Universitetet
i Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Risk Management/ Risk Management	Spring semester, 2018 Open / Restricted access
Writer: Stiffi Zukhrufany	……………………………………… (Writer's signature)

Faculty supervisor:
Prof. Eirik.B. Abrahamsen

External supervisor:
-

Title of thesis:

The Utilization of Supervised Machine Learning in Predicting Corrosion to Support Preventing Pipelines Leakage in Oil and Gas Industry

Credit (ECTS): 30 ECTS

Keywords: - Oil and Gas Pipelines - Corrosion - Supervised Machine Learning - Decision-making	Pages : 53 + enclosure : 8 Stavanger, 15th June 2018

Abstract

Pipelines have become indispensable in oil and gas industry to support transportation of flammable and poisonous fluids such as crude oil, natural gas, and refined petroleum products. They carry fluids in larger volume, safer way, and more environmental friendly compared to trucks and rails. However, like any other equipment, pipelines can have various failures to some degree. One of which is studied in this thesis work that focused on leakage. Leakage in the pipelines can initiate the occurrence of progressive accidents, such as fluid spillage, fire, and explosion. The exposure of that accidents can lead to the injuries, even worst, fatalities, environmental and asset damages, bad reputations, financial distress, and more other negative impacts. Thus, it is important to implement risk-reducing measures that can prevent pipelines leakages. Preferably, the measures must be capable to handle the root causes of the leakages.

Many incidents analysis has shown that leaking phenomena in the pipelines mainly caused by corrosion. Hence, corrosion assessment is crucial to be conducted for decision making in choosing safety measures to avoid leaking incidents. Considering, the type of corrosion, its severity, and factors that can initiate corrosion. Expectedly, preventing actions can be determined and applied based on the root-causes factors.

However, corrosion assessment in the pipelines is a difficult task to execute. This is because of the uncertainty of the future occurrence of corrosion in pipeline. Furthermore, the changing of environmental conditions nowadays, make prediction of the corrosion more difficult. The location of the pipelines for oil and gas operations, which are normally built in a great distance and located in surface and sub-surface also adding complexity in detecting corrosion accurately. Consequently, there are numerous factors that can trigger corrosion to be considered. Therefore, in order to deal with such circumstances, corrosion must be analyzed under multifarious factors per pipelines sections. The tool that can be utilized to estimate such prediction is supervised machine learning. This technology is recognized providing accurate and rapid prediction outputs based on big, various, and complex data.

The purposes of this thesis are to analyze the appropriateness of supervised machine learning in forecasting corrosion and its outputs to support the decision-making in preventing pipelines leakage. The methodologies used in the study are by reviewing literature, and studying the supervised machine learning technology, including how it processes and delivers outputs.

A suggested framework is given to improve limitations of supervised machine learning tool for better decision-making. The framework is constructed by integrating two methods. Initiated by performing a hidden uncertainty analysis method, to better reflect the aspects of uncertainty that can be neglected by the tool. Undertaking this method can minimize surprising outcomes. The second approach adopted is qualitative risk matrices, where the predicted outputs from the tool and consequences analysis outputs are compared. The results of such comparison can assist risk assessors in identifying the level of risk and suggesting recommendations and safety measures to prevent leakage in the pipelines effectively.

Acknowledgments

This master thesis is conducted to fulfill one of the requirements to accomplish Master of Science degree in Risk Management at University of Stavanger.

Beyond everything, I would like to express my praise and gratitude to God Almighty for giving me the opportunity to pursue my master’s degree and complete this thesis work.

I also want to deliver my wholehearted thanks and love to my parents Hj. Esih W and H. Solechudin and my siblings Haqiki Aplesiasfika, Yahya Januarilham, and Daisy Siti Aisyah for their love, prayers, advice, motivation, and support. Once again, thank you so much, my family, for always being there when life goes up and down, joy and sorrow.

Moreover, I would like to express my deepest gratitude to my supervisor, Prof. Eirik.B. Abrahamsen for his time, kind support, advice, precious knowledge, and guidance over the past 6 months. Without his help and guidance, this thesis work would not have been possible.

I would like to say a big thank you to my lovely best friends Mila, Erina, Annisa, Maya, Rani, Tifany, Kiki, Sandra, Tara dan Tika for their support and love. In addition, my greatest thanks to Agung Setiaji who has endlessly supported me to complete my study in Norway.

Last but not least, my most heartfelt thanks go to my Indonesian friends in Norway Kak Opit, Kak Citra, Kak Ayu, Budi, Bang Dika, Bang Anton and my Foreigner friends Andreas and Masheed for their encouragements and helps in facing bittersweet life on these past 2 years.

Stavanger, 15th of June 2018

Stiffi Zukhrufany

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Background

In oil and gas infrastructures, pipelines become one of the crucial assets for transportation. They are regarded as safer elements, lower costs, and more environmental friendly in transporting flammable and toxic fluids such as crude oil, natural gas, and refined petroleum products compared to the trucks and rails. For that reason, pipelines are built in thousands and even million miles starting from the production sites to the petroleum refineries and then continuing to the petroleum transportation hubs for further distribution to the market. However, using such important assets cannot guarantee a 100% safety in transmitting fluids. Any unwanted events can happen once the integrity of pipelines is threatened by corrosion.

Corrosion is one of the most active and dangerous damage mechanisms for pipes (Bolzon, Boukharouba, Gabetta, Elboujdaini, & Mellas, 2011). If it is not treated properly, it will reduce pipelines' wall thickness and cause leakage, which may lead to hazardous fluids released on site and worst to the environment. The volume of spillage can be higher due to the capacity of pipelines. They are able to carry about 70% more fluids than roads and rails, which are able to ship around 3-4% (Dlouhy, 2013). Once the fluids spillage associated with the combustible sources, even in the small amount, major accidents such as fire and explosion can occur. Chevron refinery fire, El Paso natural gas pipeline explosion and Sinopec gas pipeline explosion are the examples of accidents initiated by corrosion in the pipelines. Such accidents had given harmful impacts on human lives, environment, company's assets, reputation, and economic performance.

Thus, corrosion cannot be taken for granted. Corrosion prediction when using pipelines must be identified to prevent pipelines leak. Nevertheless, it is difficult to have accurate detection considering the uncertainties of future events, i.e. the specific section of pipes which will have potential of corrosion. Moreover, environment along the pipelines is likely to change (Muhlbauer, 2004) so that there will be various factors that can lead to the corrosion of the pipelines' surfaces. To have a better understanding and accurate information towards this issue, phenomena of corrosion should be prognosticated regarding the type, level of severity, and others corelated factors, such as hazardous classification of the fluids, the location and environment of the pipelines.

Corrosion prediction should be reflected on different conditions. One of the assessment tool is by using technological advancement called supervised machine learning. Such technology is considered to give accurate and fast forecasting based on a large quantity of data (Hall, 1999). However, the results of prediction that are represented in "classification" can bring skepticism for decision makers, due to the hidden uncertainty in the data, algorithm, and several assumptions that can camouflage crucial aspects of uncertainty in the real cases. Furthermore, the neglected hidden uncertainty can trigger the occurrence of surprising events and bring catastrophe to the human values. Moreover, limiting basis decision only to the classification outputs can disregard the level of risk. Since the tool seems to output classifications from big data and does not include the risk level and uncertainty aspects, decision-makers may face difficulties in taking decisions for risk-reducing measures to prevent pipelines leak effectively.

As the supervised machine learning has limitations and shortcomings, an extensive assessment beyond predicted outputs should be undertaken to strengthen decision support. By having robust decision basis, decision makers can determine appropriate risk-reducing measures to prevent corrosion in the pipelines that can cause leakage.

1.2. Objectives and Approach

The goals of presenting this thesis are to study the suitability of supervised machine learning in predicting corrosion and its predicted results as the decision-making support for preventing pipelines leakage. This study used valuable sources such as reports, scientific works and researches, journal articles, and other publications related with corrosion, pipelines leakage in oil and gas operations, supervised machine learning, treatment of uncertainty, and decision analysis.

1.3. Limitation

The limitations of this thesis are simulation and any quantitative approach, such as corrosion computation and quantitative risk analysis, are not conducted. These caused by the limitations of work scope, time, and data used. Therefore, the data and predicted results in this study are obtained from integration between related projects, published papers, and the author's perspectives.

1.4. Thesis Layout

The thesis work will consist of the information as stated in the following below:

- **Chapter 2**, covers theoretical foundations regarding risk in utilizing pipelines for transporting hazardous fluids, corrosion in the petroleum pipelines, and prediction tool called as supervised machine learning.
- **Chapter 3**, elucidates about the methodology of supervised machine learning for forecasting corrosion.
- **Chapter 4**, demonstrates the types of outputs based on supervised machine learning. In addition, there will be a discussion about its results for identifying any limitations and shortcomings by using the tool. Also, the approaches considered to deal with such weaknesses will be explained in this phase.
- **Chapter 5**, provides suggestions that shall be carried out to improve supervised machine learning results and decision support to prevent pipelines leakage.
- **Chapter 6,** discusses the suitability of supervised machine learning tool in predicting corrosion. Moreover, there will be a discussion about the role of a new framework to improve decision basis.
- **Chapter 7,** demonstrates conclusions and also suggestions for further work
- **Appendix,** delivers summary of this thesis work

2. Theoretical Foundations

2.1. Risk of Utilizing Pipelines for Transporting Hazardous Fluid

2.1.1. What is risk?

As the first consideration, running every operation can generate risk. Risk is the activity (hazard/threat) that can lead to some consequences for human values and those are uncertain (we might not know whether the event will occur and what the consequences would be like) (Aven, 2015; Rosa, 1998, 2003). The activity, consequences, and uncertainty can be constituted to the risk concept and they can be denoted in A, C, U respectively (Aven, 2015). In this part, human values can be referred to human lives, environment, company's assets, reputation, and economic performance.

Risk can bring consequences for the human values; therefore, it must be managed in an appropriate manner. However, in handling risk, we could not refer to the risk concept as it only determines risk in a general view of the situation without comprising the measurement of uncertainty in A and C. To measure the event and consequences, risk must be visualized comprehensively through risk description that suggested by Aven, (A', C', Q, K) (Aven, 2014). A' is specific events, C' is specific consequences considered, Q is a measurement of uncertainty, and K is background knowledge on which A', C' and Q are based on. The elements of A' and C' need to be specified in this section because they are uncertain in the future and the Q can be the tool to estimate those components.

By describing risk through A', C', Q, and K, overall risk picture can be defined. Detail information regarding what can go wrong in the future, how likely or severe it would be, and what the consequences of it can be acquired through this approach. Such information can be an important basis for supporting decision makers in balancing between gaining opportunities and preventing any losses, accidents, and catastrophes (Aven, 2014).

2.1.2. Risk of using pipelines for transporting fluids

In oil and gas industry, transmission of vital fluids is mostly operated by pipelines from one location to the others. They can convey higher volume of fluids safer and more environmental friendly than trucks and trains. Operating trucks and rails can result in higher serious incidents, injuries and fatalities compared to the pipelines (Dlouhy, 2013; Furchtgott-Roth & Green, 2013). However, it does not imply there will be no accidents and fatalities in the use of pipelines. As determined by the risk concept, the activity can cause some consequences. Therefore, the shipments of crude oil, natural gas, and petroleum products by pipelines will not be completely safe.

There are various failures that may occur in the pipelines. But, the failure that will be highlighted in this thesis work is leakage. The incident of pipelines leakage can release a huge amount of fluids to the surrounding areas since they can carry fluids in a large capacity. Such incidents can trigger accidents, such as fire and/or explosion to occur. The exposure of such accidents can be fatal for human values.

Preventing pipelines leakage becomes crucial task to conduct. This is to avoid any risk of leakage that can jeopardize human values. But, before doing any precautions, types of risk factors that can lead to failure in pipelines must be understood. Thereby, mitigation can be done right on the problem being faced. Based on (Dey, 2004), the causes of pipeline can experience failure are:

1) corrosion
2) external Interference
3) construction and materials defects
4) acts of God
5) human and operational error

Based on several studies (Ahammed & Melchers, 1996; Choi, Goo, Kim, Kim, & Kim, 2003; da Cunha, 2016; Dey, 2004, 2006; Vtorushina, Anishchenko, & Nikonova, 2017), corrosion is considered as the biggest cause of failure in the pipelines. It is thus important to implement some risk-reducing measures to deal with corrosion so that pipelines leakage can be prevented. To determine the measures that should be applied, fundamental knowledge about corrosion regarding its causes, consequences, preventing actions for that issues, as well as the assessment method for measuring corrosion should be comprehended.

2.2. Corrosion in The Petroleum Pipelines

2.2.1. What is corrosion?

Corrosion is defined as deterioration of a material, usually a metal, because of reaction with its surrounding environment (Chilingarian, 1989; Popoola, Grema, Latinwo, Gutti, & Balogun, 2013). That reaction can be known as electrochemical process, which contains various solid and liquid substances. The types of substances may vary as they depend on the environmental characteristics on where the pipelines are located. Basically, there are four elements that must react to lead the occurrence of corrosion such as (FluidDataReporting, 2013):

1) Anode (oxidation reaction)
   - Corrosion
2) Cathode (reduction reaction)
   - No corrosion
3) Electrolyte (cations and anions)
4) External path (usually metallic)

If any of the above elements are not available, the pipelines will not corrode or rust. Otherwise, corrosion will occur and reduce the thickness of the pipe wall to some degree. If such reaction is not terminated, the pipeline may form a rough hole (pitting), cracks on its surface and even ruptures. Figure 2.1 shows an illustration of the impacts that can be made by corrosion.

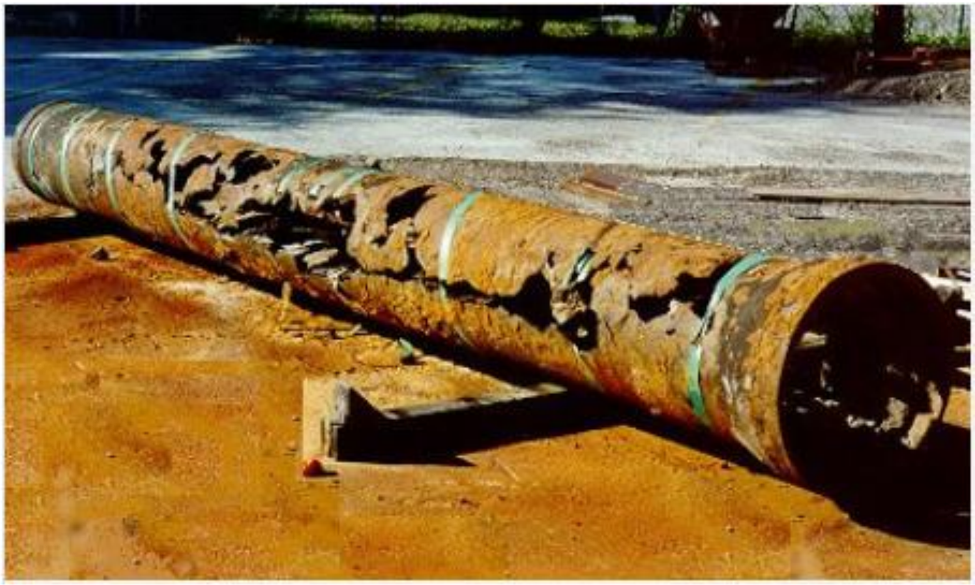


Figure 2.1 Impact of corrosion in the pipelines (Engineers)

As we can see in figure 2.1, corrosion can remove pipe surfaces in various shapes and sizes. Logically, the more severe the corrosion, the bigger the cracks or ruptures that can be created. In regards with that logics, the seriousness of leakage in the pipelines will depend on the severity of the corrosion. Hence, it is important to manage pipelines leak based upon the corrosion severity so that any accidents and consequences can be minimized in line to the problem being faced.

2.2.2. Causes of corrosion

Fundamentally, corrosion can happen because of reaction of anode, cathode, and electrolyte on the metal surface of pipelines. To control corrosion, the factors that can be those elements must be figured out. They can be identified from two parts, which are the vulnerability of the pipes material and the environment that can initiate corrosion on the pipelines wall internally and externally (Muhlbauer, 2004). However, identifying causes will be concentrated on the environmental aspect as it is the factors that can lead corrosion may occur from various factors. In this section, the environment that will be investigated are (Muhlbauer, 2004):

1) Atmospheric corrosion
2) Internal corrosion
3) Sub-surface corrosion

Atmospheric corrosion is a situation where the outer pipelines' wall experiences oxidation because of interaction between its wall and atmosphere (Muhlbauer, 2004). The atmosphere can be a weather, such as rainy, heavy wind, sunny, which the occurrences of those conditions are unsteady. This can be meant that the temperature, humidity factors, and air pollutant rate in the surrounding areas will continuously alter. The variations in those parameters that lead the external pipelines' wall encounter oxidation. In this situation, the higher rate of temperature and air moisture could enhance the process of corrosion in the pipelines (Lloyd). Besides, chemical composition either airborne chemicals (salt or $CO_{2)}$ or man-made chemicals (chlorine and SO_2 which may form H_2SO_4 and H2SO3) can also accelerate the oxidation of metal (Muhlbauer, 2004).

Internal corrosion is the condition where inside pipe wall experiences loss or damage caused by a reaction between the internal pipe's wall and a product being transported (Muhlbauer, 2004). Since the products that are transmitted through pipelines are crude oil, natural gas, and refined crude oil. Therefore, source of corrosion may be a production rates of fluid (oil, gas, water), temperature, flow velocity, CO_2 and H_2S content, water chemistry, oil or water wetting and composition, and metal surface condition (Nyborg, 2005; Papavinasam, Doiron, & Revie, 2010). Another factor that may deteriorate internal wall thickness is a microorganism. This is because sulfate and anaerobic acid are sometimes found in the petroleum pipelines (Muhlbauer, 2004). Nonetheless, such microorganism would not directly lead to corrosion in pipelines. The H2S and acetic acid that resulted from sulfate and anaerobic that can assault the metal immediately (Smart & Smith, 1991).

In this section, there will be explanation about subsurface corrosion. Subsurface corrosion attacks the pipelines that are buried underground. Identifying causes for this case is highly difficult. There are numbers of aspects that should be considered. Nevertheless, the main cause of reduction metal wall thickness in this situation is soil (Ekine & Emujakporue, 2010). The factors that may influence soil corrosion are porosity (aeration), electrical conductivity or resistivity, dissolved salts (including depolarizers or inhibitors), moisture, and PH (CORROSIONPEDIA). Each of specified factors is capable to affect anodic and cathodic polarization characteristics of a metal in soil (CORROSIONPEDIA). Soil corrosion needs to be noticed carefully as it has the capability to significantly damage the pipeline's wall if the environmental conditions are high moisture, electrical conductivity, acidity, and dissolved salts (CORROSIONPEDIA).

Overall, causes of corrosion in pipelines can be identified from three environmental areas, which are external, internal, and subsurface corrosion. The causes that have been detected must be tackled to prevent the occurrence of corrosion.

2.2.3. Consequences of corrosion

Due to corrosion can form leakage in the surface of pipelines, there are some consequences that can happen. Fluid release, fire, and explosions are the effects of such incidents. In this case, the fluids release will be an initial impact that can occur when there is a gap on the pipelines wall. If it associated with the combustible sources, the ignition can happen. The combustible sources can be dust, mist, air mixture, heat and hot surfaces, frictional sparks, auto ignition and so on (SINTEF, 2003). If the ignition is not handled appropriately, the accidents can be extended to the fire and/or explosion.

Fire and/or explosion are the most unwanted consequences. The exposure of fire and explosion can create smoke that may toxic human's health and any organisms in the surrounding area. The worst case is that it could produce thermal radiation that may majorly destruct the environment and properties also lose human lives.

In short, pipelines leak because of corrosion can pose many disadvantages for human values. Therefore, corrosion must be dealt with risk-reducing measures to avoid the occurrence of fluid release, fire, and explosions.

2.2.4. Controlling and preventing corrosion

The corrosion in pipelines can initiate fluid release, fire, and explosion to occur. Therefore, some approaches must be applied to control and prevent such problem. More specifically, the approaches must be capable to preclude anode, cathode, electrolyte to react in the metal pipelines so that an electrochemical process will not be happened.

There are various technical alternatives that can be adopted to control and prevent corrosion. That options are cathodic and anodic protection, corrosion inhibitors, material selection, chemical dosing, application of internal and external protective coatings, corrosion monitoring and inspection (Meresht, Farahani, & Neshati, 2011; Popoola et al., 2013; Samimi & Zarinabadi, 2011). In choosing which of them are suitable to tackle corrosion should be seen with respect to the assessment's outputs, which will generate information regarding the corrosion severity that may be faced.

Performing corrosion assessment are a difficult task to carry out considering the complexities of real-world situation. The assessments should be taken into account the uncertainty of future situation, the changes of environment along the pipelines, and the installation of pipelines that can be constructed in a very long distances, surface, and sub-surface. It is clear that assessment needs to be done under various context, thus, corrosion can be identified accurately. To support that assessment, one can employ predictive analytics tool such a supervised machine learning. That tool is acknowledged can provide prediction of future situation under various conditions accurately and rapidly. In the following sections, there will be comprehensive explanation about that technology.

2.3. Predictive Analytics Tool; Supervised Machine Learning

2.3.1. Predictive analytics

Forecasting corrosion in pipelines requires numerous conditions to be the references for prediction. This can be meant that big data are needed to support such assessment. Relying only on human intelligence to convert large and complex data to generate corrosion prediction could overlook many aspects that might be important in the future. This is because we have limitations in understanding thoroughly actual conditions and processing those data. As a result, predicted outputs may be wrong in representing future situations. Also, it can consume a lot of time to produce a prediction. To deal with such problems, one can use predictive analytics to generate a prediction.

Predictive analytics is technology that forecasts future behavior based on learning from experience (data) in order to drive better decisions (Siegel, 2013). To support the learning process, an algorithm is used to analyze past and present data and identify patterns to predict upcoming events (Azure). Algorithms are defined as a self-contained set of rules used to solve problems through data processing, math, or automated reasoning (Azure). Technological advancement that has the capability to perform such task using algorithm is a machine learning.

By applying machine learning, limitation in the human knowledge and abilities to produce prediction can be handled. What is more, prediction of uncertain phenomena can be done only based on available data. To clarify, that data will be the input and thus processed in the machine to establish a prediction. Such learning process can be visualized as follows.

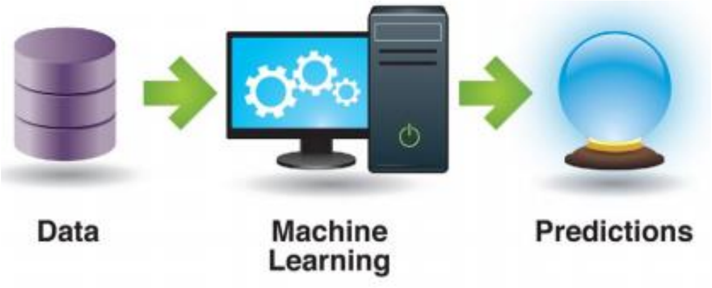


Figure 2.2 The process of prediction using predictive analytics (Siegel, 2013)

2.3.2. Machine learning

Fundamentally, machine learning is part of artificial intelligence (AI). AI has a system that is capable to learn from data, identify patterns, and produce prediction with minimal human intervention (Inc.; Kalogirou, 2001). By reflecting to the AI's capability, machine learning should be also able to establish prediction based on learning process and detecting pattern.

Machine learning is a data science technique that allows computers to use existing data to predict about future behaviors, outcomes, and trends without being explicitly programmed (Azure; Cao et al., 2016; CrashCourse, 2017; Ghahramani, 2015). More specifically, that data will be historical examples or instances for the machine to learn model of the relationship between a set of descriptive features (input) and a target feature (oustput) (Kelleher, Mac Namee, & D'Arcy, 2015).

In this case, we may curious of how such machine can learn data and thus make prediction. The computer has ability to learn data from probabilistic modeling (Ghahramani, 2015). The probabilistic modeling gives a framework for understanding what learning is, and has therefore emerged as one of the principal theoretical and practical approaches for designing computers that learn from data acquired through experience (Ghahramani, 2015). Based on such system, the machine can forecast uncertainty.

2.3.3. Supervised machine learning

Mostly, application of machine learning is premised on supervised learning. In accordance to (Guikema, 2009), supervised learning is an approach for conditions where we have record the outcome data simultaneously with the informative data, which both could be obtained from a historical operation.

To develop an understanding of how supervised machine learning measures uncertainty, let assumes and denotes informative data as input (**X**) and desired outcome data as output (**y**) (Guikema, 2009). To generate prediction from given set of input and output, we need to assess the relationship **y** = f(**X**). The f(**X**) is unknown function of input and it does not associate with any notion of uncertainty in **y** given **X**; hence, risk analysts consider f(**X**) will involve large uncertainty (Guikema, 2009). To treat that uncertainty, algorithm and training dataset are needed to be implemented into computers to learn the form and parameters of a model approximating f(**X**) so that hopefully will result in the right prediction of future circumstance based on new data (Brownlee, 2016; Guikema, 2009).

The typical predicting technique by supervised machine learning is different as done by common prediction tool, such as probabilistic risk analysis (PRA). The key differences are in the assumptions that are made to measure the relationship **y** = f(**X)** (Guikema, 2009). Supervised machine learning made assumptions to estimate such relationship based on given data. Whereas, PRA created assumptions from subjective background knowledge of the logic of condition being analyzed which thus will be used for estimating the failure scenarios or the likelihood of the event. In spite of the differences, it does not imply that supervised machine learning does not utilize PRA-based approach at all. Assumptions of PRA still be used in the supervised machine learning to give valuable insight regarding an important thing that should be taken into consideration carefully.

2.3.3.1.Techniques and algorithms

In the application of supervised machine learning, there are two techniques that can be adopted to develop a predictive model, either classification or regression. A brief explanation of classification and regression techniques can be seen as follows (MathWorks):

1. Classification techniques forecast discrete responses (e.g whether corrosion in the pipelines is “severe” or “low severe”). In this method, input data and desired outputs should be defined, collected, and organized before running supervised machine learning. Thus, classification outputs will be made based on that data.

   The algorithms that are commonly used in this technique to do classification are neural networks, support vector machine (SVM), decision trees, k-nearest neighbor, Naïve Bayes, logistic regression and many more.

2. Regression techniques forecast continuous responses. This is usually used for the case of predicting the real number of changes condition such as humidity rate and/or temperature of the environment.

   The algorithms that are usually adopted for performing regression are neural networks, linear model, nonlinear model, decision trees, and adaptive neuro-fuzzy learning.

From both techniques, the one that associates more with supervised machine learning is classification techniques. The detail explanations of algorithms that are commonly used in the classification techniques can be seen as follows (Ayodele, 2010; Osisanwo et al., 2017):

a) Linear Classifiers
   This algorithm is used to classify items that have similar feature value into classifications. Linear classifiers are rated as the fastest algorithm. Hence, it will be suitable for the situation that has a problem with speed of classification.
b) Logistic Regression
   Logistic regression is as classification function that uses class for developing model. Furthermore, it has a boundary between classes so that the class probabilities will hinge on the distance from its boundary. The more data set, the more rapid the probability. The stronger probability, the more detailed the prediction will be. Nonetheless, that detailed

prediction could be incorrect. Overall, logistic regression is the algorithm that is mostly used for applied statistics and discrete data analysis.

c) Neural Networks
Neural networks are the algorithm that is able to make a prediction by matching pattern in the training data based on a flexible, non-parametric model (Guikema, 2009). The trained network at the end will be used to forecast future condition. Besides, this algorithm could accomplish an amount of regression and/or classification tasks at once even though each network accomplish only one (Bishop, 1995).

d) Support Vector Machines (SVMs)
The purpose of implementing SVMs is to search optimal hyper plane that separates clusters of the vector. The vectors that close to the hyper plane are the support vectors. This algorithm is nearly related to the Neural Networks.

e) Decision Tree
In this part, the trees will classify examples by sorting them according to the feature values. Each node in a decision tree symbolizes a feature in an example to be classified. In addition, each branch expresses a value that node can assume. The node can be eliminated and assigned the most common class of the training examples that are sorted to it (Kotsiantis, Zaharakis, & Pintelas, 2007).

Although there are many types of algorithms, each of it has same goal that is seeking to approximate y = f(X) from the patterns observed in the given historical data (Guikema, 2009). What is more, since each algorithm has different capability in producing prediction, the accuracy of its outputs can be varied.

3. Supervised Machine Learning Methodology for Predicting Corrosion

3.1. Introduction

This chapter will elucidate the mechanism of supervised machine learning in forecasting corrosion in pipelines. There are several procedures that shall be followed to generate such prediction which can be seen in the figure below (CrashCourse, 2017; GL, 2017; Milan, 2016):

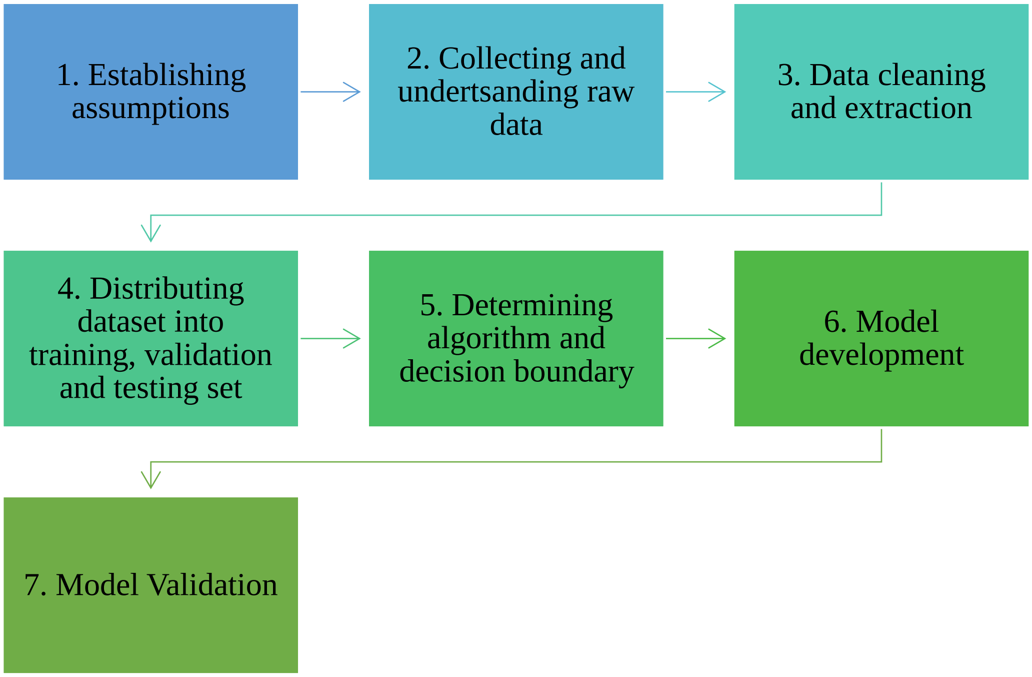

1. Establishing assumptions
2. Collecting and undertsanding raw data
3. Data cleaning and extraction
4. Distributing dataset into training, validation and testing set
5. Determining algorithm and decision boundary
6. Model development
7. Model Validation

Figure 3.1 The workflow of supervised machine learning

In real application, the process of prediction using this tool will start from step 2 to 7. Since predicting corrosion refers to real-world condition can be very difficult because of environment, lengths, and locations of pipelines. Therefore, assumptions should be made in prior to simplify the complexities of actual situations.

3.2. Establishing Basis Assumptions

Basically, assumptions are created to visualize the complexities of the real-world condition from our perspectives. Performing corrosion prediction refers to the actual conditions can be complicated. This is because pipelines can be installed in thousand and even million miles. Also, it can be constructed on the upper ground and underground. Additionally, environmental condition that always changes along the pipelines becomes corrosion can be hard to forecast.

To simplify such complexities, assumptions are made as the references in generating prediction. Taking into account the length and location of pipelines, corrosion should be predicted per pipelines section. It should be also forecasted in some degree of severity to describe corrosion phenomenon in more detail. Furthermore, regarding the changing of environment in the entire pipelines, prediction should be done under numerous factors that can lead to corrosion. Detail explanations of them will be given in the following sections.

3.2.1. Sectioning the pipelines

The length and location of pipelines installation as well as the environment that keeps changing makes potential corrosion in such asset cannot be constant and the risk picture as well. To deal with the instability of corrosion, Muhlbauer (Muhlbauer, 2004) gave suggestion to break pipelines into sections and carry out prediction per its segmentation.

The segmentations of pipelines can be divided into shorter or longer sections. According to (Muhlbauer, 2004), shorter sections can improve the accuracy of the assessment per segment but may result in higher costs of data collection, handling, and maintenance. On the contrary, longer sections may minimize costs in data but also decrease the accuracy because the average or worst case, characteristics must govern in the changeable conditions within these sections.

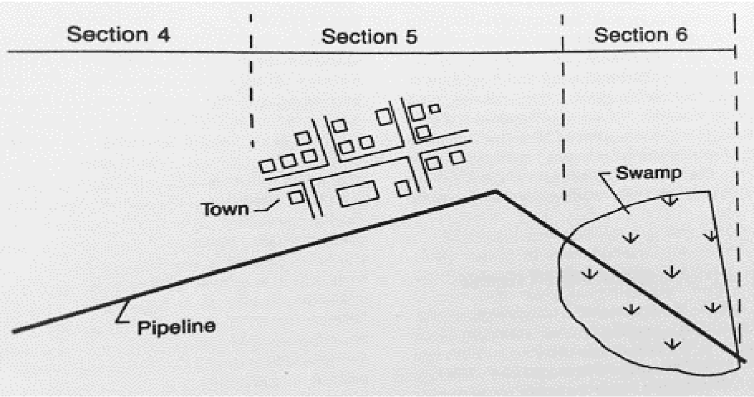


Figure 3.2 Illustration of segmentation of pipelines (Muhlbauer, 2004)

In short, by doing corrosion prediction per pipelines section, we could have a better understanding about the potentiality of corrosion in each area and also produce accurate prediction.

3.2.2. Corrosion classification criteria

In regards with corrosion can attack pipeline’s wall in various severity. It is thus essential to perform prediction based on the level of corrosion severity. We can follow standard practice by NACE International (International, 2010) in defining severity of corrosion, which can be visualized in the following below:

1). Severe, indicates having the highest likelihood of corrosion activity

2). Moderate, indicates having possible corrosion activity

3). Minor, indicates having inactive or lowest likelihood of corrosion activity

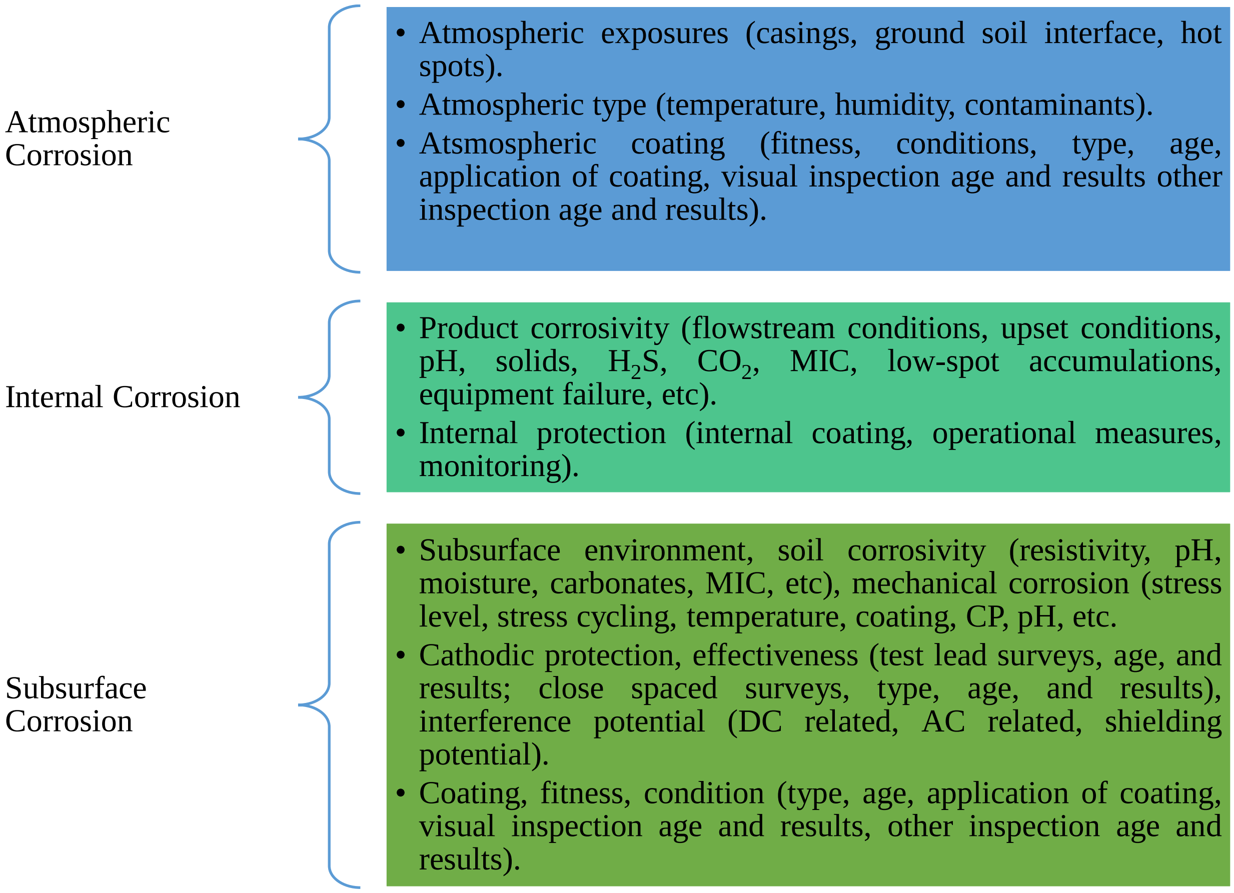
In this thesis work, the severe, moderate, and minor corrosion will be the outputs of prediction that we wish to predict using supervised machine learning. By representing corrosion in light of the degree severity, decision-making support under uncertainty can be produced in more detail way.

Moreover, it can assist risk analysts to provide suggestions of what should be done to reduce its severity.

3.3. Collecting and Understanding Raw Data

After assumptions have been made, data should be gathered and understood. Raw data can be obtained from various sources, such as inspection data, original construction, environmental condition, operating and maintenance history, historical failures and others (Miesner & Leffler, 2006; Muhlbauer, 2004). In each source, there will be many data that can be selected to support prediction of corrosion using supervised machine learning and they must be must be chosen carefully. Selecting wrong data can lead the outputs of prediction to not represent future conditions. As a result, surprising events can be likely to occur.

To have a better understanding of what data that must be selected, one can refer to the suggestion by (Muhlbauer, 2004). In that part, data are related to the causes that lead to corrosion and the physical exposures that can be degraded directly by corrosion, which the examples of them can be seen in figure 3.3.



Atmospheric Corrosion

- Atmospheric exposures (casings, ground soil interface, hot spots).
- Atmospheric type (temperature, humidity, contaminants).
- Atsmospheric coating (fitness, conditions, type, age, application of coating, visual inspection age and results other inspection age and results).

Internal Corrosion

- Product corrosivity (flowstream conditions, upset conditions, pH, solids, H_2S, CO_2, MIC, low-spot accumulations, equipment failure, etc).
- Internal protection (internal coating, operational measures, monitoring).

Subsurface Corrosion

- Subsurface environment, soil corrosivity (resistivity, pH, moisture, carbonates, MIC, etc), mechanical corrosion (stress level, stress cycling, temperature, coating, CP, pH, etc.
- Cathodic protection, effectiveness (test lead surveys, age, and results; close spaced surveys, type, age, and results), interference potential (DC related, AC related, shielding potential).
- Coating, fitness, condition (type, age, application of coating, visual inspection age and results, other inspection age and results).

Figure 3.3 Sample of data to predict corrosion (Muhlbauer, 2004)

In figure 3.3, there are numerous data that can be used to forecast corrosion in pipelines. For instance, assuming corrosion engineers would like to predict external (atmospheric) corrosion in pipelines. Hence, they can choose temperature, humidity, contaminants, type of coating, age of coating, and casing to be the sample of data in determining whether under such conditions the pipelines would have severe, medium, or minor corrosion.

However, not all data given in the figure 3.3 will be selected to be the input to forecast corrosion considering some of the them might not vital and complete. It is thus important to understand the relationship between the problem that would like to predict and the data needed. Thereby, we can have insight which potential data that should be chosen to produce accurate prediction.

3.4. Cleaning and Extracting Data

After data have been gathered and understood, they should be cleaned and extracted. This is to exclude unessential and incomplete data and to determine dataset. Dataset will contain a set of features (inputs) and classification outputs. Any values or parameters involved in the set of features should be potentially relevant for predicting performance and measurable for future application of the model (Guikema, 2009). Thereby, the predicted model can be more accurate and correct in indicating classification outputs.

Before defining the dataset, unimportance information and missing value must be eliminated from collected data. This is to produce complete data and accurate prediction. Missing value can minimize the statistical power of a study and can establish biased estimates which lead to invalid results (Kang, 2013). Once complete data has been obtained, it should be taken into further consideration by individuals (analysts, engineers, and experts) to select a powerful sample of data that will be input into the set of features. It should be noted that set of features should fit with the classification outputs that one wishes to classify.

To illustrate the determination of dataset, let assume corrosion engineers want to forecast external corrosion based on the level of severe, moderate, and minor. By discussing with some expertise, they decided to take parameters, such as temperature, humidity factors and pipelines wall thickness as the conditions that can indicate corrosion from the degree of severity. The dataset for this case will be constructed as the following below:

Table 3.1 Illustration of a dataset for the case of predicting external corrosion

Set of Features			Supervised Classification output
Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
32	55	19	minor corrosion
20	95	11	severe corrosion
29	63	20	minor corrosion
15	89	15	medium corrosion
11	90	14	medium corrosion
19	91	10	severe corrosion

3.5. Distributing Dataset into Training, Validating, and Testing Set

The dataset that has been defined must be distributed into training, validating, and testing set. In the training set, the dataset will be supplied to the learning algorithm for finding the relationship, developing understanding, making decisions, and evaluating their confidence from that training data (CrowdFlower; Ripley, 2007). In validating set, it will be utilized for evaluating an unbiased in the model that is generated from training set while tuning model hyperparameters (SHAH, 2017). Lastly, the dataset should be distributed into the testing set because they will be used to evaluate the final model that has been processed through the training and validating sets (SHAH, 2017). In this part, the testing set will contain new instances, where the machine has not learned yet and it will be loaded into a predicted model for evaluation purposes.

In terms of how much data that should be distributed into validation set will depend on the amount and complexity of hyperparameters (SHAH, 2017). If there are few hyperparameters, one will need small validation datasets and vice versa. Also, if the hyperparameters are difficult to tune, one might not need a validation set in applying supervised machine learning to create a prediction of corrosion.

3.6. Defining Algorithm and Decision Boundary

Algorithm and decision boundary are the important parts that should be implemented into computers to support the learning process. There are many types of algorithms that can be utilized to allow machines learn dataset (see section 2.3.3.1). Defining algorithm that can establish prediction accurately and correctly can be confusing and difficult. One shall run all algorithms of supervised machine learning into computers and thus choose the one that has the highest accuracy. Nonetheless, choosing proper algorithm should not be limited to the accuracy numbers. In-depth consideration under different context must be done, such as how if data increases and/or collaborates with other parameters. Comprehensive explanation about that will not be discussed in this thesis as it is not part of the scope of thesis work.

Besides algorithm, another important thing that should be set is decision boundary. The aim of setting decision boundary is to assist algorithm in classifying dataset into a particular class (Algolytics). Due to most of the algorithms are based on probabilistic models (Ahoerstemeier, Kotsiantis, Peteymills, & Zadroznyelkan); hence, decision boundary can be defined based on probability estimator. According to (Flach & Matsubara, 2008), probability estimator is a scoring classifier that gives probabilities. And it can be set based on our assumptions (University, 2015).

For example, to support algorithm in classifying corrosion severity, risk analyst, corrosion engineer, and expertise set decision boundary as follows, which they are made based on reference from DNV GL (GL, 2017):

1) If the amount of predicted corrosion shows 0 – 1%, thus it will be classified to minor corrosion.
2) If the amount of forecasted corrosion gives results between 1 – 40%, it will indicate to medium corrosion.
3) Meanwhile, if the amount of forecasted corrosion gives outputs between 40 – 100%, then it classifies to severe corrosion.

3.7. Model Development

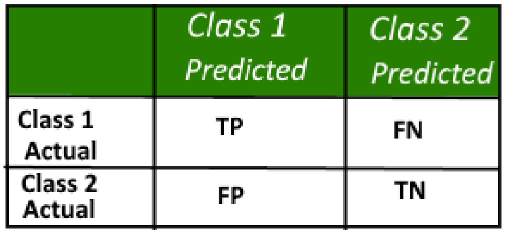
Once dataset, algorithm, and decision boundary have been prepared, then, predicting model can be built. The main objective of building the model prediction is to improve the accuracy and adjust computers to only use the defined set of features for assessment or measurement of a problem case being studied (Guikema, 2009).

The model development can be started by input training dataset into learning algorithm. At the first time, the model might generate a poor prediction. If it keeps training with the output that should have established, the predicted model can be more accurate in the next time. In this part, when the predicted model has been produced, it shall be tuned with the validating dataset and should be evaluated with the testing dataset. Nonetheless, since the validating dataset might be (not) defined because of the complexity of hyperparameters to tuned, therefore, that process can be hopped. For the testing set, it cannot be disregarded because it can determine whether the predicted model performed well or not in forecasting uncertainty.

3.8. Model Validation

Evaluating performance of the algorithm in making a prediction is a task that must be performed. This is to gain insight whether the predicted model would be correct and accurate in predicting new data about corrosion that have never been trained before. Moreover, it is to visualize how the model might perform in the real-world situation.

To carry out such evaluation, testing dataset must be input into the predicted model. What is more, decision boundary must be also loaded into the machine. By doing so, correct and false prediction can be recognized. The differences in both values will be presented through confusion matrix, which it involves actual classes in rows and predicted classes in columns (Flach & Matsubara, 2008). The basic concept of confusion matrix can be seen as follows:



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Figure 3.4 Confusion matrix (GeeksforGeeks)

By looking at figure 3.4, class 1 shows p (positive) and class 2 shows n (negative). The denotations in the confusion matrix will be described as follows (GeeksforGeeks):

- Positive (P) : Observation is positive (for instance, it is corrosion).
- Negative (N) : Observation is not positive (for instance, it is not corrosion).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Nonetheless, it should be noted that how many numbers of columns and rows in confusion matrix will be adjusted by how many classifications that individual(s) wish to classify.

For example, there are three supervised classification outputs that should be predicted, such as minor, medium, and severe corrosion. Thus, the presentation of confusion matrix would not be the same as demonstrated in the figure 3.4. Rather, it will be presented as in the figure 3.5. where the author constructed it based on the reference (Sadawi, 2014).

	Predicted			
Actual		Minor Corrosion	Medium Corrosion	Severe Corrosion
	Minor Corrosion	TP minor corrosion	E minor-medium	E minor-severe
	Medium Corrosion	E medium-minor	TP medium corrosion	E medium-severe
	Severe Corrosion	E severe-minor	E severe-medium	TP severe corrosion

Figure 3.5 Illustration of confusion matrix for multi-class classification of corrosion

By looking to the figure above, we can notice that there is only information about true positive prediction. True positive demonstrates the prediction is correct. To have information about the false negative, true negative, and false positive in the multi-class confusion matrix, one can follow computation as follows (Sadawi, 2014):

- False Negative (FN)
  The total number of false negative for a class can be obtained by summing values in the corresponding row without including the TP in that class.
- False Positive (FP)
  To gain a total number of false positive for a class, one should sum values in the corresponding column without including TP in that class.
- True Negative (TN)
  The total number of true negative for a certain class will be acquired by summing all columns and rows except the value in that class's column and row.
- Total number of test examples of any class
  To have insight of how many test instances in any class, one can sum of the corresponding row including TP in that class.

After information about true positive, false negative, true negative and false positive have been known, how well the performance of the algorithm in generating a prediction can be evaluated. The evaluation can be done by measuring accuracy, precision, and recall, which their formulas can be visualized as follows (GeeksforGeeks):

- Accuracy

The accuracy of prediction can be calculated by:

$$\frac{TP+TN}{TP+TN+FN+FP} \cdots\cdots \tag{3.1}$$

The outcomes of accuracy shall not be trusted completely. A 90% accuracy of prediction can show that it is correct, average (between true and false), or wrong. The corrosion that has predicted to be minor corrosion, may be severe in the real cases. Therefore, broader assessment should be done to strengthen prediction outputs and avoid misclassification. In this case, misclassification can bring harmful impacts not only to the human lives and environment but also to the company's assets, reputation, and economical performances.

- Recall

Recall is the ratio of total number of true classified positive examples divides to the total number of positive examples. The formula can be seen as follows:

$$\frac{TP}{TP+FN} \cdots\cdots \tag{3.2}$$

High recall can define the class is correctly identified or it can be meant that there is small number of FN (False Negative).

- Precision

In order to obtain a precision value, one should divide the total number of true classified positive examples by the total number of predicted positive examples. The formula of precision is given below:

$$\frac{TP}{TP+FP} \cdots\cdots \tag{3.3}$$

In this part, high precision can identify that a positive classification output is truly positive (small number of FP (False Positive)).

We can integrate the recall and precision values to generate conclusion of such prediction as follows (GeeksforGeeks):

- High recall, low precision can be meant that most of the positive examples are correctly recognized but there are a lot of false positives.
- Low recall, high precision can be meant that we lose numerous of positive examples (high FN) but those we forecast as positive are actually positive (low FP).

Overall, model evaluation is conducted to examine predicted model based on the algorithm's performance. The most crucial quantities in classification performance, such as true positive, false positive, and accuracy can be acquired as well (Flach & Matsubara, 2008). Those values can assist individual(s) in defining how many correct and false prediction. Furthermore, it can determine the appropriateness of such algorithm to make a further prediction. If the algorithm can produce high accuracy of prediction, it can be implied that the predicted model is robust to be used for predicting actual condition in the future.

4. Supervised Machine Learning Results and Discussion

4.1. Introduction

This chapter provides a demonstration of supervised machine learning outputs in forecasting corrosion in pipelines. The results will be thus discussed from the risk management perspectives to investigate whether the tool will be suitable for predicting corrosion and whether it will be sufficient for being decision-making support to prevent pipelines leakage. The approaches that shall be considered to improve decision supports will be provided after limitations and shortcomings of the tool have been detected.

4.2. Supervised Machine Learning Results

Based on theoretical foundations in chapter 2, supervised machine learning will result prediction in classification (discrete response). What will be the classification depends on what we wish to predict. For the case of forecasting corrosion in pipelines, the classifications shall be reflected to the corrosion severity. In this thesis work, the degree of severity that shall have to prognosticate are minor, moderate, and severe corrosion. Since operating supervised machine learning is not the part of this work. Therefore, an illustration of how such technology describes the defined corrosion severity will be presented in this section.

Before demonstrating the illustration, let us assume the situation where corrosion engineers would like to forecast external corrosion in the specified pipelines. By following to the NACE Standard International, corrosion will be foreseen based on the degree of severity, such as severe, moderate, and minor. Moreover, based on discussion with some experts and referring to the sample of data collection (see figure 3.3), the potential data that will be used to generate prediction are temperature, humidity factors, and pipelines wall thickness. After all important parameters have been observed, the dataset should be created and then fed into a different set (training and testing) for model development and validation purposes. The outputs of this prediction can be seen in table 4.1, which it is constructed based on collaboration from several literatures (dataminingincae, 2014; GL, 2017; Mahjania, Jalilia, Jafariana, & Jaberia; Maini, 2017; Montgomery, 2016; Supriyatman, Sidarto, Suratman, & Dasilfa, 2012; University, 2015)

In table 4.1, the outputs of prediction using supervised machine learning are displayed in the row “testing dataset after input into evaluated predicted model”, column “supervised classification output”. Such classifications can be obtained from the learning process that is done by algorithm. To be more clearly, by learning information in the row “training dataset”, the algorithm can be able to generate prediction and classification about that data. It should be noted that, in this example, the values given in each parameter of temperature, humidity factors, and pipelines wall thickness are only illustrative because of the limitations in the data availability.

Table 4.1 Illustration of predicted outputs based on supervised machine learning

	Set of Features			Supervised Classification output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
Training dataset	32	55	19	minor corrosion
	20	95	11	severe corrosion
	29	63	20	minor corrosion
	15	89	15	medium corrosion
	11	90	14	medium corrosion
	19	91	10	severe corrosion
	Set of Features			Supervised Classification output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
Testing dataset before input into evaluated predicted model	20	80	20	?
	32	67	18	?
	13	94	12	?
	22	90	15	?
	25	95	10	?
	17	88	13	?
	Set of Features			Supervised Classification output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
Testing dataset after input into evaluated predicted model	20	80	20	minor corrosion
	32	67	18	minor corrosion
	13	94	12	severe corrosion
	22	90	15	medium corrosion
	25	95	10	severe corrosion
	17	88	13	medium corrosion

By describing corrosion as in table 4.1, we can be more understanding of what can go wrong in the future under diverse conditions of e.g temperature, humidity factors, and pipelines wall thickness. In practice, corrosion can be predicted based on more than three features. It can be ten or even larger, which it will depend on the context of the assessment. The point is that although there are a lot of data or conditions that should be learned by the algorithm to predict corrosion, that technology still capable to find pattern recognition and make automate indication accurately.

Overall, by adopting supervised machine learning, corrosion can be forecasted under various severity and factors that can lead to corrosion. From my point of view, this approach can help risk analysts in improving their knowledge regarding severity of corrosion that may occur under different situations. Risk-reducing measures to prevent corrosion can be also defined based upon its severity being faced, which, hopefully, they can avoid pipelines leakage effectively.

4.3. Is Supervised Machine Learning Fruitful for Predicting Corrosion?

It is known that many measurements of uncertainty generate prediction in a probability or expected value. Meanwhile, supervised machine learning produce prediction in a classification. That differences lead to the curiosity whether such predictive analytics tool will be useful to foresee corrosion in pipelines?. To answer such question, we must be remembered that predicting corrosion throughout the pipelines is quite difficult. The severity of corrosion that may deteriorate pipeline's wall thickness are uncertain because of several factors, such as changing environment, length and location pipelines.

To deal with such uncertainty, corrosion should be predicted per pipelines section with respect to its severity and numerous factors that may cause corrosion to occur. By doing so, phenomena of corrosion can be captured under different context, which that is good to develop understanding of what can go wrong in the upcoming event. However, performing prediction under those conditions only using human intelligence can lead to several problems.

There will be a big and complex data as well as several assumptions that we need to process for generating such prediction. Indeed, it would be complicated and frustrating to convert all available background knowledge (data and assumptions) into information about corrosion in the future. Our knowledge has a limitation in understanding and integrating overall aspects related to pipelines corrosion. As a result, prediction can be not accurate and important aspects related to future event can be overlooked. Furthermore, it can consume a lot of time to process this prediction. In practice, the assessment results need to produce promptly and precisely because decisions must be taken immediately to resolve the issues being faced.

To assist human intelligence in prognosticating corrosion under numerous conditions, we can use technological advancement such as a supervised machine learning. That tool is capable to make accurate and quick predictions based on learning from data even it is a big data. The type of outputs that will be generated by this tool can be seen in table 4.1. By considering the way of corrosion is described as in that table, we can be more understanding about the factors that can cause corrosion to occur in some degree of severity. For instance, if the temperature, humidity, and pipelines wall thickness are 25ºC, 95%, and 10 mm respectively; thus, this will indicate severe corrosion. Meanwhile, when the conditions of those variables show 22ºC, 90%, 15 mm then the severity of corrosion that may attack the surface of pipeline is medium.

By having the ability to predict corrosion based on multifarious situations accurately and instantaneously, in my opinion, supervised machine learning seems fruitful to be used to forecast corrosion in pipelines. Furthermore, the outputs form this tool can help risk analysts in providing suggestions of what needs to be done to handle a different level of corrosion.

4.4. Are Supervised Machine Learning Outputs Robust to be The Decision Support?

If the purpose of predicting corrosion is to support decision makers in avoiding pipelines leakage, predicted results based on supervised machine learning should not be entirely believed. This is because such tool must have some drawbacks that can affect the accuracy of the prediction's results. Hence, we may wonder whether the classification outputs are strong enough to be the decision-making support to prevent leakage incidents in pipelines?. To answer this question, we must first identify the shortcomings of supervised machine learning.

It is known that the predicted results based on this tool are underlying on the data, learning algorithm and several assumptions. Data that is used by the algorithm to learn and generate prediction can be inherent with uncertainty. The instances in the training dataset are made based on individuals' background knowledge. Thus, once they gave wrong examples, the algorithm will produce incorrect prediction. Moreover, the parameters values and/or other information that obtained from historical data may not reflect the actual or original situations.

In this case, algorithm can be also associated with the uncertainty. The technique of algorithm in finding pattern recognition between inputs and outputs to generate prediction is not transparent. That is why, the truths of predicted outputs will be uncertain. The algorithm can be called as black boxes prediction as it has ability to learn data easily and quickly and thus find solutions for those who have a limitation or nothing knowledge in its inner workings (Kamalnath, 2017). Apart from that choosing wrong algorithm to create prediction can result in incorrect and inaccurate classification outputs. As the consequences, the predicted classifications are not representing actual conditions.

Furthermore, assumptions can be involved with uncertainty as it is made based on our knowledge to simplify the complexity of the actual situation. Meanwhile, our knowledge can overlook the aspects of uncertainty (Abrahamsen, Aven, & Iversen, 2010). Thus, it may be wrong in making representation of actual conditions. This is because we have limitations in visualizing the world as a whole.

Considering background knowledge such as data, algorithm, and assumptions can likely to collaborate with uncertainty. Therefore, the classification outputs should be used with caution because the aspects of uncertainty are not reflected comprehensively. In this part, what has been predicted to be minor corrosion can be severe corrosion in the real-world situations. It is thus crucial to not overlook uncertainty because it can lead to the occurrence of surprising outcomes which they can cause more serious disaster to human values. This is why, uncertainty is assumed as dominant factors of risk (Abrahamsen et al., 2010).

Besides neglecting uncertainty, supervised machine learning results are also not reflected the aspect of risk. More specifically, it is not taken into considerations the degree of risk. As a result, decision makers may have difficulties in understanding which severity of corrosion that may bring high or unacceptable risk when it should occur. It is also lead to the problem in deciding safety measures that should be implemented shortly.

By considering the weaknesses of this tool, supervised machine learning results are not strong enough to be the decision basis to support preventing pipelines leakage. This due to the uncertainty and risk are not reflected comprehensively. Meanwhile, in managing safety of an operation, awareness to the both aspects are vital because they can be the references in reducing the occurrence of unwanted accidents and other consequences that can harm human values. It is thus important to develop decision basis based on this predictive tool. It can be done by performing extensive analyses that can cover the aspects of uncertainty and risk in the decision-making support.

4.5. The Need for Performing Extensive Analyses Beyond Supervised Machine Learning Results

Producing decision basis based on supervised machine learning can establish comprehensive information about corrosion. Detail preventing actions can be defined in line to the problem being faced. However, it is not the perfect tool to be the only decision support for preventing leakage in the pipelines. The reason is because the uncertainty and risk aspects are not taken into account properly by this tool. In the meantime, accidents, losses, and catastrophes can be avoided by reducing risk and uncertainty involved in its activity (Aven, 2014).

Thus, uncertainty and risk need to be considered in the decision-making support. There are many approaches that can be used to reflect both aspects. In terms with uncertainty, the method should be able to capture the aspects of uncertainty in a detail way. Thus, the occurrence of surprising events can be avoided. In accordance to (Gross, 2010), an event is regarded as a surprise if the occurrence of it is not expected and also contradicted to the accepted knowledge. Meanwhile, based on Aven (Aven, 2014), surprising event (with severe consequences) can be known with black swan, which that is related to the present knowledge/beliefs. Envisioning both experts' point of views, it can be highlight that surprising outcomes can occur because of the current knowledge/belief that is not considered about such events. In my opinion, it can happen because, naturally, human intelligence has a limitation in knowing thoroughly about what will occur in the upcoming event. Thereby, such surprising events are not included when performing analyses and/or assessments.

For reflecting the aspects of risk in the decision basis, we shall adopt the method that can diagnose the level of risk that may be confronted. Knowing the risk level can assist risk analysts in producing more detail information regarding to risk in that activity and suggestions of measures to reduce its risk. Moreover, they can have an insight of which risks that are not acceptable and acceptable. In addition, they can construct better communication about risk assessment's results to the decision makers. It is important to produce clear and understandable information about risk in the activity so that decision makers can easily review, understand thorough phenomena, and weigh decisions that should be taken.

All things considered, to deal with the weaknesses of supervised machine learning, extensive analyses more than supervised machine learning should be carried out. This is to involve uncertainty and risk aspects in the decision basis. Hence, it can be more robust to support decision makers preventing pipelines leakage.

4.6. The Need for Undertaking Consequences Analysis

In order to support decision basis reflects the aspects of risk, especially the risk level of an event. The element of the risk itself must be described properly. By referring to the section 2.1.1, risk can be described through A’, C’, Q, and K. Meanwhile, in this thesis work, implementation of supervised machine learning that is used to support preventing pipelines leak only covered the elements of A’, Q, and K. The A’ are the level severity of corrosion (minor, medium and severe), Q is supervised machine learning tool and the K are the data, performance of algorithm, assumptions and suppositions.

Indeed, limiting information based on supervised machine learning will not reflect the aspects of risk comprehensively. Considering such problems, therefore, specified consequences (C’) should be analyzed to complete the information about the risk being confronted. This can be done by performing a consequences analysis. The objectives of performing such analysis according to the NORSOK Z-013 are to (Association, 2010):

a). assess the possible outcomes of identified and related initiating events that may contribute to the overall risk picture;

b). analyze potential event sequences that may evolve following an initiating event that happen, define the influence of the performance of barriers, the degree of the physical impacts and the extent of damage to personnel, environment, and assets, corresponding to the specified context of assessment.

The approaches that can be used to assess the possible outcomes and examine the potential sequent consequences are varied. There are qualitative and quantitative approaches such as coarse judgmental assessment (extrapolation based on available data or experimental studies), event tree analysis (involving detail assessment of the various branches) and so on (Vinnem, 2014). Those qualitative and quantitative methods will generate results in expected judgments and values respectively.

All in all, analyzing the specified consequences in each specified initiating event can improve the insight about the overall risk in such activity. Also, it can support in defining what kind of safety measures that should be addressed to avoid the occurrence of specified consequences.

5. Suggestions to Improve Decision Support based on Supervised Machine Learning

5.1. Introduction

It is known that decision basis based on supervised machine learning is not robust to be decision support for preventing pipelines leakage. That tool ignores the crucial aspects of uncertainty and risk. Therefore, extended analysis should be carried out to improve decision support.

Before determining what kinds of analyses that shall be implemented, one should notice that uncertainty and risk have its own issues that should be concerned. For uncertainty, the issues are related to fuzziness in current knowledge that is used to forecast uncertain condition. Meanwhile, the risk is centralized more to the stage of jeopardy that would be faced when an event occurs. In spite of the differences, gathering both aspects can support decision makers in determining what kinds of treatment that must be taken to manage safety under the conditions that are associated with vagueness.

Due to by reflecting the aspects of uncertainty, we can have insight in how to better handle risk. Hence, broader analyses shall be done following to the suggested framework below:

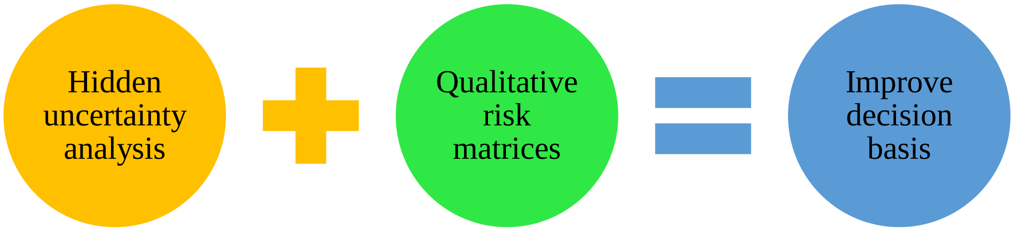


Figure 5.1 A suggested framework to improve decision basis from supervised machine learning

As we can see in figure 5.1, the framework contains with two methods; hidden uncertainty analysis and qualitative risk matrices, which they will be performed gradually. In this part, hidden uncertainty analysis will be performed in prior considering uncertainty is the fundamental problem that may cause surprising outcomes occur. By applying such method, we could have an insight in how supervised machine learning outputs and risk should be interpreted with respect to the uncertainty involved. Qualitative risk matrices are the second method that will be executed to generate information about the degree of risk that may be confronted in the future. It should be noted that interpretation of risk level should be regarded to the degree of uncertainty that obtained from first method.

5.2. The Application of Hidden Uncertainty Analysis

Based on information in section 4.5, to anticipate the occurrence of surprising events, one must put more attention to knowledge/beliefs that are used to make prediction. This is because they are the main sources that uncertainty can be overlooked. It is thus crucial to examine uncertainty in the knowledge bases. It can be done by performing a hidden uncertainty analysis. The methodology of this approach can be seen in figure 5.2. It is created based on collaboration of the author's perspective with the papers from (Selvik & Aven, 2009) and (Abrahamsen et al., 2010).

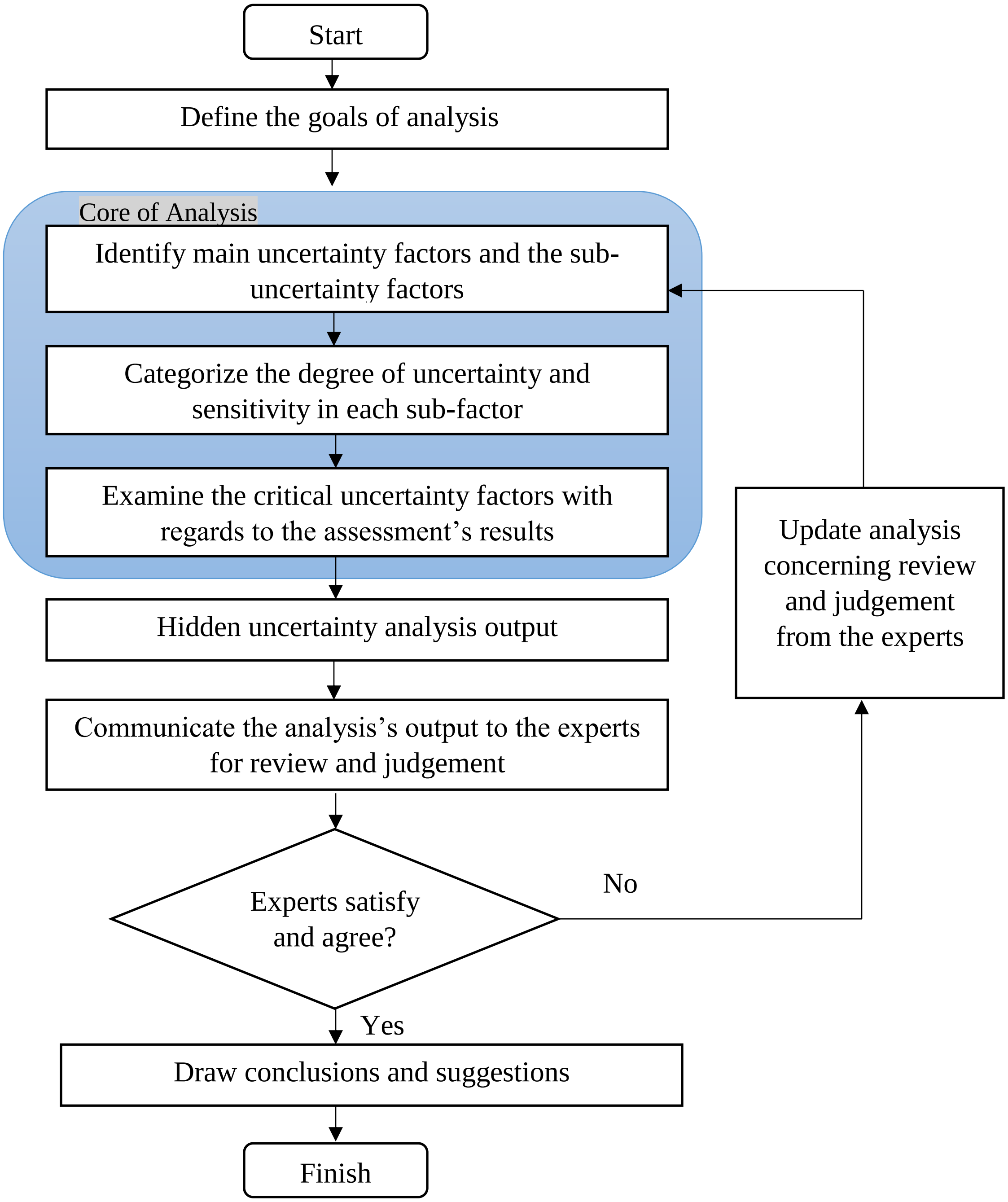

Start
Define the goals of analysis
Core of Analysis
Identify main uncertainty factors and the sub-uncertainty factors
Categorize the degree of uncertainty and sensitivity in each sub-factor
Examine the critical uncertainty factors with regards to the assessment's results
Hidden uncertainty analysis output
Communicate the analysis's output to the experts for review and judgement
Experts satisfy and agree?
No
Yes
Update analysis concerning review and judgement from the experts
Draw conclusions and suggestions
Finish

Figure 5.2 Procedure of performing hidden uncertainty analysis

In figure 5.2, there are several stages that should be accomplished to figure out the aspects of uncertainty. The first step is to determine the objectives of analysis. By specifying the goals in the early phase, we could have insight about what should be achieved and also what actions that should be taken to accomplish the targets.

The next stage is to identify main uncertainty factors and the sub-components of it. We might be confused about what factors that should be discovered in this case. Following the perspectives by (Abrahamsen et al., 2010; Aven, 2014; Gross, 2010) that mentioned surprising events can happen because of camouflaged uncertainties in the background knowledge. Therefore, the factors that should be identified is the uncertainty in the knowledge base that are utilized to generate a prediction. The specification of both elements should be performed meticulously and carefully because it is the core of the hidden uncertainty analysis. The more uncertainty factors that are detected, the more we can anticipate and find proper solutions to deal with that. Hence, the predicted results from running supervised machine learning can be more valid and robust.

Once the second step has been finished, the identified uncertainty factors should be categorized with respect to the degree of uncertainty and sensitivity (Aven, 2008). In this part, the level of uncertainty should be considered as it could define whether the basis knowledge to support making a prediction involved with large uncertainty. High uncertainties in the background knowledge may lead the predicted outputs to give misclassification about what can go wrong in the future. Under such circumstances, human values are at a stake. In the meantime, the degree of sensitivity should be reflected to have an insight about the effects on the prediction's outputs when e.g different data and assumptions are performed. All in all, to ensure the consistency in the process of categorization, some guidelines are needed to implement. In this part, the categorizations of the degree of uncertainty and sensitivity will refer to the (Flage & Aven, 2009) as can be seen in table 5.1 and 5.2.

Table 5.1 Guidelines for categorizing degree of uncertainty and sensitivity (Flage & Aven, 2009)

Aspect	**Score**	**Interpretation**
Uncertainty	Significant	*At least one of the following condition is fulfilled*
		• The phenomena involved are not well understood; models are non-existent or known/believed to give poor predictions
		• Data are not available, or are unreliable
		• The assumptions made represent strong simplifications
		• There is lack of agreement/consensus among experts

Table 5.2 Guidelines for categorizing degree of uncertainty and sensitivity (continued) (Flage & Aven, 2009)

Aspect	**Score**	**Interpretation**
Uncertainty	Moderate	*Condition between level uncertainty of high and low*
		•The phenomena involved are well understood, but the models used are considered simple/crude.
		•Some reliable data are available
	Minor	*All following conditions are fulfilled*
		•The phenomena involved are well understood; the models used are known to give predictions with the required accuracy
		•The assumptions made are seen as very reasonable
		•Much reliable data are available
		•There is broad agreement among experts
Sensitivity	Significant	• Relatively small changes in base case values needed to bring about altered conclusions.
	Moderate	• Relatively large changes in base case values needed to bring about altered conclusions.
	Minor	• Unrealistically large changes in base case values needed to bring about altered conclusions

After categorizing sub-uncertainty factors with respect to the guidelines above, we need to examine the importance of uncertainty factors. It can be done by averaging the score of the degree uncertainty and sensitivity (Aven, 2013) in each factor. By doing so, risk analysts can detect potential factors that may affect the predicted outputs to give a false representation of the actual condition and trigger surprising outcomes to happen.

When all processes in the core of analysis have been carried out, risk analysts could establish comprehensive information about uncertainty with respect to the level of uncertainty, sensitivity, and criticality. That information should be informed to the experts for reviewing and judging about whether such information is robust, the specified factors are rigorous, and there are missing aspects that the risk analysts neglected to identify. If they are not satisfied and agreed about that results due to e.g there are some uncertainty factors that are still ignored, hence, risk analysts should perform update analysis considering the advices and suggestions that are given by the experts. Otherwise, we can proceed to the final step that is to make conclusions and suggestions. The conclusion and suggestions are created to inform how predicted outputs based on supervised machine learning should be used as a decision-making support under uncertainty. All in all, an illustration of a hidden uncertainty analysis related for this thesis case can be seen in table 5.3.

Table 5.3 Illustration of hidden uncertainty analysis approach

Goal				
To identify the overall degree of uncertainty involved in the predicted outputs and uncertainty factors that can significantly trigger the occurrence of surprising outcomes				
Hidden Analysis Uncertainty Outputs				
Main uncertainty factors	Sub-main uncertainty factors	Degree of uncertainty	Degree of sensitivity	Degree of criticality
Data	Quality of dataset	Significant	Significant	Significant
	Age of dataset	Moderate	Moderate	Moderate
Algorithm	Performance of algorithm	Moderate	Significant	Moderate
	The operator(s) that perform supervised machine learning	Moderate	Significant	Moderate
Assumptions and suppositions	Segmentation of pipelines	Significant	Significant	Significant
	Degree of corrosion (e.g minor, moderate, severe)	Moderate	Moderate	Moderate
	Defined set of features (e.g temperature, humidity factors, pipelines wall thickness)	Moderate	Significant	Moderate
Conclusions and Suggestions				
Conclusions	Since sub-main uncertainty factors are mostly inherent with moderate uncertainty, it can be concluded that overall classification outputs involve with moderate uncertainty. Under these circumstances, surprising events may likely to occur.			
	Based on the degree of criticality, the uncertainty factors that can generate the occurrence of surprising outcomes are the quality of dataset and segmentation of pipelines.			
Suggestions	Due to the overall classification outputs are contained with moderate uncertainty, they must be interpreted in overestimated way. For instance, minor corrosion must be seen as moderate and moderate corrosion to severe.			

By having information as given in table 5.3, we could have a broader insight about the aspects of uncertainty that may involve in the classification outputs and that could provoke surprising outcomes to happen. Furthermore, any suggestions can also be defined such as in how adopting supervised machine learning results to be the decision support concerning many inputs contained with uncertainty. It should be noted that when the overall uncertainties are assessed as moderate or significant, hence, risk should be overestimated rather than underestimated because uncertainty can increase risk (Muhlbauer, 2004). That is why, in table 5.3 suggestion is made to interpret corrosion in an overestimated way. In addition, by referring to the outputs of this method, any improvements for further analysis to reduce uncertainty can be also determined. More importantly, it can be utilized as well for guidance in managing risk under uncertainty.

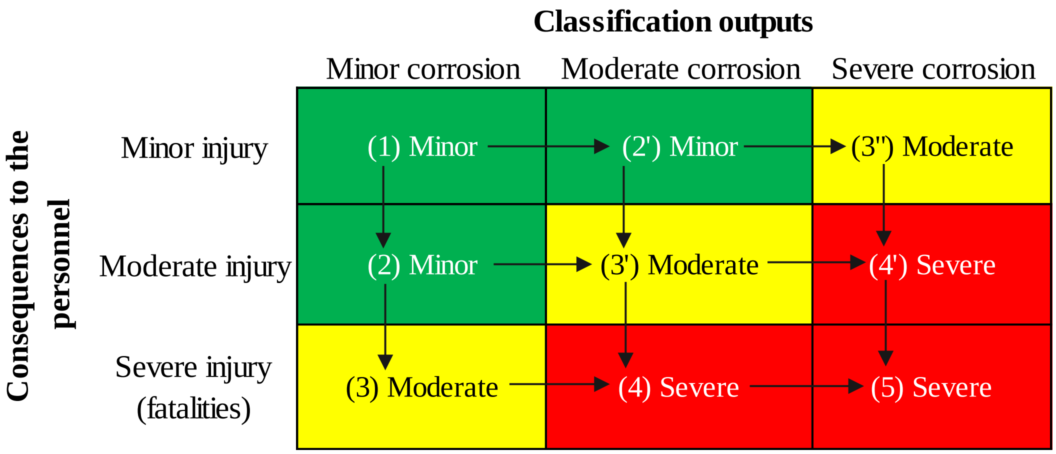
5.3. The Implementation of Qualitative Risk Matrices

In order to reflect the risk aspects in the decision support, one can adopt a method such qualitative risk matrices. Risk matrices have been commonly used in risk management to provide a clear framework in ranking and prioritizing risk (Anthony Tony Cox, 2008). To be more clearly, it can produce valuable information about setting risk priorities, identify which risks that are needed to take into consideration deeply and which risks that decision makers can disregard or postponed because it is judge as low (Anthony Tony Cox, 2008). Furthermore, the implementation of risk matrices can assist the decision makers in evaluating risk, whether it is acceptable or not acceptable (Muhlbauer, 2004). Thereby, they can decide assuredly which risk reducing measures that should be applied immediately to mitigate risk being faced.

Risk matrices can be applied by comparing risk assessment outputs with the consequences analysis outputs (Aven, 2015; Lu et al., 2015). It is thus important to perform consequences analysis before employing this approach. That analysis should be reflected to the performance of barriers, the level of the physical impacts and damage to personnel, environment and assets as stated by NORSOK Z-013 (Association, 2010). Also, it should be assessed based on the specified event from risk assessment outputs.

To illustrate the application of qualitative risk matrices for this case, assuming risk analysts have carried out consequences analysis to the personnel when each severity of corrosion (minor, moderate, severe) should occur. The results of consequences analysis are given in the qualitative such as minor, moderate, and severe injury (fatality). Once information about corrosion and its consequences have been produced, then, they can be compared through risk matrices to generate an insight about the risk level being confronted.

In this part, risk matrices (see figure 5.3) that will be used as an illustration are built based on reference from (Elmontsri, 2014) where arrows and numerical grade are included to direct region that has low to high risk. The arrows that are set in the multi-color box is to give direction from region of lower risk to higher risk (Elmontsri, 2014). Specification of risk level can be done by visualizing through multi-colors boxes (green, yellow, and red). Each box is fulfilled with a number from 1 to 5 to grade the risk, with 1 is indicated as the lowest risk and 5 is the highest risk. The numerical grades that are denoted by prime (‘) and double prime (“) are to demonstrate they have the same relative risk level regarding the close regions that connected by arrows, the risk in these regions can be highly different and not necessarily identical (Elmontsri, 2014).



Classification outputs

Consequences to the personnel	Minor corrosion	Moderate corrosion	Severe corrosion
Minor injury	🟩 (1) Minor	🟩 (2') Minor	🟨 (3'') Moderate
Moderate injury	🟩 (2) Minor	🟨 (3') Moderate	🟥 (4') Severe
Severe injury (fatalities)	🟨 (3) Moderate	🟥 (4) Severe	🟥 (5) Severe

Figure 5.3 Illustration of risk matrices for the case of corrosion in the pipelines

As we can see from the figure above, risk matrices comprise with predicted outputs based on supervised machine learning and consequences analysis outputs to the personnel. To interpret the level of risk from this method, let us assume corrosion in the specified segment of pipelines is forecasted to be minor corrosion. Since our intelligence has a limitation in foreseeing future situation, the consequences of minor corrosion are uncertain. It can be minor, moderate, or severe injuries. If the impacts to the personnel is minor or moderate, the risk level under such conditions would be placed at number 1 and 2 respectively which they are minor risk. Meanwhile, if the consequences are severe and it can lead to loss of human lives, the risk level would be placed at number 3 which means risk within the moderate level. Table 5.4 shows detail information about the level of risk under a various degree of corrosion and its consequences.

Table 5.4 Illustration of interpretation of risk level based on qualitative risk matrices approach

Risk Matrices Region		Risk Level
Classification Outputs of Corrosion	Consequences to the personnel	
Severe corrosion	Severe injury	5
Severe corrosion	Moderate injury	4
Moderate corrosion	Severe injury	4'
Moderate corrosion	Moderate injury	3'
Minor corrosion	Severe injury	3''
Severe corrosion	Minor injury	3
Moderate corrosion	Minor injury	2
Minor corrosion	Moderate injury	2'
Minor corrosion	Minor injury	1

It is clear that by adopting qualitative risk matrices, risk analysts can produce clear information about the risk level under various corrosion severity and its consequences (see table 5.4). What is more, they can provide suggestions that can support the decision makers in preventing pipelines leakage. That suggestions can contain information about:

- Type of risk reducing measures that shall be implemented to reduce risk level
- Risk that should be addressed immediately by risk reducing measures

By referring to the suggestions above, decision makers can have an insight on what prevention actions that should be taken shortly. But, before establishing such suggestions, it would be essential to setting the degree of risk that shall be prioritized. The illustration of risk priority can be seen as follow:

- 1st priority, risk is in the level of 5, 4, and 4'.
- 2nd priority, risk is in the level of 3' and 3".
- 3rd priority, risk is in the level of 3, 2, 2', and 1.

It should be noticed that, in practice, defining requirements for prioritizing risk can be varied. It will hinge on to the context of the assessments. After priority have been set, risk analysts can construct information about risk reducing measures that shall be chosen with respect to the risk level and risk priority as can be seen in the table 5.5 and 5.6.

Table 5.5 Illustration of suggested risk reducing measures to prevent pipelines leakage

Risk Level	Risk Priority	Risk Reducing Measures Strategies
5	1st priority	— Choose high quality material of pipelines that can withstand with severe corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Set corrosion inhibitors. — Do inspection and maintenance. — Implementation of measures should be done immediately
4	1st priority	
4'	1st priority	

Table 5.6 Illustration of suggested risk reducing measures to prevent pipelines leakage (continued)

Risk Level	Risk Priority	Modified Risk Reducing Measures Strategies
3'	2nd priority	— Choose high quality material of pipelines that can withstand with moderate corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Perform maintenance — Implementation of measures should be also done shortly to avoid risk becomes significant.
3''	2nd priority	
3	3rd priority	— Choose high quality material of pipelines that can withstand with minor corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Implementation of measures can be postponed. But, it still need to be addressed to keep risk level within minor.
2	3rd priority	
2'	3rd priority	
1	3rd priority	

We can notice that the overall suggestions of safety measures in the tables above are to reduce the severity of corrosion instead of minimizing its consequences. This is because defining measures to decrease potential consequences is more complex and there would be changes in some aspects of product streams and/or surrounding pipelines that may contribute the greatest change (Muhlbauer, 2004). Thereby, it is preferable to minimize the risk by decreasing the failure potential (Muhlbauer, 2004). Furthermore, communicating decision supports as in table 5.5 and 5.6 to the decision makers can assist them in weighing risks that should be promptly mitigated with defined safety measures.

Considering the outputs of supervised machine learning and consequences are likely to contain with uncertainty because they are obtained from prediction. The level of risk must be seen with caution, particularly if the uncertainty involved are assessed to be moderate or significant. Under such circumstances, the risk level should be overestimated by considering the risk level in the 3rd priority to be 2nd priority and 2nd priority to be 1st priority. Since there is a changing in the interpretation of risk level, suggestions of safety measures should be modified before delivering to the decision makers as can be seen in table 5.7.

Table 5.7 Illustration of modification in the suggested risk reducing measures due to considering uncertainty

Risk Level	Original Risk Priority	Modified Risk Priority Concerning Uncertainty	Modified Risk Reducing Measures Strategies
5	1st priority	1st priority	— Choose high quality material of pipelines that can withstand with severe corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Set corrosion inhibitors. — Do inspection and maintenance. — Implementation of measures should be done immediately
4	1st priority	1st priority	
4'	1st priority	1st priority	
3'	2nd priority	1st priority	— Choose high quality material of pipelines that can withstand with severe corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Set corrosion inhibitors. — Do inspection and maintenance. — Implementation of measures should be done immediately
3''	2nd priority	1st priority	
3	3rd priority	2nd priority	— Choose high quality material of pipelines that can withstand with moderate corrosion. — Seal internal and external pipelines wall with coating, anodic and cathodic protection. — Perform maintenance. — Implementation of measures should be also done shortly to avoid risk becomes significant.
2	3rd priority	2nd priority	
2'	3rd priority	2nd priority	
1	3rd priority	2nd priority	

Overestimating risk level can provide both advantage and disadvantage. The benefit is that we can anticipate properly on events and consequences that may happen (surprisingly) with stronger safety measures. Hence, the losses and accidents might not generate to the large extent. However, the drawback is that it can be costly for the companies because they have to allocate their resources for more risk reducing measures.

Principally, the company addresses risk reducing measures to balance between gaining opportunities and avoiding losses and accidents. If the attentions are more focused on avoiding losses and accidents by applying more measures, they can obtain a lower opportunity of what is expected. In contrast, if they put more considerations on achieving opportunities, they may get more losses if accidents occur. In practice, the companies often face this gambling situation so that economical perspectives need to be considered before making a decision to define measures that should be selected.

Overall, by applying risk matrices we could reflect the aspects of risk thoroughly. Information about the risk level can be acquired based upon the degree of corrosion and the consequences of it. Furthermore, by referring to the identified risk level, risk analysts can provide suggestions for decision makers regarding preventing actions that should be taken immediately to reduce risk level until within an acceptable criteria of company and/or authority. However, risk level that is obtained from risk matrices should be used with caution because it is associated with uncertainty to some extent and it is based on qualitative judgments.

6. Final Discussion

6.1. The Suitability of Supervised Machine Learning for Predicting Corrosion

Corrosion in the pipelines is regarded as serious issues that cannot be taken for granted. It has the capability to reduce pipeline's wall thickness until causing leakage if it is not taken care of properly. The magnitude of leak sizes will vary starting from minor perforation to breaks the pipes, which it will depend on corrosion severity.

The consequences of leakage, even in a small pinhole, can initiate subsequent accidents to occur. Initial accident that may occur is fluid release. Fluid release can generate to fire and/or explosion if it reacted with the combustible sources. Even the mist or dust can be one of the sources. Meanwhile, in real life, we often regard those factors as trivial things. Considering to that issues, it is thus crucial to treat corrosion properly, therefore, the incident of pipelines leakage during transportation of hazardous fluid can be avoided and so too does the accidents that can endanger human lives, environment, company's assets, and reputation.

In order to keep the pipelines from corrosion, appropriate risk reducing measures should be implemented. There are various types of measures that can be applied to impede electrochemical process reacts on the surface of pipes, such as corrosion inhibitors, internal and external coatings protection, corrosion inspections, and many more. To support decision makers in defining which of them must be addressed, the severity of corrosion must be predicted.

However, forecasting corrosion in the pipelines can be acknowledged as a very difficult task to perform because of some factors. The pipelines that are installed in surface and sub-surface within thousand and even million miles can lead to difficulties in identifying which asset that may experience corrosion. Moreover, potential corrosion can be hardly to detect due to environment in the entire pipelines always changes because of weather, composition fluids, and so on. To deal with these circumstances, corrosion should be prognosticated per pipelines section under numerous causes that may lead such issues to occur. A breakthrough approach that has ability to make prediction under various situations should be adopted to solve these problems. For this case, one can employ predictive analytics tool such a supervised machine learning.

Supervised machine learning has been recognized can measure uncertainty only from the data even it is big data. That technology is part of artificial intelligence that has the capability to establish prediction with minimal human intervention. By using this tool, the prediction will be made based on the learning process by the algorithm. More specifically, the algorithm analyzes the dataset in the training set to recognize the relationship between the inputs (set of features) and outputs (classifications). In this part, although there are numerous inputs and outputs, the predictive analytics tool still capable to find its patterns.

The process of learning from the training set will generate predictive models that can be used to make automate indication of new data. However, before that, such models should be evaluated. It can be done by input dataset from the testing set into the predictive models. By doing so, we can have an insight on how many data that gives wrong or correct classifications. Also, the accuracy, precision, and recall of the performance of the algorithm can be known. Thereby, we may determine whether such algorithm would be suitable to apply for making a prediction of the condition being analyzed. If it is assessed to be appropriate, then it can be used for making prediction of actual condition. The output of this approach will be presented in classification. To have more understanding about the results of supervised machine learning, we can visualize table 4.1.

By considering the type of the prediction outputs as given in table 4.1, it can be understood that supervised machine learning has the ability to describe the occurrence of corrosion based on different degree of severity and conditions synchronously. Indeed, by performing this tool, risk analysts can identify corrosion on almost all pipelines under numerous phenomena. Comprehensive information about corrosion can thus be acquired and they can be more understanding on what can go wrong in the future. Moreover, the techniques of predictions that measure uncertainty by classification can help them in prioritizing corrosion that needs to be handled immediately.

What is more, forecasting corrosion using this predictive tool can generate results accurately and quickly even though the data are large and variant. This is because algorithm of supervised machine learning has the ability to observe relationships between inputs and outputs that should have been produced. The more often the learning algorithm is trained by the data, the more accurate and faster the prediction will be on the next time around. In addition, by considering the algorithm’s abilities, this tool can be used to monitor corrosion in the pipelines that have been identified whether it keeps on the same level of severity or gets lower when risk reducing measures are applied or otherwise. If it is detected higher, thus, modification of preventing measures that have been addressed should be done to keep corrosion within a safety level. Moreover, it can be utilized to track new corrosion in the pipelines, therefore, further mitigation plan can be defined.

Nevertheless, like other measurement of uncertainty tools, supervised machine learning has some weaknesses. To measure uncertainty, this tool requires background knowledge such as data, algorithm, and assumptions. Meanwhile, such background knowledge can associate with uncertainty due to several factors. The factors that can lead them integrated with uncertainty will be discussed in the paragraph below.

Data, it can collaborate with uncertainty because the dataset is made based on individual(s) knowledge. If they have lack of knowledge and give wrong examples in the dataset about future phenomena, hence, this predictive tool will produce an incorrect prediction. Also, the data that is collected based on historical performance may not describe the actual situation.

For the algorithm, it can also be inherent with uncertainty due to the mechanism of the algorithm in learning data and generating prediction cannot be completely understood by human. It is thus like a black boxes prediction and we may wonder whether the prediction results are true.

In the meantime, the assumptions are the factors that can likely to involve with uncertainty. This is due to fundamentally we, as a human, have some limitations in understanding the real-world situation thoroughly. Therefore, the important aspects related to the upcoming event can be neglected.

The uncertainty involved in each background knowledge can cause the prediction outputs does not represent accurately phenomena that can happen in the future. Furthermore, utilizing these results to be the decision basis would not be robust. This is because the aspects of uncertainties are not taken into account properly using this predictive analytics tool. As a result, surprising outcomes can happen. It is known that the impacts of surprising events can be more disasters for human values. Besides, supervised machine learning results also do not reflect the aspect of risk. Therefore, we have no insight whether such corrosion is acceptable when it occurs and which safety measures that need to be implemented immediately.

All things considered, by weighing the benefits and drawbacks of supervised machine learning, this tool is considered can be appropriate for predicting corrosion in the pipelines. This is because such tool can forecast corrosion under various severity and factors accurately and fast. However, if the purpose of performing prediction is to support decision makers in choosing risk reducing measures to prevent pipelines leak, the classification outputs that are generated based on this technology should not be trusted entirely to be the basis of the decision. The aspects of uncertainty and risk are overlooked. Therefore, it is not robust to be the only decision support because both aspects are important to consider when managing hazard of an operation. Therefore, some approaches are needed to strengthen the decision basis.

6.2. The Role of A Suggested Framework in Improving Decision Basis

Based on the previous discussion, supervised machine learning is judged to be appropriate for predicting corrosion in the pipelines. Nonetheless, the classification outputs that are generated from such tool are considered not powerful enough to be the decision support, especially for preventing pipelines leakage. The reason is because those results do not fully reflect the important aspects such as uncertainty and risk.

There are specified aspects of uncertainty and risk that are ignored by this tool. In this case, measuring uncertain conditions using supervised machine learning can disregard the uncertainty in the background knowledge used such as the data (e.g temperature, humidity factors, pipelines wall thickness, dataset, etc.), algorithm (artificial neural network, decision tree, logistic regression, etc.) and assumptions (segmentation of pipelines, degree severity of corrosion, etc.). Therefore, the results can be skeptical for the decision makers whether it is true and accurate to be the decision supports. Also, by simply using such results, surprising outcomes can be likely to occur.

For the risk, the aspects that are not considered is the acceptability of the occurrence of corrosion and its consequences when it should happen. In the meantime, such information is needed to be produced so that the level of risk being faced can be identified. Furthermore, it can help risk analysts in defining more assuredly which circumstances that need implementation of isk reducing measures shortly.

It is thus clear that by ignoring uncertainty and risk there would be some problems that can exist. Decision makers may not trust the classification outputs to be used as the basis of decision even if the accuracy, precision, and recall of that predictive tool are good. Moreover, surprising events could happen and bring more severe accidents. Furthermore, decision makers can be difficult to decide preventing actions that should be selected immediately due to risk analysts do not provide information regarding to it.

To deal with those problems, more extensive analyses beyond supervised machine learning results should be implemented to improve decision support. In this case, a suggested framework that contains with two methods are recommended, which are: the hidden uncertainty analysis and risk matrices will be used to develop decision basis. Both methods will not be performed simultaneously because each of them has its own issues to dealt with. But, it does not imply that both elements cannot be collaborated.

In this thesis work, the hidden uncertainty analysis is performed to lead the supervised machine learning outputs have a better reflection towards the aspects of uncertainties. It is done by identifying uncertainty and sub-uncertainty factors which then they are assessed with respect to the level of uncertainty, sensitivity and criticality. By performing this method, the overall degree of uncertainty that involved in the predicted outputs can be detected. Hence, the suggestions of how the results of such predictive analytics tool should be interpreted can be defined under the consideration of uncertainty. In this part, if the overall level of uncertainty is examined to be moderate or significant. Therefore, the classification results must be diagnosed in overestimated way, for example by visualizing minor to medium corrosion, and medium to severe corrosion. This is done to anticipate the uncertainty that are inherent in the prediction outcomes and to avoid the occurrence of surprising events.

After the aspects of uncertainty have been treated, we continue to follow the next stage which is to perform qualitative risk matrices. That method is conducted to consider the risk aspects in making decision support to prevent pipelines leak. However, to use this approach we need to figure out the consequences in each identified severity of corrosion. Therefore, the risk level can be obtained by comparing the classification results and its consequences. Having information about the risk level can assist risk analysts in defining which risks that are acceptable and not acceptable. It also can help them in determining recommendations of safety measures that shall be taken. More importantly, they can produce information which prevention actions that should be chosen due to the risk is not within the safety level. Nonetheless, it should be noted that in prioritizing risk level, we need to consider the degree of uncertainty. If it assigned to be moderate or significant based on a hidden uncertainty analysis, then, the risk level must be seen in underestimated way because uncertainty can increase risk. It can thus be meant that the more risk reducing measures that should be implemented to minimize risk.

Overall, by employing such a suggested framework, decision support based on supervised machine learning can be more robust. The aspects of uncertainty and risk are included. Also, the suggestions of how they must be handled can be known. Thus, the decision makers can be more understand and easily review and judge such analyses outputs for taking decision.

7. Conclusions and Suggestion for Further Work

7.1. Conclusions

The present thesis has conducted to analyze the appropriateness of supervised machine learning to be a tool for prognosticating corrosion in pipelines and its predicted results for being decision supports in avoiding pipelines leakage. Based on literature review and brainstorming with supervisor, the main findings of this work will be summarized as follows:

1) Identifying corrosion in pipelines is quite difficult due to changing of environment in almost all pipelines and their installation that are commonly constructed in very long distances also placed in surface and sub-surface. In such circumstances, potential corrosion that may attack the surfaces of the pipes becomes unstable and so too does the risk picture.
2) Forecasting corrosion only with human intelligence may hide important aspects of uncertainties because we have limitations in understanding the overall real-world situations and in converting all background knowledge (e.g data, assumptions, etc.) to foresee what can go wrong in the future. As the consequences, predicted results may deviate from the actual conditions and surprising outcomes may likely to occur.
3) Supervised machine learning has the ability to generate a prediction and classification about the data accurately and promptly, even though it is massive and complex data.
4) Several benefits can be acquired by performing supervised machine learning to predict corrosion. Detail information regarding the type, degree of severity, and factors that can lead to corrosion can be detected simultaneously. Moreover, the ability of the tool in classifying data makes risk analysts can identify, prioritize, and monitor corrosion in pipelines under numerous factors rather than only single cause without taking much effort of human intervention and consume so much time.
5) The drawbacks of using supervised machine learning are the important aspects of uncertainty and risk are not reflected comprehensively in the predicted outputs. As a result, what has been forecasted about corrosion might be wrong (e.g minor corrosion may turn out severe in reality). Furthermore, decision-making in defining preventing actions to mitigate pipelines leakage can be difficult because lack of information about the risk being faced. In short, misinterpretation in visualizing future phenomena and in choosing safety measures can cost companies on a large scale.
6) By weighing the advantages and disadvantages of supervised machine learning, that tool would be suitable for predicting corrosion in pipelines. This is because the instability of potential corrosion as a result of changing environment, lengths and locations of pipes can be captured by using this tool. Therefore, the assessment of corrosion can be carried out in more detail as well as in a wide context. However, in view of its drawbacks, the outputs based on this tool are considered weak to be the decision-making support in preventing leakage in pipelines. Therefore, some approaches should be implemented to strengthen the decision basis from such predictive tool.

7.1.1. A suggested framework to develop decision basis for preventing pipelines leakage

Preventing leakage in the pipelines by simply using decision basis based on supervised machine learning may lead to the failure in managing safety. This is because two crucial aspects such as uncertainty and risk are not reflected well. Meanwhile, the uncertainty is vital to be considered properly to avoid the occurrence of surprising events, whereas the risk is taken into account for diagnosing the level of hazards being faced. Considering by reflecting both aspects decision-making supports can be more robust; thus, they should be examined in this case. Involving them in the basis of decision can assist decision makers to weigh properly and choose assuredly safety measures that should be taken to prevent pipelines leakage.

To reflect the aspects of uncertainty and risk in the decision support, one can adopt a suggested framework. The framework involves two methods, hidden uncertainty analysis and qualitative risk matrices, which both of those shall be carried out gradually. But, the first approach that should be performed is hidden uncertainty analysis. By performing this method, we can have insight regarding the overall degree of uncertainty involves in the predicted outputs from supervised machine learning. Afterwards, qualitative risk matrices become the next method that should be taken. The outcome of that approach is information regarding risk level that would be faced in the future can be obtained. By integrating both methods, risk analysts can define the prevention actions considering the degree of uncertainty and risk involved. In this part, if the overall level of uncertainty diagnosed to be moderate or significant, therefore, the level of risk should be interpreted in an overestimated way. Seeing risk in such way can be implied that there would be larger number of risk reducing measures that should be implemented to anticipate the occurrence of surprising events and other hazards. This can be good for the companies in reducing losses, accidents, and other disasters. However, it can be much costly for them and may reduce their expected benefits to some extent.

7.2. Suggestion for further work

Since the work of this thesis is based on a qualitative approach and there are also limitations in the scope of work, data, and amount of time. Thus, the information produced to answer the issues being analyzed may not be complete. Therefore, several suggestions for further work are provided to improve and strengthen the suitability of supervised machine learning approach to forecast corrosion and its outcomes to be the basis of the decision in avoiding pipelines leak.

1) Performing quantitative analysis
   Due to running supervised machine learning is beyond the scope of this thesis, hence, the predicted outputs that given in this work are based on illustration. Thus, some important information can be neglected and mislead by the author. By executing quantitative analysis, the examination of whether this tool is appropriate to be the tool for predicting corrosion can be more robust and understood well.
2) Analyzing cost-effective analysis and ALARP principle
   By adopting a suggested framework to improve decision support based on supervised machine learning, the companies are advised to overestimate the occurrence of corrosion severity when the overall degree of uncertainty is moderate or significant. In other words, it can be implied that they should invest more in risk reducing measures, meanwhile they

have limitations in resources. It is thus important to use methods such as cost-effective analysis and ALARP principle to investigate whether the suggestion that is proposed from the framework is fruitful when considering the trade-off perspectives.

References

Abrahamsen, E., Aven, T., & Iversen, R. (2010). Integrated framework for safety management and uncertainty management. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 224*(2), 97-103.

Ahammed, M., & Melchers, R. (1996). Reliability estimation of pressurised pipelines subject to localised corrosion defects. *International Journal of Pressure Vessels and Piping, 69*(3), 267-272.

Anthony Tony Cox, L. (2008). What's wrong with risk matrices? *Risk analysis, 28*(2), 497-512.

Association, N. O. I. (2010). Risk and Emergency Preparedness Assessment, NORSOK Standard Z-013, Rev. 3. In: Lysaker, Norway: Standards Norway.

Aven, T. (2008). A semi-quantitative approach to risk analysis, as an alternative to QRAs. *Reliability Engineering & System Safety, 93*(6), 790-797.

Aven, T. (2013). Probabilities and background knowledge as a tool to reflect uncertainties in relation to intentional acts. *Reliability Engineering & System Safety, 119*, 229-234.

Aven, T. (2014). *Risk, surprises and black swans: fundamental ideas and concepts in risk assessment and risk management*: Routledge.

Aven, T. (2015). *Risk analysis*: John Wiley & Sons.

Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*: InTech.

Azure, M. Introduction to Machine Learning in the Azure cloud. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-machine-learning

Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Bolzon, G., Boukharouba, T., Gabetta, G., Elboujdaini, M., & Mellas, M. (2011). *Integrity of Pipelines Transporting Hydrocarbons: Corrosion, Mechanisms, Control, and Management*: Springer science & business media.

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms. Retrieved from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W., & Jeyachandra, B. (2016). *Data driven production forecasting using machine learning.* Paper presented at the SPE Argentina Exploration and Production of Unconventional Resources Symposium.

Chilingarian, G. V. (1989). Corrosion and water technology for petroleum producers: Loyd W. Jones, Oil and Gas Consultants International, Inc.(OGCI), Tulsa, Oklahoma, 202 pp. In: Elsevier.

Choi, J., Goo, B., Kim, J., Kim, Y., & Kim, W. (2003). Development of limit load solutions for corroded gas pipelines. *International Journal of Pressure Vessels and Piping, 80*(2), 121-128.

CORROSIONPEDIA. Soil Corrosion. Retrieved from https://www.corrosionpedia.com/definition/1465/soil-corrosion

CrashCourse (2017). [Machine Learning & Artificial Intelligence: Crash Course Computer Science #34].

CrowdFlower. What is training data? Retrieved from https://www.crowdflower.com/what-is-training-data/

da Cunha, S. B. (2016). A review of quantitative risk assessment of onshore pipelines. *Journal of Loss Prevention in the Process Industries, 44*, 282-298.

dataminingincae (Producer). (2014). Gretl Tutorial 5: Forecasting and Confusion Matrix. Retrieved from https://www.youtube.com/watch?v=YHIfgO8H7Xo

Dey, P. K. (2004). Decision support system for inspection and maintenance: a case study of oil pipelines. *IEEE transactions on engineering management, 51*(1), 47-56.

Dey, P. K. (2006). Integrated project evaluation and selection using multiple-attribute decision-making technique. *International Journal of Production Economics, 103*(1), 90-103.

Dlouhy, J. A. (2013). Pipelines are safer than trains and trucks, report says. Retrieved from https://fuelfix.com/blog/2013/10/17/pipelines-safer-than-trains-and-trucks-report-says/

Ekine, A., & Emujakporue, G. (2010). Investigation of corrosion of buried oil pipeline by the electrical geophysical methods. *Journal of Applied Sciences and Environmental Management, 14*(1).

Elmontsri, M. (2014). Review of the strengths and weaknesses of risk matrices.

Engineers, r. k. C. Pipeline Integrity Assessment. Retrieved from http://www.rkconsult.nl/services/pipeline-integrity/

Flach, P., & Matsubara, E. (2008). *On classification, ranking, and probability estimation.* Paper presented at the Dagstuhl Seminar Proceedings.

Flage, R., & Aven, T. (2009). Expressing and communicating uncertainty in relation to quantitative risk analysis. *Reliability & Risk Analysis: Theory & Application, 2*(13), 9-18.

FluidDataReporting (Producer). (2013). Internal Corrosion Control for Oil and Gas Pipelines. Retrieved from https://www.youtube.com/watch?v=9bb-B357oQA&t=182s

Furchtgott-Roth, D. E., & Green, K. P. (2013). Intermodal safety in the transport of oil.

GeeksforGeeks. Confusion Matrix in Machine Learning.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452.

GL, D. (2017). DNV GL Machine Learning – Predict external corrosion on oil and gas pipelines. Retrieved from https://github.com/readyforchaos/Predict-external-corrosion-on-oil-and-gas-pipelines

Gross, M. (2010). *Ignorance and surprise: Science, society, and ecological design*: MIT Press.

Guikema, S. D. (2009). Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory. *Reliability Engineering & System Safety, 94*(4), 855-860.

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Inc., S. I. Machine Learning What it is and why it matters. Retrieved from https://www.sas.com/en_us/insights/analytics/machine-learning.html

International, N. (2010). *Standard Recommended Practice: Pipeline External Corrosion Direct Assessment Methodology*: NACE International.

Kalogirou, S. A. (2001). Artificial neural networks in renewable energy systems applications: a review. *Renewable and sustainable energy reviews, 5*(4), 373-401.

Kamalnath, T. B. a. V. (2017). Controlling machine-learning algorithms and their biases.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology, 64*(5), 402-406.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*: MIT Press.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

Lloyd, G. O. *Atmospheric Corrosion*. Retrieved from http://www.npl.co.uk/upload/pdf/atmospheric_corrosion.pdf.

Lu, L., Liang, W., Zhang, L., Zhang, H., Lu, Z., & Shan, J. (2015). A comprehensive risk evaluation method for natural gas pipelines by combining a risk matrix with a bow-tie model. *Journal of Natural Gas Science and Engineering, 25*, 124-133.

Mahjania, M., Jalilia, S., Jafariana, M., & Jaberia, A. Prediction of metal corrosion using feed-forward neural networks.

Maini, V. (2017). Machine Learning for Humans, Part 2.2: Supervised Learning II. Retrieved from https://medium.com/machine-learning-for-humans/supervised-learning-2-5c1c23f3560d

MathWorks. What Is Machine Learning? 3 things you need to know. Retrieved from https://se.mathworks.com/discovery/machine-learning.html

Meresht, E. S., Farahani, T. S., & Neshati, J. (2011). Failure analysis of stress corrosion cracking occurred in a gas transmission steel pipeline. *Engineering Failure Analysis, 18*(3), 963-970.

Miesner, T. O., & Leffler, W. L. (2006). *Oil & gas pipelines in nontechnical language*: PennWell Corporation.

Milan, J. (2016). supervised-workflow-machine-learning. Retrieved from http://blog.bidmotion.com/2016/06/23/good-morning-have-you-used-machine-learning/

Montgomery, S. (Producer). (2016). Data Mining and Machine Learning via Support Vector Machines. Retrieved from http://slideplayer.com/slide/9493622/

Muhlbauer, W. K. (2004). *Pipeline risk management manual: ideas, techniques, and resources*: Elsevier.

Nyborg, R. (2005). Controlling internal corrosion in oil and gas pipelines. *business briefing: exploration & production: the oil & gas review, 2*, 70-74.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128-138.

Papavinasam, S., Doiron, A., & Revie, R. W. (2010). Model to predict internal pitting corrosion of oil and gas pipelines. *Corrosion, 66*(3), 035006-035006-035011.

Popoola, L. T., Grema, A. S., Latinwo, G. K., Gutti, B., & Balogun, A. S. (2013). Corrosion problems during oil and gas production and its mitigation. *International Journal of Industrial Chemistry, 4*(1), 35.

Ripley, B. D. (2007). *Pattern recognition and neural networks*: Cambridge university press.

Rosa, E. A. (1998). Metatheoretical foundations for post-normal risk. *Journal of risk research, 1*(1), 15-44.

Rosa, E. A. (2003). The logical structure of the social amplification of risk framework (SARF): Metatheoretical foundations and policy implications. *The social amplification of risk, 47*.

Sadawi, N. (Producer). (2014). Evaluating Classifiers: Confusion Matrix for Multiple Classes. Retrieved from https://www.youtube.com/watch?v=FAr2GmWNbT0

Samimi, A., & Zarinabadi, S. (2011). An Analysis of Polyethylene Coating Corrosion in Oil and Gas Pipelines. *Journal of American science, USA*.

Selvik, J., & Aven, T. (2009). An extended Bayesian updating approach to support product selection based on performance testing—a drilling jar case. *Reliability, risk and safety: theory and applications, 2*, 813-819.

SHAH, T. (2017). *Train, Validation and Test Sets*.

Siegel, E. (2013). Predictive analytics. *Hoboken: Wiley*.

SINTEF, S. (2003). *Handbook for Fire Calculations and Fire Risk Assessment in the Process Industry*. In. Retrieved from https://ia800506.us.archive.org/5/items/SINTEF2003HandbookForFireCalculationsAndFireRiskAssessmentInTheProcessIndustry/SINTEF%20-%202003%20-%20Handbook%20for%20Fire%20Calculations%20and%20Fire%20Risk%20Assessment%20in%20the%20Process%20Industry.pdf

Smart, J. S., & Smith, G. L. (1991). Pigging and chemical treatment of pipelines.

Supriyatman, D., Sidarto, K. A., Suratman, R., & Dasilfa, R. (2012). Artificial Neural Networks for Corrosion Rate Prediction in Gas Pipelines.

University, A. (Producer). (2015). Confusion Matrix & Model Validation. Retrieved from https://www.youtube.com/watch?v=bpsmoQdoYpQ

Vinnem, J.-E. (2014). Offshore Risk Assessment vol 1. In: Springer.

Vtorushina, A. N., Anishchenko, Y. V., & Nikonova, E. (2017). *Risk Assessment of Oil Pipeline Accidents in Special Climatic Conditions.* Paper presented at the IOP Conference Series: Earth and Environmental Science.

Appendix A

A New Framework for Improving Decision Basis based on Supervised Machine learning to Support Preventing Pipelines Leakage due to Corrosion

Abstract

Leakage in the pipelines that is caused by corrosion can pose a significant hazard not only to the company's assets and reputations but also to the human lives and environment. It is thus crucial to implement some safety measures that can handle corrosion effectively. To support decision-makers choose appropriate measures, phenomena of corrosion should be foreseen in prior. Thereby, the measures can be addressed in line to the problem being faced. However, performing such prediction can be very difficult considering environment along the pipelines keeps changing and installation of such assets can be in thousands and even million miles also in the surface and sub-surface. To deal with such circumstances, a prediction under various corrosion severity and its causes should be carried out per pipelines section. Nevertheless, generating such prediction only using human intelligence may ignore many important aspects that may occur in actual situations. As a result, a surprising outcome can likely to occur. To minimize crucial aspects being overlooked, one can adopt technology such a supervised machine learning to assist in making prediction. That tool is recognized as to be able to produce accurate and fast prediction based on big, various, and complex data. Nonetheless, like many other tools, supervised machine learning has some drawbacks that makes the outputs are not robust to be the decision support. The drawbacks are the aspects of uncertainty and risk are not reflected comprehensively. In this paper work, a new framework is proposed to handle such drawbacks and develop decision basis. The proposed framework contains two methods; hidden uncertainty analysis and qualitative risk matrices which they will be integrated to improve decision support.

1. INTRODUCTION

In oil and gas industry, in comparison to trucks and trains, pipelines are regarded as the safer instruments, lower costs, and more environmentally friendly for transmitting crude oil, natural gas, and petroleum products from place to place. However, when there is an accident such as leakage, the amount of fluids spillage are bigger from pipelines because it can carry fluid in capacity about 70% than trains and roads that are able to transport around 3-4% (Dlouhy, 2013). Taking into account that volume of fluid release, hence, pipelines can be assumed not entirely safe in transporting hazardous fluids. This is because major accidents can happen starting from that pipelines leak incidents.

Flammable and poisonous fluids that release from pipelines can trigger sequent accidents to happen. Fire and/or explosion may occur if that fluids associated with the combustible sources (e.g dust, mist, frictional sparks, auto ignition, etc.) (SINTEF, 2003). They can be expanded to the large extent if the defined safety barriers cannot handle such accidents. Obviously, under such situations, health, safety, and environment are at a stake.

Considering the risk of pipelines leak may jeopardize human values, hence, factors that lead to leakage should be dealt properly. Generally, there are various causes that can lead pipelines to experience leakage. But, the main causes of it is corrosion (Ahammed & Melchers, 1996; Choi et al., 2003; da Cunha, 2016; Dey, 2004, 2006; Vtorushina et al., 2017). This is due to the fact that the base material of pipelines is mostly made of metal. Metal pipelines can encounter corrosion because there is an electrochemical process such as anion, cathode, and electrolyte that react on its surface (FluidDataReporting, 2013). As the consequences, the thickness of pipelines' wall may decrease. If it is not handled appropriately, it can cause leakage.

Due to corrosion has the capability to remove metal pipelines until forming a leakage, risk reducing measures should be implemented to avoid such issue to occur. To have a better understanding about the measures that should be chosen to tackle corrosion, an assessment toward that issue must be conducted. However, performing corrosion assessment is not an easy task in light of actual conditions that are full of complexities to understand.

In reality, the environmental condition in the entire length of pipes changes intensely due to weather, fluid compositions, and so on. This makes the potential corrosion are hard to detect because there are numerous factors that can induce the occurrence of such incident to some degree. What is more, the location and length of pipelines installation can be also the problems in identifying corrosion. To deal with such situations, corrosion shall be predicted under multifarious conditions. More specifically, it should be forecasted based on different severity and several causes per pipelines segment. Indeed, there would be a huge volume, variant, and complex data that should be gathered and processed to support this prediction.

Only utilizing human intelligence to process such big data can lead to some problems. The first problem is that it may take so much time to produce prediction of what can go wrong in the upcoming event. Meanwhile, in practice, decision supports are needed to be established quickly so that the problems can be solved shortly. Another problem is that many crucial aspects of uncertainty can be overlooked. As the consequences, a prediction may highly deviate from the actual situations and a surprising event may happen. It should be noted that the occurrence of surprising event can bring more disaster to human values.

To confront with the problems above, one can utilize supervised machine learning as the tool to predict corrosion based upon various conditions. This is because that tool is known can generate prediction based on large and complex data accurately and fast. Hence, establishment of decision basis to support prevention of pipelines leak can be done without much effort of human intervention and consume so much time. However, limiting decision support only to this method can ignore uncertainty and risk aspects. In the meantime, both aspects are vital to consider during managing safety of an operation. They can be valuable references in reducing the occurrence of unwanted accidents and other consequences that may harm human lives, environment, assets, and reputation.

Therefore, in this paper, a new framework that involves with two methods: hidden uncertainty analysis and qualitative risk matrices are provided to reflect the aspects of uncertainty and risk. What is more, it is also used to improve decision basis based on supervised machine learning by integrating both methods. All in all, by performing this framework, decision-making supports can be more robust in assisting decision makers prevent pipelines leakage.

This paper is organized as follows. The second section describes existing framework based on supervised machine learning. The third section discusses the method of supervised machine learning in predicting corrosion and being the decision support to avoid pipelines leak, whereas the fourth section provides the suggested framework to develop decision basis based on supervised machine learning. Lastly, the fifth section elucidates the conclusion of this paper.

2. EXISTING FRAMEWORK OF SUPERVISED MACHINE LEARNING

First of all, machine learning is a data science technique that allows computers to use existing data to predict about future behaviors, outcomes, and trends without being explicitly programmed (Azure; Cao et al., 2016; CrashCourse, 2017; Ghahramani, 2015). For the most part, the application of machine learning is based on supervised learning. Supervised learning is an approach for conditions where we have record the outcome data (output) simultaneously with the informative data (input) that could be acquired from historical operation (Guikema, 2009). By using this tool, a prediction will be generated based on finding the relationship between output and input that have been set.

To be more clearly, let assumes and denotes informative data as input (**X**) and desired outcome data as output (**y**) (Guikema, 2009). In this part, the relationship **y** = f(**X**) should be assessed to produce a prediction based upon a given set of input and output. The set of input and output can be known with a dataset. The f**(X)** is unknown function of input and it does not associate with any notion of uncertainty in **y** given **X** (Guikema, 2009). Therefore, f**(X)** is considered will contain large uncertainty. To treat the uncertainty in that parameter, the algorithm and training dataset must be applied into computers to learn the form and parameters of a model approximating f**(X)** so that hopefully will generate results in the right prediction of future circumstance based on new data (Brownlee, 2016; Guikema, 2009).

It is obvious that the technique of supervised machine learning describes uncertainty is different as done by probabilistic risk analysis (PRA), the tool that is commonly used for making prediction. PRA finds the relationship **y**=f(**X)** based on the assessors' background knowledge such as assumptions, historical data, expertise judgments, and many more which then will be used for estimating the failure scenarios or the likelihood of the event. Meanwhile, supervised machine learning observes that relationship by learning from the dataset which afterwards that data will be classified into a particular class. That is why, the outputs of performing supervised machine learning are presented in some classifications.

To illustrate the typical outputs of supervised machine learning, assuming corrosion engineers would like to forecast external corrosion based on the degree severity such as severe, moderate, and minor. To support such prediction, potential factors that can trigger the occurrence of external corrosion should be identified. By having a discussion with some experts, the main causes of the external wall of pipes experiences deterioration are temperature and humidity factor. Furthermore, since the outer surface will be exposed directly to corrosion, thus, it is crucial to include wall thickness as the potential factors that should be concerned in identifying corrosion in pipelines. After all set of inputs and outputs have been defined, they should be set as training dataset and then fed into learning algorithm to discover relationship between inputs and outputs. That learning process will result predictions in classifications as in the table 1.

Table 1 Illustration of supervised machine learning outputs

Training dataset	Set of Features			Supervised Classification output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
	32	55	19	minor corrosion
	20	95	11	severe corrosion
	11	90	14	medium corrosion
new dataset	Set of Features			Supervised Classifiication output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
	20	80	20	?
	13	94	12	?
	22	90	15	?
Predicted results of new dataset	Set of Features			Supervised Classification output
	Temperature in Celcius	Humidity Factor in %	Pipelines Wall Thickness in mm	
	20	80	20	minor corrosion
	13	94	12	severe corrosion
	22	90	15	medium corrosion

By describing corrosion as in the table 1, we can be more understanding of what can go wrong in the future under diverse conditions of e.g temperature, humidity factors, pipelines wall thickness, instead of one factor of failure. In practice, there can be large features that are used to make prediction which it will depend on the context of the assessment. In this part, although there are a lot of conditions that should be learned by the algorithm to make prediction, that technology still capable to find pattern recognition and make automate indication accurately.

3. DISCUSSION

Basis knowledge about supervised machine learning and its type of predicted output have been elaborated in previous section. Thus, in this part, there will be discussions regarding the appropriateness of such tool in predicting corrosion and its results for supporting decision makers in avoiding pipelines leak.

In view of complexities of the real-world situation, it is thus crucial to predict corrosion under miscellaneous conditions. This is done to produce accurate prediction under situation where the environment alters intensely, the location of pipes that can be in surface and sub-surface and the length of pipes that can be installed in great distances.

By considering the ability of supervised machine learning that can make prediction based on big data quickly and accurately, hence, this tool can be judged suitable to be a tool for predicting corrosion in pipelines. This is because not only single but numerous conditions (level of severity, factors, etc.) can be forecasted by this tool. Thus, information about corrosion can be produced more comprehensively as corrosion can be identified in different contexts. Furthermore, the predicted outputs that are presented in classifications can help risk analysts in prioritizing corrosion for management purposes. The speed of algorithm in generating prediction make this tool can be used for monitoring corrosion. Thereby, any changes can be diagnosed and actions to adjust the changes can be planned and addressed immediately.

However, as many measurement tools, supervised machine learning have drawbacks. This tool does not reflect the aspects of uncertainty thoroughly. The background knowledge that is used to make a prediction using this tool such as data, learning algorithm, and assumptions can likely to associate with the uncertainty.

The data that is utilized by algorithm for learning process can involve with uncertainty because it is created based ones' knowledge. If they have lack of understanding about the phenomena being analyzed, they may provide incorrect examples in the dataset. As a result, the predicted results can be also wrong in representing future actual condition. Also, the data that is gathered from historical performance may not represent the real-world situation. Furthermore, algorithm can contain uncertainty because the detail process of learning and making prediction are not transparent. It seems like a black boxes prediction. The decision makers might be skeptical whether the results are correct despite the model evaluation has examined the accuracy, recall, and precision are good. Moreover, assumptions can be also inherent with uncertainty because basically we, as a human, cannot foresee and visualize a whole world situation. A plenty of important aspects of uncertainty can be neglected once we have lack of knowledge towards the issues being analyzed.

Since each background knowledge that will be used to make a prediction from supervised machine learning can likely to associate with uncertainty, therefore, the predicted results based on this predictive analytics tool should not be trusted completely for being the only decision support. The classification outputs can produce wrong prediction, e.g a specified pipelines section is forecasted minor corrosion but in reality, it may turn out to be severe corrosion.

What is more, restricting decision basis only to the supervised machine learning can ignore the aspects of risk. This is because such tool can only measure uncertainty of corrosion based upon specified event (e.g severe or minor corrosion). The specified consequences are not reflected (e.g the impacts of the occurrence of severe corrosion). That is why risk are not considered properly using this tool. Meanwhile, to support the decision makers, such aspect need to be provided to describe risk comprehensively, thus, they can have an insight which risks that are significant and need to address measures promptly.

Concerning both important aspects such as uncertainty and risk are overlooked by this tool; thus, it can be said that decision basis based on supervised machine learning is not robust to support decision makers in preventing leaking phenomena in pipeline. Some approaches are required to apply for improving decision basis.

4. THE NEW FRAMEWORK TO IMPROVE DECISION-MAKING

A suggested framework is given in this paper work to reflect the aspects of uncertainty and risk. In addition, it can be used to develop decision basis that is established from supervised machine learning. In this part, the framework contains two methods which are a hidden uncertainty analysis and qualitative risk matrices, which they should be carried out progressively due to uncertainty and risk have different aspects that should be covered. For uncertainty, the aspects that should be reflected are related to vagueness in the background knowledge used to predict uncertainty. In the meantime, the aspects of risk that should be indicated is the stage of risk that would be faced when an event should occur. However, although both aspects have its own issues, uncertainty can be useful for risk aspects. Especially, for defining what kinds of preventing actions that should be chosen to manage risks under uncertainty.

Indeed, to achieve that information, a hidden uncertainty analysis and qualitative risk matrices shall be collaborated. Hidden uncertainty analysis will be conducted firstly to identify the overall degree of uncertainty involved in the prediction outcomes as well as the factors that can significantly lead to the deviation from actual situation. That can be done by specifying uncertainty factors and sub-uncertainty factors which then they will be assessed in terms with the degree of uncertainty, sensitivity, and criticality. In this case, if the overall level of uncertainty is appraised to be moderate or significant, thus, predicted outputs should be interpreted in an overestimate way. For instance, by visualizing minor as moderate corrosion and moderate as severe corrosion.

After, a hidden uncertainty analysis has been performed, the next method that should be performed is qualitative risk matrices. It should be noted that in operating this approach, consequences in each corrosion severity should be analyzed. Also, it should be reflected to the personnel, environment, assets and so on as stated by NORSOK Z-013. Once the outputs of consequences analysis have been obtained, then, they should be compared to the classification outputs that produce from supervised machine learning. The outcomes of doing this approach is we can have an insight about the level of risk that might be confronted in the future. By referring to risk level, risk analysts can establish suggestions regarding safety measures that shall be implemented to deal with the risk being faced. Furthermore, employing this method can help the decision makers in deciding assuredly which safety measures that should be addressed immediately.

It should be noted that, interpretation of risk level should be seen based on the overall degree of uncertainty. If a hidden uncertainty analysis result showed the degree of uncertainty is moderate or significant, thus, risk level must be interpreted in an overestimated way. It is thus meant that there will be more risk reducing measures that should be implemented to handle risks.

Actually, there is benefit and drawback of applying more safety measures in an operation. The advantage is we can be more prepared and aware if what have been predicted does not occur in the future or it happens more severely (surprising outcomes). But, the disadvantage is that the companies need to spend more resources on that measures meanwhile they have some limitations too. In reality, such gambling situations are often happened, especially in balancing between safety and cost. That is why, the suggestion that require to overestimate risk must be considered in line to the trade-off aspect.

5. CONCLUSIONS

Overall, performing supervised machine learning to predict corrosion in pipelines can bring advantages and disadvantage. The advantages are that corrosion can be predicted simultaneously with respect to the type, severity, and numerous causes that can lead such issue to occur rather than only single factor. Therefore, information about corrosion can be acquired comprehensively. Furthermore, the ability in generating classification about data makes this tool can support risk analysts in identifying, prioritizing, and monitoring corrosion without taking much effort from human intervention and a long time. Meanwhile, the disadvantage is that this tool does not reflect the important aspects of uncertainty and risk. Ignoring the uncertainty can lead to the occurrences of surprising events. Whereas, overlooking the aspects of risk in decision-making can cause difficulties in selecting safety measures to prevent pipelines leakage. This is because the degree of jeopardy is not taken into consideration properly so that makes it hard to define which measures that should be implemented immediately or postponed. By weighing the benefits and drawback, it can be considered that predicted outputs based on this tool would not be robust to be the only decision support for preventing pipelines leakage.

Thus, the decision basis based on supervised machine learning needs to be improved before delivering to the decision makers. The aspects of uncertainty and risk should be examined to strengthen the decision-making support. In order to reflect both aspects, one can adopt a new framework that consists with two methods: hidden uncertainty analysis and qualitative risk matrices. In this case, the hidden uncertainty analysis method should be performed in prior to examine the overall degree of uncertainty involved in the predicted outputs based on this tool. The second approach that shall be adopted is qualitative risk matrices, where the predicted outputs and consequences analysis outputs are compared to identify the risk level of future event. By integrating both methods, risk analysts can determine the risk reducing measures based upon the degree of uncertainty and risk involved. In this part, if the overall degree of uncertainty involved in the prediction's outputs are moderate or significant, thus, the risk level should be interpreted in an overestimated way. Visualizing risk in such way can lead the companies to invest more in the safety measures for anticipating the occurrence of surprising events and other hazards. On one hand, it can be good for them because any losses, accidents, and other catastrophes can be avoided effectively. On the other hand, it can take so much cost only for preventing pipelines leakage. Meanwhile, there are other incidents and accidents that must be treated as well. This situation can decrease their expected benefits. Therefore, in taking decisions to select prevention's actions, decision makers must reflect to the economic aspects as the companies have limitations in the resources.

REFERENCES

Ahammed, M., & Melchers, R. (1996). Reliability estimation of pressurised pipelines subject to localised corrosion defects. *International Journal of Pressure Vessels and Piping, 69*(3), 267-272.

Azure, M. Introduction to Machine Learning in the Azure cloud. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-machine-learning

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms. Retrieved from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W., & Jeyachandra, B. (2016). *Data driven production forecasting using machine learning.* Paper presented at the SPE Argentina Exploration and Production of Unconventional Resources Symposium.

Choi, J., Goo, B., Kim, J., Kim, Y., & Kim, W. (2003). Development of limit load solutions for corroded gas pipelines. *International Journal of Pressure Vessels and Piping, 80*(2), 121-128.

CrashCourse (2017). [Machine Learning & Artificial Intelligence: Crash Course Computer Science #34].

da Cunha, S. B. (2016). A review of quantitative risk assessment of onshore pipelines. *Journal of Loss Prevention in the Process Industries, 44*, 282-298.

Dey, P. K. (2004). Decision support system for inspection and maintenance: a case study of oil pipelines. *IEEE transactions on engineering management, 51*(1), 47-56.

Dey, P. K. (2006). Integrated project evaluation and selection using multiple-attribute decision-making technique. *International Journal of Production Economics, 103*(1), 90-103.

Dlouhy, J. A. (2013). Pipelines are safer than trains and trucks, report says. Retrieved from https://fuelfix.com/blog/2013/10/17/pipelines-safer-than-trains-and-trucks-report-says/

FluidDataReporting (Producer). (2013). Internal Corrosion Control for Oil and Gas Pipelines. Retrieved from https://www.youtube.com/watch?v=9bb-B357oQA&t=182s

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452.

Guikema, S. D. (2009). Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory. *Reliability Engineering & System Safety, 94*(4), 855-860.

SINTEF, S. (2003). *Handbook for Fire Calculations and Fire Risk Assessment in the Process Industry*. In. Retrieved from https://ia800506.us.archive.org/5/items/SINTEF2003HandbookForFireCalculationsAndFireRiskAssessmentInTheProcessIndustry/SINTEF%20-%202003%20-%20Handbook%20for%20Fire%20Calculations%20and%20Fire%20Risk%20Assessment%20in%20the%20Process%20Industry.pdf

Vtorushina, A. N., Anishchenko, Y. V., & Nikonova, E. (2017). *Risk Assessment of Oil Pipeline Accidents in Special Climatic Conditions.* Paper presented at the IOP Conference Series: Earth and Environmental Science.