Table 4-16 Algorithm results comparison with two derived features

4.4.1.4 LSTM-Autoencoder neural network performance

By looking at the result of the LSTM-Autoencoder with 500 epochs and appropriate threshold, we realize that neural network acquired lower results compare with other algorithms on this dataset. The following figure shows the AUC diagram:

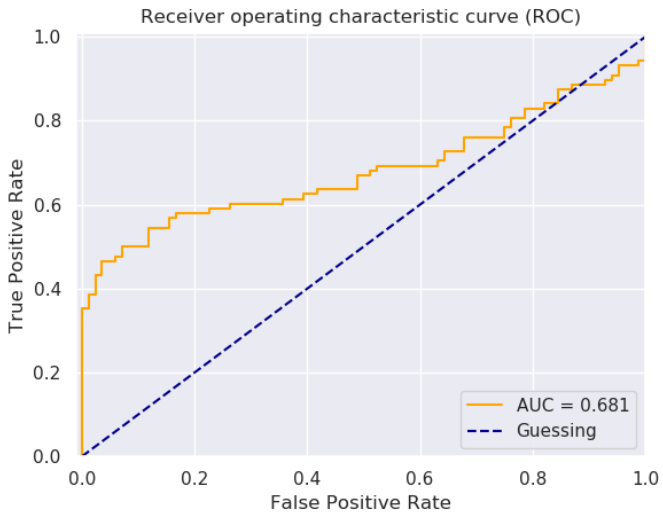
LSTM-Autoencoder	
AUC: 0.6808	

Figure 4-11 LSTM-Autoencoder result

We made several datasets to test our experiments as different sampling strategies can decrease bias in all the events.

Chapter 5 Discussion

Leak detection of the pipelines is one of the important challenges in today's life and water as a fundamental primary material of living things is much worth to give emphasis to this topic. On the other hand, growing technology in different sensors and combination of retrieved data with machine learning techniques, has opened another different chapter in research and market field to offer solutions to the problem.

In this study, we have reported our trial and error attempts in the context of design science methodologies in information system. Therefore, we pointed one of the many nominated solutions to the problem from the earlier researches and case studies in order to investigate the best way of applying machine learning techniques into the data for similar scenarios. Apart from that, we tried to make a point of better data collection in terms of related attributes and important features.

In this defined context, we pushed towards that goal by testing the first nominated acoustic type of sensors along with different approaches of machine learning to estimate the performance of the offered scenario.

In this project we faced different problems in different stages of the project, like sensor installation, data collection and data analysis and we tried to report them well in such a way that could be helpful for further researches.

Back to our research questions mentioned in previous chapters we had:

RQ1: What are the appropriate machine learning methods in acoustic leak detection?

The result of the researches shows that, machine learning techniques are the inseparable part of any detection specially when it come to the huge data monitoring and data collection in large scale projects.

The precise detection of the leakage projects has been started from oil and gas field and extended to the water leak detection as todays one of the most important challenges.

In real time water monitoring systems, the role of the anomaly detection method is clearly impressive in compare to the traditional approach leak detection to identifying anomalies in domestic usage by referring to the water meter at the end of the month [2].

The anomaly detection of machine learning approach in real time scenarios performs more effective when it comes to domestic water usage monitoring. It detects the abnormal usage for each household and the suspected area can be verified immediately right after the detection. On the other hand, it helps the system in large scales for pipe burst occurrences and not small leakages.

If we put together all parts of different experiments and approaches performed in this project, we simply realized that essentially from the general aspect, applying machine learning techniques on a specific set of data, requires some knowledge about that data and fundamentally depends on the type of the data we deal with.

Sometimes in case of unlabelled and labelled datasets, we know from the theory that unsupervised and supervised learning is the techniques that must be applied on the corresponding dataset respectively, but the result of these two are completely different when we can't evaluate our result in case of unsupervised type of learning, hence we don't have true labelled dataset to refer them.

In this project, especially in our third dataset, we achieved the results with high variations in each attempt in case of semi supervised learning with unacceptably low performance. The reason is that, we had few labelled instances and that makes the dataset imbalanced. In case of imbalanced datasets, it is true that sampling techniques can change the entire output in case of supervised and semi supervised learning in our datasets. Stratified type of sampling proved to be a promising method for that purpose as declared earlier in some cases.

Dealing with all datasets, first we tried to apply k-means clustering to find out best possible data classification of the dataset to have histogram of the data with the classified ranges. Among all the normalization strategies on our dataset, the effect of Gauss Rank normalization is significantly different in our third dataset. We got the information that our collected dataset in Yorkshire dataset is truly involves with 2 different classes as leakage and no leakage points with respect to the noise data.

The reason we didn't try standardization on the Yorkshire dataset is that, from the theoretical aspect by checking the data histogram, the dataset does not follow the Gaussian distribution. So, we applied normalization instead of standardization and the theory is confirmed once again with significant result by applying the normalization.
As shown in the evaluation section, we tried to apply 5 different algorithms. XGB, RF, KNN, Adaboost and Bagging are the nominated algorithms for result comparison. XGB and Adaboost, from the boosting algorithms performed better in compared with the other techniques specially after normalization techniques of pre-processing section. Even though all our algorithms performed well in our first assumption in Yorkshire dataset but the highest performance after hyperparameter tuning optimization is for the extreme gradient boosting (XGB).
It may implies by checking the comparison results from the first assumption in Yorkshire dataset, that the outcomes are too good to be true, but that's not the case and the dataset is imbalanced from the few number of the leakage instances vs large number of "no leakage" events.
In our fourth dataset, dealing with noise components, the XGB algorithms could not perform better than RF, even after the hyperparameter tuning optimization. This shows the rigidness of the Random Forest algorithm which is a suitable choice for datasets without normalization and less attributes.
Although our LSTM-Autoencoder neural network didn't performed like others on this dataset, but we can verify the normalization effect on it from the AUC diagrams showed earlier. Neural network techniques give a better result when dealing with more attributes, especially in some cases of unusual flows in datasets with proper balanced instances.

RQ2: Which attributes are playing important role in detection of the leakages in urban water pipeline data analysis?

The important part of the study is focused on the set of attributes and their impact on the corresponding algorithm performance. It is tried to verify the feature importance strategies in different case studies to collect the suitable features for the leakage detection project scenarios while the second dataset in our project, is a real evidence of the water flow attribute collection for anomaly detection in monitoring water distribution systems.

Although applying feature importance algorithms helps to identify the effective attributes of the related dataset having more impact on the algorithm performance, but the importance of these attributes can be changed by small changes in the datasets on that specific period. This is the case specially when the numbers of the features are less.

Considering the comparison algorithm result between two and four features, it implies, even though the date and noise average level was identified from the feature importance algorithm as the most effective attributes, but with elimination of Date and ID feature in two features test results, we received better algorithm outcome performances.

On the other hand, elimination of standard deviation among three features like, Max, Min and Std which is calculated from the other features and is added to the dataset in feature selection process, effects the overall performance represented in the Table 4-15.

Table 4-16 shows another comparison with two derived features and the algorithm performance is significantly lower in compare with the Min & Max features algorithm performances.

The maximum and minimum intensity of the acoustic noises are represented as the main features of the acoustic datasets. Comparing these two features with the main data attributes of the Yorkshire dataset as average level and spread level of the noises, conveys the message that, more feature selection from the noise parameters are extremely helpful technique in acoustic analysis of the data.

It also implies that, other attributes like Date, different IDs, location, and time are attributes with less effect in classification project scenarios.

The anomaly detection techniques required more direct data gathering from the sensors and the flow of the water for reliable performances, especially in real time detection scenarios.

Chapter 6 Conclusion and future work

The leak detection strategies in water transportation pipeline networks is an essential research topic in today's life. It helps to avoid the challenges in different fields like public health, resources wastages, agricultural deficiency parameters and financial problems.
The acoustic signals collected by the listening devices for detection of the leakages in buried water pipelines is one of the many nominated techniques to deal with this problem and Norway like other countries, is seeking for the solutions in most of the country municipalities.
In this study, from the methodological perspective, working with two different datasets as our case studies, we examined different machine learning approaches and the results are reported. Towards the mentioned experiments, we tried to identify the best possible helpful features from the available case study datasets for our project. The project is started in institute for energy technology (IFE) with noise data collection as the first nominated solution experiment to the problem.
In this thesis, the machine learning approaches examined on several datasets from unsupervised, supervised and semi supervised approaches with respect to the available datasets. Decision tree algorithms like XGB and RF show promising values on water flow dataset which we received from the municipality guard system on specific period, while applying k-means clustering in an unsupervised learning method, detects the anomaly behaviour of the water flow.
Applying decision tree algorithm techniques on two case studies, one with acoustic leak detection dataset and other with noise analysis of industrial machine malfunctioning noises, shows better performances in decision tree algorithms in compare with the neural network approaches.
LSTM-Autoencoder neural network, which is counted as ensemble learning type of neural network, used for noise detection analysis, needs some more data features to compete with boosting techniques like XGB and Adaboost.

Testing the algorithm performances with different features, after applying feature selection techniques, shows that the feature engineering and feature selection techniques are extremely useful techniques in noise classification method analysis.

Result comparison tables of algorithm performances for the same dataset in chapter 4 with 2 features, 3 features and 5 features shows, even though the attributes like minimum and maximum noise intensity are the two main attributes of the dataset, but the derived attributes from these two can also play a crucial role in the algorithm performances.

Combination of the water flow attributes with the acoustic attributes can give us a better precise leak localization result. We will combine all the received features from these two aspects, data flow feature details and acoustic features, to have a better prediction and detection of the water leakages in our future work.

Bibliography

[1] P. Aramane, A. Bhattad, and N. Aithal, “Iot and Neural Network Based Multi Region and,” vol. 10, no. 6, pp. 61–68, 2019.

[2] I. S. Herath, “Smart Water Buddy: IoT based Intelligent Domestic Water Management System,” in *2019 International Conference on Advancements in Computing (ICAC)*, Dec. 2019, pp. 380–385. doi: 10.1109/ICAC49085.2019.9103379.

[3] T. Ravichandran, K. Gavahi, K. Ponnambalam, V. Burtea, and S. J. Mousavi, “Ensemble-based machine learning approach for improved leak detection in water mains,” *J. Hydroinformatics*, pp. 1–17, 2021, doi: 10.2166/hydro.2021.093.

[4] S. W. Oh, D. B. Yoon, G. J. Kim, J. H. Bae, and H. S. Kim, “Acoustic data condensation to enhance pipeline leak detection,” *Nucl. Eng. Des.*, vol. 327, pp. 198–211, Feb. 2018, doi: 10.1016/j.nucengdes.2017.12.006.

[5] M. Fagiani, S. Squartini, L. Gabrielli, M. Severini, and F. Piazza, “A Statistical Framework for Automatic Leakage Detection in Smart Water and Gas Grids,” *Energies*, vol. 9, no. 9, p. 665, Aug. 2016, doi: 10.3390/en9090665.

[6] R. P. da Cruz, F. V. da Silva, and A. M. F. Fileti, “Machine learning and acoustic method applied to leak detection and location in low-pressure gas pipelines,” *Clean Technol. Environ. Policy*, vol. 22, no. 3, pp. 627–638, Apr. 2020, doi: 10.1007/s10098-019-01805-x.

[7] V. Marra, “MULTIPHYSICS ANALYSIS ADVANCES WATER MAIN LEAK DETECTION,” p. 3, 2017.

[8] W. Chalgham, A. Seibi, and F. Boukadi, *Simulation of Leak Noise Propagation and Detection Using COMSOL Multiphysics*. 2016. doi: 10.1115/IMECE2016-68163.

[9] H. Purohit *et al.*, “MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection,” *ArXiv190909347 Cs Eess Stat*, Sep. 2019, Accessed: Apr. 20, 2021. [Online]. Available: http://arxiv.org/abs/1909.09347

[10] A. M. Sadeghioon, N. Metje, D. N. Chapman, and C. J. Anthony, “SmartPipes: Smart Wireless Sensor Networks for Leak Detection in Water Pipelines,” *J. Sens. Actuator Netw.*, vol. 3, no. 1, Art. no. 1, Mar. 2014, doi: 10.3390/jsan3010064.

[11] R. Pérez, V. Puig, J. Pascual, A. Peralta, E. Landeros, and Ll. Jordanas, “Pressure sensor distribution for leak detection in Barcelona water distribution network,” *Water Supply*, vol. 9, no. 6, pp. 715–721, Dec. 2009, doi: 10.2166/ws.2009.372.

[12] G. S. Galloway, V. M. Catterson, T. Fay, A. Robb, and C. Love, “Diagnosis of tidal turbine vibration data through deep neural networks,” in *Proceedings of the Third European Conference of the Prognostics and Health Management Society 2016*, I. Eballard and A. Bregon, Eds. ESP: PHM Society, 2016, pp. 172–180. Accessed: Apr. 27, 2021. [Online]. Available: https://strathprints.strath.ac.uk/57127/

[13] P. Chumchu, “A Leak Detection in Water Pipelines Using Discrete Wavelet Decomposition and Artificial Neural Network,” in *Advances in Signal Processing and Intelligent Recognition Systems*, Singapore, 2021, pp. 49–65. doi: 10.1007/978-981-16-0425-6_4.

[14] "Norsk Vann - Reporting of mini treatment plants - need for clarification." https://www.norskvann.no/index.php/10-nyheter/1481-rapportering-av-minirenseanlegg-behov-for-presisering (accessed May 03, 2021).
[15] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.
[16] K. Potdar, T. Pardawala, and C. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *Int. J. Comput. Appl.*, vol. 175, pp. 7–9, Oct. 2017, doi: 10.5120/ijca2017915495.
[17] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.
[18] Z. Reitermanova, "Data splitting," in *WDS*, 2010, vol. 10, pp. 31–36.
[19] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification," *Remote Sens.*, vol. 11, no. 2, Art. no. 2, Jan. 2019, doi: 10.3390/rs11020185.
[20] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.
[21] J. L. Bacca Acosta, S. M. Baldiris Navarro, R. Fabregat Gesa, S. Graf, and Kinshuk, "Augmented Reality Trends in Education: A Systematic Review of Research and Applications," Oct. 2014, Accessed: May 10, 2021. [Online]. Available: https://dugi-doc.udg.edu/handle/10256/17763
[22] D. Zaman, M. K. Tiwari, A. K. Gupta, and D. Sen, "A review of leakage detection strategies for pressurised pipeline in steady-state," *Eng. Fail. Anal.*, vol. 109, p. 104264, Jan. 2020, doi: 10.1016/j.engfailanal.2019.104264.
[23] H. Luo, K. Wu, R. Ruby, F. Hong, Z. Guo, and L. M. Ni, "Simulation and Experimentation Platforms for Underwater Acoustic Sensor Networks: Advancements and Challenges," *ACM Comput. Surv.*, vol. 50, no. 2, p. 28:1-28:44, May 2017, doi: 10.1145/3040990.
[24] S. Seyoum, L. Alfonso, S. J. van Andel, W. Koole, A. Groenewegen, and N. van de Giesen, "A Shazam-like Household Water Leakage Detection Method," *Procedia Eng.*, vol. 186, pp. 452–459, Jan. 2017, doi: 10.1016/j.proeng.2017.03.253.
[25] A. Pal and K. Kant, "Water Flow Driven Sensor Networks for Leakage and Contamination Monitoring in Distribution Pipelines," *ACM Trans. Sens. Netw.*, vol. 15, no. 4, p. 37:1-37:43, Aug. 2019, doi: 10.1145/3342513.
[26] J. Wang *et al.*, "Smart Water Lora IoT System," in *Proceedings of the 2018 International Conference on Communication Engineering and Technology*, New York, NY, USA, Feb. 2018, pp. 48–51. doi: 10.1145/3194244.3194260.
[27] Y. Kim, T. Schmid, Z. M. Charbiwala, J. Friedman, and M. B. Srivastava, "NAWMS: nonintrusive autonomous water monitoring system," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, New York, NY, USA, Nov. 2008, pp. 309–322. doi: 10.1145/1460412.1460443.
[28] M. Y. Aalsalem, W. Z. Khan, W. Gharibi, M. K. Khan, and Q. Arshad, "Wireless Sensor Networks in oil and gas industry: Recent advances, taxonomy, requirements,

and open challenges," *J. Netw. Comput. Appl.*, vol. 113, pp. 87–97, Jul. 2018, doi: 10.1016/j.jnca.2018.04.004.

[29] Y. Zhong *et al.*, "Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 207–222, May 2019, doi: 10.1016/j.isprsjprs.2019.02.021.

[30] S. Tariq, Z. Hu, and T. Zayed, "Micro-electromechanical systems-based technologies for leak detection and localization in water supply networks: A bibliometric and systematic review," *J. Clean. Prod.*, vol. 289, p. 125751, Mar. 2021, doi: 10.1016/j.jclepro.2020.125751.

[31] W. Zhang, P. Wang, K. Sun, C. Wang, and D. Diao, "Intelligently detecting and identifying liquids leakage combining triboelectric nanogenerator based self-powered sensor with machine learning," *Nano Energy*, vol. 56, pp. 277–285, Feb. 2019, doi: 10.1016/j.nanoen.2018.11.058.

[32] L. Berardi, D. B. Laucelli, A. Simone, G. Mazzolani, and O. Giustolisi, "Active Leakage Control with WDNetXL," *Procedia Eng.*, vol. 154, pp. 62–70, Jan. 2016, doi: 10.1016/j.proeng.2016.07.420.

[33] S. Pandya and H. Ghayvat, "Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence," *Adv. Eng. Inform.*, vol. 47, p. 101238, Jan. 2021, doi: 10.1016/j.aei.2020.101238.

[34] D. I. Hefft and F. Alberini, "A step towards the live identification of pipe obstructions with the use of passive acoustic emission and supervised machine learning," *Biosyst. Eng.*, vol. 191, pp. 48–59, Mar. 2020, doi: 10.1016/j.biosystemseng.2019.12.015.

[35] I. Stoianov, L. Nachman, S. Madden, and T. Tokmouline, "PIPENETa wireless sensor network for pipeline monitoring," in *Proceedings of the 6th international conference on Information processing in sensor networks*, New York, NY, USA, Apr. 2007, pp. 264–273. doi: 10.1145/1236360.1236396.

[36] W.-Y. Chuang, Y.-L. Tsai, and L.-H. Wang, "Leak Detection in Water Distribution Pipes Based on CNN with Mel Frequency Cepstral Coefficients," in *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, New York, NY, USA, Mar. 2019, pp. 83–86. doi: 10.1145/3319921.3319926.

[37] N. K. Banjara, S. Sasmal, and S. Voggu, "Machine learning supported acoustic emission technique for leakage detection in pipelines," *Int. J. Press. Vessels Pip.*, vol. 188, p. 104243, Dec. 2020, doi: 10.1016/j.ijpvp.2020.104243.

[38] S. Moulik, S. Majumdar, V. Pal, and Y. Thakran, "Water Leakage Detection in Hilly Region PVC Pipes using Wireless Sensors and Machine Learning," in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, Sep. 2020, pp. 1–2. doi: 10.1109/ICCE-Taiwan49838.2020.9258144.

[39] M. Ahadi and M. S. Bakhtiar, "Leak detection in water-filled plastic pipes through the application of tuned wavelet transforms to Acoustic Emission signals," *Appl. Acoust.*, vol. 71, no. 7, pp. 634–639, Jul. 2010, doi: 10.1016/j.apacoust.2010.02.006.

[40] S. El-Zahab, E. Mohammed Abdelkader, and T. Zayed, "An accelerometer-based leak detection system," *Mech. Syst. Signal Process.*, vol. 108, pp. 276–291, Aug. 2018, doi: 10.1016/j.ymssp.2018.02.030.

[41] J. Merta and J. Fikejz, "Utilization of Machine Learning to Detect Sudden Water Leakage for Smart Water Meter," in *2019 29th International Conference*

Radioelektronika (RADIOELEKTRONIKA), Apr. 2019, pp. 1–5. doi: 10.1109/RADIOELEK.2019.8733447.

[42] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.

[43] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[44] J. Eekels and N. F. M. Roozenburg, "A methodological comparison of the structures of scientific research and engineering design: their similarities and differences," *Des. Stud.*, vol. 12, no. 4, pp. 197–203, Oct. 1991, doi: 10.1016/0142-694X(91)90031-Q.

[45] H. Muccini and K. Vaidhyanathan, "Software Architecture for ML-based Systems: What Exists and What Lies Ahead," *ArXiv210307950 Cs*, Mar. 2021, Accessed: May 16, 2021. [Online]. Available: http://arxiv.org/abs/2103.07950

[46] B. Tang, "Data Collection and Feature Extraction for Machine Learning," *Medium*, Nov. 15, 2019. https://medium.com/ai%C2%B3-theory-practice-business/data-collection-and-feature-extraction-for-machine-learning-98f976401378 (accessed May 16, 2021).

[47] D. Steen, "A Gentle Introduction to Self-Training and Semi-Supervised Learning," *Medium*, Aug. 31, 2020. https://towardsdatascience.com/a-gentle-introduction-to-self-training-and-semi-supervised-learning-ceee73178b38 (accessed May 25, 2021).

[48] J. Brownlee, "Semi-Supervised Learning With Label Propagation," *Machine Learning Mastery*, Dec. 29, 2020. https://machinelearningmastery.com/semi-supervised-learning-with-label-propagation/ (accessed May 25, 2021).

[49] P. Heckbert, "Fourier Transforms and the Fast Fourier Transform (FFT) Algorithm," p. 13.

[50] "1. Supervised learning — scikit-learn 0.24.2 documentation." https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed May 24, 2021).

[51] "NumPy Documentation." https://numpy.org/doc/ (accessed May 24, 2021).

Appendix A Abbreviations

STD standard deviation

Min minimum

Max maximum

RF random forest algorithm

XGB eXtreme Gradient Boosting

KNN K-nearest neighbours' algorithm

Appendix B Pre-processing codes (MIMII)

Attaching all the source codes is too bulky and not necessary to be attached.

```
########################################################################
# import default python-library
########################################################################
import pickle
import os
import sys
import glob
from tqdm import tqdm
########################################################################


########################################################################
# import additional python-library
########################################################################
import numpy
import pandas as pd
import librosa
import librosa.core
import librosa.feature
import yaml
import logging
# from import
from tqdm import tqdm
from sklearn import metrics
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
import librosa.display
import matplotlib.pyplot as plt
########################################################################

n_mels = 64
frames = 5
n_fft = 1024
hop_length = 512
power = 2.0
dims = n_mels * frames

"""
Standard output is logged in "baseline.log".
"""
logging.basicConfig(level=logging.DEBUG, filename="baseline.log")
logger = logging.getLogger(' ')
handler = logging.StreamHandler()
formatter = logging.Formatter('%(asctime)s - %(levelname)s - %(message)s')
handler.setFormatter(formatter)
logger.addHandler(handler)
```

```
# normal_files = sorted(glob.glob(".\\data\\normal\\*.wav"))
normal_files =
sorted(glob.glob("/home/mohammed/separated_dataset/db+6_id_04/normal/*.wav"))

normal_labels = numpy.zeros(len(normal_files))
if len(normal_files) == 0:
    logger.exception("no_wav_data!!")

# 02 abnormal list generate
abnormal_files =
sorted(glob.glob("/home/mohammed/separated_dataset/db+6_id_04/abnormal/*.wav"
))

abnormal_labels = numpy.ones(len(abnormal_files))

if len(abnormal_files) == 0:
    logger.exception("no_wav_data!!")

def datset_constructor(dataset):
    df = pd.DataFrame()
    df["min"] = dataset.min(axis=1)
    df["max"] = dataset.max(axis=1)
    df["mean"] = dataset.mean(axis=1)
    df["median"] = dataset.median(axis=1)
    df["quantile1"] = dataset.quantile(0.25)
    df["quantile2"] = dataset.quantile(0.5)
    df["quantile3"] = dataset.quantile(0.75)
    df["std"] = dataset.std(axis=1)
    df = df.reset_index()
    df.drop(["index"], axis=1, inplace=True)
    return df

train_files = normal_files[:]
y_train = normal_labels[:]
test_files = abnormal_files[:]
y_test = abnormal_labels[:]
# print("normal label shape : ", y_test.shape)

i = 0
df_train = pd.DataFrame()
df_test = pd.DataFrame()
for idx in range(len(train_files)):
    try:
        multi_channel_data, sr = librosa.load(train_files[idx], sr=None,
mono=True)

        if i != 0:
df1 = pd.DataFrame(multi_channel_data.reshape(1, -1))
    df_train = df_train.append(df1)

else:
    df_train = pd.DataFrame(data=multi_channel_data.reshape(1, -1))
    i = i + 1
except ValueError as msg:
    logger.warning(f'{msg}')
x_train = df_train.reset_index()
```

```
n_result = pd.concat([n_x_dataset, y_dataset], axis=1)
print(" result : \n", result)
print(" result_abs : \n", result_abs)
print(" n_result : \n", n_result)
result.to_csv("/home/mohammed/result/separated_dataset/result_db+6_id_04.csv"
, index=True)
n_result.to_csv("/home/mohammed/result/separated_dataset/result_Normalized_db
+6_id_04.csv", index=True)

data_dict = {
    "x_dataset": x_dataset,
    "y_dataset": y_dataset,
    "result": result,
    "result_abs": result_abs,
    "n_result": n_result,
}
print(data_dict.keys())
f_t_write =
open('/home/mohammed/separated_pickles/preprocessed_dataset_db+6_id_04.pickle
', "wb")
pickle.dump(data_dict, f_t_write)
f_t_write.close()
# return datat_dict
```

Appendix C MIMII XGB source code

Attaching all the source codes is too bulky and not necessary to be attached.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from xgboost import plot_tree
import pickle
import time
from matplotlib import dates as mpl_dates
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.metrics import plot_confusion_matrix
import sys
import evaluator
import hyperparameter_tuner
from sklearn.metrics import roc_curve
import seaborn as sns
import xgboost as xgb
from sklearn import metrics
from xgboost import XGBClassifier

file_to_read =
open('/home/mohammed/pickle/preprocessed_dataset_id_00.pickle', "rb")
loaded_object = pickle.load(file_to_read)
file_to_read.close()
dataset = loaded_object
result = dataset["result"]
n_result = dataset["n_result"]
print("result : \n", result)
result.to_csv('result.csv')
print("n_result : \n", n_result)
print("result shape : \n", result.shape)

y_dataset = result.loc[:, ["label"]]
x_dataset = result.drop(["label"], axis=1)
n_x_dataset = n_result.drop(["label"], axis=1)
max_min_mean_median_std = n_x_dataset.drop(['quantile1', 'quantile2',
'quantile3'], axis=1)
max_min_mean = n_x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'median', 'std'], axis=1)
max_min_mean_std = n_x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'median'], axis=1)
max_min_mean_median = x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'std'], axis=1)
```

```
mean_median_std = x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'min', 'max'], axis=1)
min_max = x_dataset.drop(['quantile1', 'quantile2', 'quantile3', 'median',
'std', 'mean'], axis=1)
mean_std = x_dataset.drop(['quantile1', 'quantile2', 'quantile3', 'median',
'min', 'max'], axis=1)
median_std = x_dataset.drop(['quantile1', 'quantile2', 'quantile3', 'mean',
'min', 'max'], axis=1)
min_max_median = x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'std', 'mean'], axis=1)
min_max_std = n_x_dataset.drop(['quantile1', 'quantile2', 'quantile3',
'median', 'mean'], axis=1)



x_train, x_test, y_train, y_test = train_test_split(min_max_std,
                                                    y_dataset,
                                                    test_size=0.25,
                                                    shuffle=True,
                                                    stratify=y_dataset,
                                                    random_state=42)

clf = xgb.sklearn.XGBClassifier(n_estimators=20,
                                max_depth=8,
                                learning_rate=0.1,
                                subsample=0.9,
                                colsample_bytree=0.9,
                                booster="gbtree",
                                eval_metric="map",
                                verbosity= 0,
                                n_jobs= -1)
# clf = xgb.sklearn.XGBClassifier(n_estimators=50)
# params = hyperparameter_tuner.xgb_hyperparameter_tuner(clf, x_train,
y_train)
# clf.set_params(**params)
clf.fit(x_train, y_train)
# xgb_pred = clf.predict(x_test)

evaluator.evaluate_preds(clf, x_train, y_train, x_test, y_test)

plt.figure(figsize=(7, 4))
xgb.plot_importance(clf, ax=plt.gca())
plt.show()
plt.show()
```