Faculty of Engineering and Science

Department Of Electrical and Computer Engineering

Optimal Strategy in Predicting Equipment Sensor Failure Using Genetic Programming and Histogram of Residual

Pauline Wong

0000-0001-6033-4916

This thesis is presented for the Degree of
Master of Philosophy
of
Curtin University

December 2023

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature:

(Pauline Wong)

Date: 24th May 2023

Acknowledgements

I would like to express my sincere thanks and appreciation to my supervisors, Dr Wong Wei Kitt, Dr Amaluddin Yusoff, Assc. Prof. Lenin for their continuous guidance and support throughout my research period. Dr. Wong Wei Kitt, with his dedication and guidance, has inspired me throughout my research and thesis writing.

I would like to thank Sarawak Shell Berhad for my course sponsorship and Curtin University Malaysia for providing all the necessary resources and useful webinars throughout my studies.

In addition, I would like to thank my colleagues from the Engineering and Reliability teams for their inspiring suggestions and discussions during the work of this research.

Finally, I would like to give my biggest thanks to my family for their continuous support and encouragement during the good and bad times.

Abstract

Sensors are crucial in detecting equipment problems in various plant systems. However, detecting sensor abnormality is challenging because the data are normally acquired using IoT approach and stored offline in a dedicated server (data logs). Hence this situation presents both opportunities and challenges when exploring the sensor abnormality detection task. In this thesis, the concept is to apply a multistage compressor sensor fault detection method using data logs. The objectives of this research are to device an approach to detect sensor abnormality and perform this in a "white box" approach. The motivation of this research thesis stems from a need of developing a transparent ("white box" model for detecting sensor abnormality using low frequency data (10 minutes interval). Hence, this only allows for static machine learning model implementation. In the proposed approach, the compressor sensor output is modelled as a function of other sensors using static approach. Subsequently, the model output is used for detecting abnormality by observing the residuals. It has been shown that the histogram of residuals offers rich information to predict abnormality of the targeted sensor. In particular, to explore the concept using genetic programming to generate regression model which offers more "white box" solution to the operators. There are various advantages in this approach. Firstly, the conventional "black box" approach lacks model transparency and, thus, is highly unfavored in critical systems. Secondly, equations are more easily applied in Programmable Logic Controller (PLC) in the case where autonomous flagging is required. Comparison between regression results with Multiple Linear Regression (MLR) and

Neural Network Regression (ANN) are discussed. Results show that the best generated models are comparable with the latter but with more crisp "white box" mathematical equations utilizing lesser feature inputs (four features only). In terms of performance, the best model generated yielded a $R^2 = 0.9044$ which is comparable to ANN and MLR. Comparative results indicate that the performance of MLR, GP and ANN differs slightly thereby showing that the models is linear as MLR is known to perform optimally in such situations. However, the main advantage in the GP generated model is that it only manage to generate an solution model consisting only of 4 features with relatively well performance as indicated from the R^2 on independent test data. The research concludes that the 1)proposed approach to identify the abnormally was feasible albeit with simulated sensor faulty data. The second finding suggest that a mathematical model to predict the sensor output (for calculating the residual histogram) was feasible as the $R^2 = 0.9044$ was acquired on the independent test dataset.

Publications Arising from the Thesis

Pauline Wong, W.K. Wong, Filbert H. Juwono, Lenin Gopal, Mohd Amaluddin Yusoff, A minimalist approach for detecting sensor abnormality in oil and gas platforms, Petroleum Research, Volume 7, Issue 2, 2022,
ISSN 2096-2495, https://doi.org/10.1016/j.ptlrs.2021.09.007.
(https://www.sciencedirect.com/science/article/pii/S2096249521000740) (Elsevier) (published)

Pauline Wong, W.K. Wong, Filbert H. Juwono, Andy Lease Basil,Lenin Gopal, Sensor Abnormality Detection in Multistage Compressor Units: A "White Box" Approach Using Tree-Based Genetic Programming, Complex System Modeling and Simulation (E-Prime, Elsevier) , 2022, (Accepted for publication, Pending minor revision)

Contents

List of Figures

List of Tables

Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
ANFIS	**A**daptive **N**euro-Fuzzy **I**nference **S**ystem
CGP	**C**artesian **G**enetic **P**rogramming
HH	**H**igh **H**igh
IoT	**I**nternet of **T**hings
ML	**M**achine **L**earning
MLR	**M**ultiple **L**inear **R**egression
MSE	**M**ean **S**quared **E**rror
NPT	**N**ational **P**ipe **T**hread
RMSE	**R**oot **M**ean **S**quared **E**rror
ND	**N**ormal **D**istribution
PCA	**P**rincipal **C**omponent **A**nalysis
PI	**P**lant **I**nformation
PLC	**P**rogrammable **L**ogic **C**ontroller
PSO	**P**article **S**warm **O**ptimization
RMSE	**R**oot **M**ean **S**quare **E**rror
RPM	**R**evolution **P**er **M**inute
SCADA	**S**upervisory **C**ontrol **D**ata **A**cquisition
UD	**U**niform **D**istribution

List of Symbols

y_n	**T**arget value
x_n	**V**ector of **F**eatures
β	**W**eights **F**or **L**inear **R**egression
α	**L**earning **R**ate for ANN
$\hat{\alpha}$	**G**enetic Programming **I**nput
$\hat{\beta}$	**G**enetic Programming **I**nput
ϵ	**B**ias **T**erm
i	**I**ndex
j	**I**ndex
R^2	Statistical Measure (Goodness of Fit)
f	Noise Generation Parameter

Chapter 1

Introduction

1.1 A Relevant Industrial Dilemma

Compressors are important in the oil and gas industry and have been in constant research for detecting failure as exemplified in research works demonstrated in R. Kurz and K. Brun (2012) and Priyanka et al. (2020). Primarily, these machines are used in the export pipeline for transporting gas to the shore. Compressors are often not spared due to economic reasons, making them vital for reliable operation and continuous production. However, owing to issues with the operation in matured gas fields, the compressors are constantly tripped, causing the operations to come to a halt. Therefore, various sensors can be mounted to monitor the operation of the compressors.

Sensors are commonly used in industrial automation to detect abnormalities and control system characteristics (Thangavel et al., 2021). The sensors measure the compressor characteristics, such as shaft vibration, rotor position, gas temperatures, pressures to determine thermodynamic performance, etc. It is worth mentioning that the placement of the sensors is crucial to get useful information related to the dynamic features of the machine (Goyal et al., 2019). Using the correlation of sensor data, the operator can have an early detection system to avoid early faults. For these reasons, in-depth knowledge of sensor data is

needed. Due to the harsh environmental conditions, methods for continuously tracking sensor health are needed to ensure data validity. However, apart from adding redundancy to the sensors, no other solutions have been proposed. It is difficult for plant operators to maintain the equipment in this condition.

An ideal scenario is to have a model that alerts the user if a sensor drift or fault is observed. A high level concept of this research is shown in Figure 3.1 and the research boundaries will be further discussed. Let's assume that there are six sensors (V1-V6) which correlate and respond to the various states that the compressor is currently operating at. Data are collected and multiple regression models are trained such that the sensor outputs can be modelled. In this scenario, the six regression models, which correspond to V1-V6 sensors, are predicted and compared to observe the output of the sensors. Residuals are the differences between the sensor model's prediction and the observed data. In ideal sensor conditions, the residual distribution would be a normal curve with mean 0 for V1-V6. The residuals' standard deviation and change in the mean can discover potential abnormalities of the compressors. For the sake of brevity, this thesis investigation focuses on the shaft's RPM sensor, it is widely considered as the primary sensor for monitoring the compressor's working condition. In practice, the definition can be generalized to model all other sensors and can be extended to other equipment like pumps and motors. There is essentially no restriction in the use of ML models for predicting faults.

In stating this, algorithm in application would only be applied based on suitability such as unbalance data set, limited data set or the presence of only positive or negative class in faults. An example of this case can be seen in Priyanka et al. (2020) and Thangavel et al. (2021) which applied unsupervised clustering algorithms for detection of failure in gas transportation pipeline. The importance of sensor in detecting machine faults cannot be overstated. In general, data analysis or fault detection can be categorized into several categories. The first category uses the dynamic approach and may require high-frequency data. This method

may be costly due to high-frequency data collection and sampling, as shown in Yazar et al. (2017). The second approach makes use of feature generation of data and applies the static nature of the data, as presented in Jiang et al. (2019). In any of the both stated case, a fault of intermittence may have adverse effects causing false alarms of low detection of faults.

1.2 Research Motivations

This thesis proposes a system that is able to detect abnormalities in data logs to predict equipment sensor failure. In particular, the use of regression model with a more transparent and mathematical formulation which is well-known as "white box" approach, as mentioned previously. This allows operators to periodically check on the health of sensors without relying on the high frequency data and computation. Deviation from regression models indicated the health and integrity of a particular sensor. Note that higher frequency data is more computationally expensive and is not feasible for all systems, especially given the cost of implementation.

The proposed method relies on "offline" data logs and a low update rate (10 minute interval), making data streaming less computationally expensive. This data logging system is a typical case for most operators, given the high cost of operations. In this thesis, a GP tree approach (Cavagliá et al., 2020; Kai, 2017) is used to mathematically model the compressor RPM sensor output as a function from other sensors. Note that other than the RPM sensor, the compressor is equipped with 46 sensors as shown in the Appendix A.

As previously stated, MLR and ANN models are excellent working choices, but they lack model transparency or "white box" sense. These models are used as benchmark models for comparison purposes. Summary of the contributions is as follows:

1. A method for predicting sensor failure using a mathematical model. The

model is developed using a tree-based GP defined by the program length, and it is then used to predict the compressor's RPM sensor abnormality.

2. The proposed method is compared with the MLR and ANN models with regards to model fitness metrics, i.e., Mean Squared Error (MSE) and R^2.

3. The residuals and augmented actual data is used to predict various types of faults in the sensor. In specific, actual data are augmented using the approach proposed by (Tsai et al., 2019).

1.3 Research Contributions

The following research question arises: "how sensor failures can be identified, the primary instrument that detects machine problems?" With such a vast focus on predictive maintenance, some of the concepts can be borrowed from the field of predicting machine fault for sensor fault detection. The configuration of sensors is typical for applications on an offshore gas compression facility, and thus, the investigation sufficiently generalize to cases of similar nature. In fact, most of the aforementioned concepts applied some form of machine learning (ML) that map sensor output to fault flags. Hence the ML models may be similarly deployed to map sensor reading to a predicted output in which deviation (residuals) can be observed and analyzed to trigger sensor abnormality. In another words, methods to model machine failure can be used to similarly flag abnormally as well with a slight change in thinking. The purpose of this research is twofold. First, this research investigates the application of multiple regression and curve fitting neural networks models to model machine sensor outputs as a function of the other sensor readings. As such, the predicted model can be used and compared with the observed output. Two shallow models were selected for a compressed model to include every possible state of the compressor sensor (i.e., data approach). In justifying the application of the aforementioned model of ML, minimal complexity models are preferred due to "Occam's Razor" principle and to avoid over-fitting

for more reliable sensor output modelling and prediction. For this case, the obvious choices are shallow learners such as the selected curve fitting neural network and multiple linear regression. In essence, a neural network may be considered as a 'black box' model, whereas a multiple regression is an equation-based model in which both have their respective pros and cons. Neural networks operate effectively by mapping inputs to outputs in a nonlinear fashion due to their connection structure. On the other hand, multiple regression models can only map linearly, but they may provide a clearer perspective on which sensors (features) are important. The contributions of this thesis can be summarized as follow:

1) A new approach in modelling sensor output using sensor output residual method

2) Mathematical model for generating residual in (1) by modelling the expected output using as a function of other sensor output using Genetic Programming (GP)

1.4 Aim and Objectives

Based on the research questions discussed above, sensors are the primary instrument that detects machine problem, hence accurate identification of sensor failure is important. The objectives of this research are as follows:

1) Implement an approach to detect the faults in sensors by only using low frequency data logs. Data collected at an interval of 10 minutes.

2) Implement and modelling using "White box" model approach. The definition of "white Box" model refers to transparent mathematical model. The details can be found in the subsequent chapters of the thesis

An approach using low frequency data logs is needed because most of the industrial system has low update rate. This is definitely an important advantage that meets current system availability and practicality, for the system does not need to "overdrive" in data collection. This thesis also aims to implement and modelling using "white box" approach as this criteria is important especially in critical industrial system. Three machine learning models were considered, and as a result, neural networks can offer a more complex and efficient prediction while multi regression models can provide a mathematical rationale for the sensor correlation. The third model using GP is proposed in view that it offers a "white box" approach. Limitations and operational challenges win the current practices will be discussed in the next chapter, so the approaches in this thesis can be further supported by the detailed industrial operation and scientific literature reviews.

1.5 Overview of The Thesis

The remaining section of this thesis is organized as follow. Chapter 2 discusses the state- of-the-art of predictive maintenance/sensor fault detection research. In Chapter 3, the methodology leading up to the proposed solutions is provided. This includes the data collection and the compressor unit which is in discussion. In Chapter 4, the results of the experiments are presented and discussed. Finally, the conclusion and future work are presented in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction to Chapter

In order to provide the reader with a greater understanding of the idea proposed, this section addresses the general concept, sensor analysis, and the use of ML to model sensor output. Despite the attempt to investigate the sensor abnormality, it is inevitable that the relationship between sensor and predictive maintenance are intertwined, and many of the concepts of detecting component fault and sensor fault share the same principles. Therefore some investigation of the research work related to predictive maintenance is discussed as a possible reference to detect sensor faults. The term predictive maintenance is a concept that has been studied by automation researchers and academics, and it is part of early fault detection of machines. Sensors are used as part of the model building for predicting states of the machine. This corresponds to the implementation of Industry 4.0. Internet of Things (IoT) and artificial intelligence have been utilized in the deployment of Industry 4.0 in oil and gas platforms to identify and fix problems, such as pipeline faults.

2.2 Predictive Maintenance

Predictive maintenance is a technique used to predict when a machine or equipment is likely to fail, so that maintenance can be scheduled before the failure occurs. The goal of predictive maintenance is to avoid unexpected downtime and reduce maintenance costs, while improving reliability and safety.

Predictive maintenance is based on the use of data analysis techniques, such as machine learning and statistical analysis, to identify patterns and trends in machine performance data. The data used for predictive maintenance can come from various sources, such as sensors, logs, and historical records.

The predictive maintenance process typically involves the following steps:

1) Data collection: Collecting data on machine performance, including sensor data, logs, and historical records.

2) Data pre-processing: Cleaning and preparing the data for analysis, such as filtering out noise and removing outliers.

3) Data analysis: Analyzing the data to identify patterns and trends, and building a predictive model to predict future machine performance.

4) Maintenance planning: Using the predictive model to schedule maintenance activities before a failure occurs, based on the predicted time to failure or remaining useful life.

5) Maintenance execution: Performing maintenance activities as scheduled, based on the predictive model.

Predictive maintenance has many benefits, including increased equipment reliability, reduced maintenance costs, and improved safety. It is widely used in various industries, such as manufacturing, transportation, and energy, to improve the performance and longevity of equipment and reduce downtime.

On application level, some of these research conducted is demonstrated in (Priyanka et al., 2021b; Bhaskaran et al., 2021, 2020; Priyanka et al., 2020). In particular, predictive maintenance has been used as a means of preventing a severe failure that often results in a loss of productivity and, ultimately, revenue.

In most cases, predictive maintenance uses sensor reading and data processing to forecast machine malfunction or under-performance. Some research work about predictive maintenance has been available in the literature. Authors in Hanachi et al. (2017) used an adaptive neuro-fuzzy inference system (ANFIS) model for predicting degradation of the compressor section in gas turbine engines (GTE). In particular, the work focused on a particular problem called fouling. The fouling state was characterized by the relative change of the ratio of the compressor mass flow and efficiency against ideal conditions. Using the ANFIS model, the overall forecast findings showed an increase of around 50% and 28% for the 2-hour and 120-hour forecast cycles, respectively, compared to the logarithmic regression model.

In Jegadeeshwaran & Sugumaran (2015), the authors used various classification models to classify 10 hydraulic brake faults. With the inclusion of feature reduction methods, the authors reported 96% recognition rate. There are different fine-tuning models that need to be taken into account and, thus, one ML model does not indicate an absolute match for a particular application. This involves a certain 'try-all' approach, thereby validating further research studies on the application of different ML models for the good of the manufacturing community.

In Jose (2018), the authors proposed and advocated heterogeneous computing involving edge devices, massive wireless networks, ML, cloud computing, and advanced statistical approach among others for predictive maintenance. This approach would go beyond traditional methods, requiring significant investment in infrastructure. In the case of the modern automation setup and the declining price of the computing equipment, this is a viable solution considering the possible savings in terms of production shutdowns.

A similar trend was observed in Cachada et al. (2018), where the authors introduced the use of various Industrial 4.0 concepts, such as Internet-of-Things (IoT) and cloud computing, in the system architecture for the predictive main-

tenance tasks. It should be noted that the authors in Cachada et al. (2018) focused on a specific type of maintenance, called condition-based maintenance, where machine performance was monitored, as opposed to event-based maintenance (Naskos et al., 2020), which focuses on preventing failure. Both use the same sensor technology, but the focus is on different aspects.

Researchers in Steurtewagen & Van den Poel (2019) applied several ML models for root cause analysis. The purpose of this case study was to investigate the root cause of high vibration-scenarios that triggered the failure of the compressor units in the oil refinery. In Rosli et al. (2018), the authors used principal component analysis (PCA) feature reduction and multiple linear regression model for investigating the main cause of air booster compressor failure. Using R^2 and root mean square error (RMSE) values, it was shown that multiple linear regression was more suitable for determining the influential parameters causing the failure of the compressor. In Sakthivel et al. (2014), the authors applied linear and nonlinear feature reduction techniques for classifying centrifugal pump faults. From Rosli et al. (2018) and Sakthivel et al. (2014), the authors deduced that feature reduction may be highly dependent on the type of rotating machine and the data points used. This does not, however, result in absoluteness in the improvement of the ML model.

In Rosli et al. (2019), the authors applied a stochastic optimization technique for detecting air booster compressor failure. In particular, the authors applied particle swarm optimization (PSO) for tuning neural network structures. The results showed that there was slight improvement when compared to the gradient based approach. Noted that some researchers have emphasized their prediction model based on static analysis of sensor reading. However, some of them have used time series data for analysis. For example, a long-term memory (LSTM) model, a type of recurrent neural network, was used in J. S. Rahhal and D. Abualnadi (2020) to predict compressor failure. The disadvantage of LSTM is that time-based models are more dynamic and would require high frequency data.

2.3 Sensor Fault Detection

Due to the importance of sensors with respect to predictive maintenance, it is therefore imperative to ensure that sensors are in good working conditions such that it does not trigger false notifications of equipment failures. The challenge being faced is that not much accepted industrial practices apart from applying redundant sensors for monitoring. This is by no mean an economical practice albeit acceptable due to the importance it bears. An alternative and efficient method would further create opportunities to ensure more efficiency in sensor monitoring. In Byun et al. (2019), authors applied a logic and condition based approach in detecting sensor faults in particular by sensor triangulation method. The authors in Wang et al. (2017) proposed three methods to detect sensor fault by observing transient state errors. The authors in X. Jia and Q. Cheng and Y. Hou (2018) applied state estimations using fuzzy systems and compared with the observed output. Residuals, which is the difference between sensor measurements and the estimated outputs of the system based on an observer, were applied to decide on the sensor faults.

An extensive review of sensor fault was examined in Li et al. (2020). In the article, the authors cited three main categories of sensor detection methods, namely model-based, data-driven, and knowledge-based methods. The model-based method requires some forms of comparison between model and observed data. This would require extensive modelling of a plant. Meanwhile, knowledge-based requires prior knowledge of sensor faulty condition that may be indicated. On the other hand, a data-driven approach would require some data mining approach for determining if the sensor is at fault state. However, as ML algorithms evolve, more hybrid techniques are applied, thereby blurring the lines between the methods. As an example, this suggested approach is a combination of the first and third categories. Another elaborate review on sensor abnormality detection was presented in Gaddam et al. (2020). Authors in the paper have attempted to classify strategy for detection of sensors into three main categories namely

the network level strategy, heterogeneous strategy and homogeneous strategy. In particular, worth to highlight the authors' explanation of the heterogeneous strategy which may be related to the proposed approach in this paper. The author describes the strategy as applying other sensors to predict the sensor output of interest. The *principalis* idea here is to use sensor values to gauge sensor of interest "correctness", this essentially can be implemented in various ways. In Kapitanova et al. (2012), authors have applied the state of activity in a smart home environment to predict what the sensors should be in ideal normal sensor state. This essentially agrees with the principle outlined by Gaddam et al. (2020).

By exploring the previous works that relate directly to sensor abnormality detection, there is still room to explore, especially regarding the sensors representing complex and dynamic states of the system. Noted that the research area in particular abnormality detection is highly context-dependent, particularly in the system which is being applied on. It is imperative to explore these concepts especially for the case of compressors which play such a critical role in the oil and gas platforms. In this case, data acquired is from a lower frequency interval making dynamic analysis undesirable. Due to the high correlation between the sensors, it might be suitable to apply some forms of sensor correlation concept, known as "heterogenous strategy" (Gaddam et al., 2020). By borrowing concepts from machine failure prediction, ML can be applied to correlate sensor output as a function of other sensor outputs. Since the sensors are correlated, ML would be able to model various states of operation enabling the model output to be compared with the observed output over a prescribed span of time. This is further discussed in the subsequent section.

Sensors have been widely adopted in system state and fault monitoring. The bulk of research applies multiple sensor correlation to detect plant abnormality (Sujeong & Duck-Young, 2019). Despite the importance of sensor integrity, not many strategies have been proposed to check the abnormality of the sensors. Oil and gas industries commonly apply redundancy of sensors as a fail-safe strategy

(Johansen et al., 2021). In-depth discussion on this topic could be challenging due to few reasons. First, sensor applications are numerous, and discussions in specific applications can be considered niche discussions within an industrial/scientific community. Second, sensor data is typically highly classified due to the stakeholder's critical operations, making it difficult to hold discussions between sensor domain experts, operators, and machine learning specialists (the three invested parties within this scope of research).

Despite these limitations, some important investigations on the fundamentals of this topic of discussion which emphasizes from the industry point of view need to be further investigated. On the other hand, from an economic point of view, global uncertainty has caused operators to employ more automation strategies when it comes to plan system maintenance. In particular, sensors enable operators to gain a better understanding of the state of operations, from the platform to the refinery plant. Nevertheless, the industry is skeptical of using machine learning algorithms for operation due to a variety of reasons, including the "black box" nature, algorithmic biases, and uncertainty in machine decision making, which are impeding growth in this area (Cavagliá et al., 2020). It is worth mentioning that the term "black box" refers to algorithmic models that are not readable or transparent in nature, such as Deep Learning (DL) and Artificial Neural Networks (ANN). Regardless of preference, economic pressure has caused operators to reconsider the balancing act between increasing productivity, safety concerns, and economic benefits. A preliminary research on the use of the "black box" approach for sensor abnormality detection is presented in **?**.

In the oil and gas industry, some research works focus on the algorithmic (machine learning or model-based) approaches for pipeline monitoring as demonstrated in Priyanka et al. (2020),Priyanka & Thangavel (2020), Priyanka et al. (2021b), Priyanka et al. (2021a). This aspect is considered as a critical area of application where a mathematical model-based approach is preferred. Furthermore, in Rosli et al. (2019) and Sakthivel et al. (2014), the authors investigated pre-

dictive maintenance for air booster compressor motor failure with the intention of sensor-based monitoring of the equipment conditions. Note that, predictive maintenance generally can be categorized as event-triggered (Bousdekis et al., 2017) and condition-based maintenance (Li et al., 2017). In the latter, data are stored and monitored consecutively. As a result, it may incur more computations as compared to the event-triggered maintenance. Another example of the event-triggered predictive maintenance is shown in Naskos et al. (2020) where a factory setting was applied as a case study.

In Cachada et al. (2018), the authors investigated how early investigation and identification of faults could lead to lesser maintenance time which subsequently lead to lesser economical impact due to fault related uncertainty. In Hanachi et al. (2017), the authors emphasized the significance of various gas turbine degradation and it prompted an investigation into the use of the predictive maintenance concept for mitigation purposes. In Byun et al. (2019), sensor fault detection and signal restoration in intelligent vehicles were investigated. The paper shows another example where critical integrity of sensor may cause safety and health concerns. Specifically, condition-based approaches were implemented to detect sensor faults. The paper demonstrated a simple yet effective method in diagnosis of sensor but the implementation was based on specific application and other complex systems, such as a plant, may require a more complex modeling to encapsulate the more complex states of the system.

A comprehensive survey of methods in detecting sensor abnormality is presented in Gaddam et al. (2020). In the report, the authors highlighted that generally they exist three strategies in classifying the faults of sensors which are network level strategy, heterogeneous strategy, and homogeneous strategy. A homogeneous strategy in this context refers to correlating the output of the same type of sensors to detect sensor abnormalities. A heterogeneous technique, on the other hand, employs a variety of sensor types to detect failure. The authors also highlighted various machine learning models that could help feature researchers

develop a strategy for detecting faults. In Li et al. (2020), the authors provided an in-depth perspective on recent advances pertaining to sensor abnormality detection. Similar to Gaddam et al. (2020), the authors highlighted various types of sensor faults which can be considered for similar types of sensor abnormality research work such as sensor drift, abrupt failure, random faults, short circuit, and open circuit faults. The authors also highlighted similar mitigation strategies in identifying these faults namely model-based, knowledge-based, and data-driven.

There is a growing community of researchers that applies model deviation with actual measurement for sensor faults diagnosis. Cha et al. (2017) applied various regression models such as support vector regression models and auto associative neural network for predicting sensor drift. The research work focus on a particular type of sensor faults which are sensor drift. The approach uses machine learning to model sensor output and compared with measured reading. A method for intelligent sensor validation in real time situations is presented in Ibarguengoytia et al. (2001). For the purpose of locating a malfunction in a collection of sensors, the program makes use of a Bayesian network. The relationships and inter dependencies that exist between all of the sensors are represented by this Bayesian network. A second Bayesian network is used to single out the malfunctioning sensor from among all of the other apparent malfunctioning sensors. An online sensor defect detection method that uses Auto-Associative Kernel Regression (AAKR) and the Generalized Likelihood Ratio Test is presented in Sairam & Mandal (2016). The AAKR technique is used to estimate the data, and the GLRT method is used as a measure to identify the malfunctioning sensor on the residual space, which is defined as the amount by which the approximated data deviates from the original data. In Alves et al. (2021), authors implemented similar model based approach for validation of sensors in oil and gas platform scenario, in particular an injection pump. The authors provided strong mathematical justification on the proposed approach by citing that even though metrics such as summed squared error is reached, one or more of the individual performance metrics, including: i) accu-

racy; ii) robustness; iii) spillover and iv) filtering of the neural network, may not be satisfactory while validating sensor measurements.

In Galotto et al. (2015), authors presented a 10 years experience of data driven models for sensor validation applied for petroleum and natural gas industry. Auto-associative kernel regression has been used as the main modeling method. The models achieved were embedded in a software called Sentinell, which is used for sensors diagnosis. The software is being used in a natural gas compression station, and it has been evaluated in other industries such as: refineries, offshore petroleum platforms, and thermoelectric power plants.

Another method based on modelling was presented in Galotto et al. (2007). However, apart from detection of fault, this approach also considers compensation. The goal of this study is sensor fault tolerance as well as sensor fault compensation. These two goals are intertwined. In a typical method to fault tolerance, the defect would first be identified, and then the sensor would be removed from the system. It's possible that the faulty sensor has an off-set or scaling mistake, but that error may be adjusted for such that it can still be utilized. This is accomplished with the use of a mathematical solution based on kernel regression that is shown in this study. This approach is able to adjust for measurement error, resulting in estimates that are more accurate and dependable. In the following, discussion and then application of this method can be expanded to motor drives. The findings of both simulations and experiments are given and discussed here.

2.4 Common Machine Learning

A machine learning model is a mathematical algorithm that is trained on a dataset to make predictions or decisions without being explicitly programmed to do so. It is a key component of machine learning and artificial intelligence applications, as it enables computers to learn and improve from experience.

There are different types of machine learning models, such as regression mod-

els, classification models, clustering models, and neural networks. Each type of model is suited to different types of tasks and datasets. The most relevant machine learning models for this project is to predict a continuous value. This is also known as regression type machine learning models.

The process of building a machine learning model involves selecting an appropriate algorithm, preparing and cleaning the data, training the model, and evaluating its performance. Once a model is trained, it can be used to make predictions on new data, and its performance can be measured and improved through feedback and further training. In this section we discuss some relevant regression based machine learning models.

2.4.1 Multi Linear Regression

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and two or more independent variables. It is an extension of simple linear regression, which only considers one independent variable. In multiple linear regression, the relationship between the dependent variable and the independent variables is represented by a linear equation of the form:

$$y = 0 + 1x1 + 2x2 + ... + nxn + \tag{2.1}$$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, 0 is the intercept, 1, 2, ..., n are the coefficients, and is the error term.

The goal of multiple linear regression is to estimate the values of the coefficients that minimize the sum of squared errors between the predicted values of the dependent variable and the actual values. This is typically done using a method called ordinary least squares (OLS) regression. Multiple linear regression can be used for various applications, such as predicting sales based on advertising spending, analyzing the impact of different factors on customer satisfaction, or determining the factors that influence the performance of a machine learning algorithm.

2.4.2 Artificial Neural Network

A regression neural network is a type of artificial neural network used for predicting a continuous output variable, such as a numerical value or a real number. It is a machine learning model that is based on the principles of feedforward neural networks, and is particularly useful for modeling complex nonlinear relationships between input and output variables. In a regression neural network, the input layer consists of one or more input variables, which are connected to one or more hidden layers of neurons via weighted connections. The hidden layers use nonlinear activation functions, such as the sigmoid or the rectified linear unit (ReLU), to transform the input signals and produce a set of output signals. Finally, the output layer produces a single continuous output value based on the activations of the previous layers. The training of a regression neural network typically involves minimizing a cost function that measures the difference between the predicted output values and the actual output values. This is typically done using an optimization algorithm, such as stochastic gradient descent (SGD), to adjust the weights of the connections between the neurons. Regression neural networks have many applications, including in finance, healthcare, and engineering. For example, they can be used to predict stock prices, model the behavior of patients with a specific disease, or forecast energy consumption in a building.

2.4.3 Regression Support Vector Machine

A support vector machine (SVM) is a machine learning algorithm that can be used for both classification and regression tasks. In regression tasks, SVM is known as regression support vector machine (R-SVM) or support vector regression (SVR). The goal of R-SVM is to find a function that best fits the training data while minimizing the margin violations. The margin violations represent the difference between the actual value and the predicted value, which the algorithm tries to minimize. The basic idea behind R-SVM is to transform the input data into a higher-dimensional space using a kernel function, and then find a hyperplane that

best fits the transformed data. The hyperplane is chosen in such a way that it maximizes the margin between the hyperplane and the data points closest to it, which are called support vectors. The choice of kernel function is important in R-SVM, as it determines the transformation of the input data. Commonly used kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. R-SVM is a powerful technique for regression tasks, especially when dealing with nonlinear relationships between the input and output variables. It has many applications, such as in finance, economics, and engineering, for predicting stock prices, estimating demand for a product, or modeling the behavior of a system. Let$(\overrightarrow{x}, y)$ be the feature pair representing the features and $\overrightarrow{x} = [x_1 x_2 x_3]$ representing the features extracted from the $ph(x_1), temp(x_2), sc(x_3)$ respectively.

subsequently, the polynomials of higher order (1,2,3) were derived. we update the $x' = [x_1^1 x_2^1 x_3^1 x_1^2 .....x_1^3 x_2^3 x_3^3]$

projecting x' to a higher dimension, the RBF function $\phi(x')$ was deployed in as $\phi(x') = \exp(-\gamma(\|x' - c\|))$. In order to achieve higher non-linear mapping capabilities, several changes were made. firstly, the polynomial augmented inputs were weighted. Secondly, γ values that controls the Rbf projection as also considered for optimization. Thirdly, weights are added to augment x' and c such $\exp(-\gamma(\|x' - c\|))$ can achieve more non linear mapping. These 3 strategies were deployed such that better non linear mapping can be achieved to correlate the original features x to y.

2.5 Limitation of Non-Interpretable Machine Learning Models

Thus far in the discussion, it is noted that many researchers have presented various approaches that are mostly specific to a particular type of sensors. Despite limited works in generalized sensor approach, the principles implemented may provide a general approach. A validation reading can be acquired by using ma-

chine learning models such as neural network and etc. As stated in previous discussion, most researchers have preference for such models. These models lack transparency despite cited to be highly efficient in various domains of implementation. Rudin (2019) discuses the limitation and dangers of using non interpretable machine learning models with case study from medical and other industrial approach. Based on the highlighted issues, the term "black box" refers to simplified representations of mathematical models that are so difficult for human to comprehend that they are referred to as "black box models." A lack of interpretability in predictive models may erode confidence in such models, which is particularly problematic especially in the field of health care, where the many choices are all literally a matter of life and death. In Loyola-González (2019), the comparison of "white box" vs "black box" was further discussed with certain industrial practitioners having preference for fully explainable models while justifying certain cases where black box models may be advantages. In the case of the targeted application (compressor sensor fault detection), there is an obvious advantage in using white box modelling given the stakes involved and the industry preference to understand ability of the model.

2.6 Summary of Literature Review and Research Gap

Based on the literature reviews discussed in the chapter, research gaps were identified in this area. This thesis proposes a system that is able to detect abnormalities using data logs. Regression model is preferred due to the ability to produce more transparent and mathematical formulation, also well-known as the "white box" approach. This feature is important to allow operators to periodically check on the health of sensors without necessitating high frequency data and computation. Deviation from regression models will indicate the health and integrity of a particular sensor. Note that higher frequency update data is more computationally

expensive and will not be feasible for all systems, especially given the cost of implementation. The proposed method relies on "offline" data logs and a low update rate, making data streaming less computationally expensive. This data logging system is a typical case for most operators, given the high cost of operations. In this thesis, a GP approach (Cavagli´a et al., 2020; Kai, 2017) is used to mathematically model the compressor RPM sensor output as a function from other sensors. Note that other than the RPM sensor, the compressor is equipped with 46 sensors as shown in the Appendix A. As previously stated, MLR and ANN models are excellent working choices, but they lack model transparency or "white box" sense. These models are used as benchmark models for comparison purposes. Summary of the contributions is as follows:

1. A method for predicting sensor failure using a mathematical model. The model is developed using a tree-based GP defined by the program length, and it is then used to predict the compressor's RPM sensor abnormality.

2. The proposed method is compared with the MLR and ANN models with regards to model fitness metrics, i.e., Mean Squared Error (MSE) and R^2.

3. The residuals and augmented actual data is used to predict various types of faults in the sensor. In specific, actual data are augmented using the approach proposed by (Tsai et al., 2019).

Chapter 3

Methodology

3.1 Concept

This chapter explains the methodology in carrying out this research work. First section describes data acquisition process involved covering the multistage compressor system and the sensor system. Next section starts with a high level process flow overview, then followed by discussion of the coverage of the research data which includes mechanical, auxiliary, and overall process systems of the compressor. Each of these categories including machine functions are discussed briefly. Strategies from data collection to the pre-processing steps are discussed, then followed by the modeling approaches for MLR, ANN and GP. Phase 2 discusses the identification of abnormality. The overall concept of the proposed sensor fault detection is illustrated in Figure. 3.1. There are several assumptions in implementing the concept proposed. Firstly in Phase 1, data training is assumed to be pristine as no faulty sensor was reported within the time frame of data collection. It is noteworthy that in this research thesis, the proposed Genetic programming approach was selected due to the "white box" nature and the MLR and ANN are selected as comparative models as it is a common bench-marking exercise to do so. The second assumption pertains to the implementation in Phase 2. The implementation of Phase 2 using residual histogram to evaluate the proposed sensor

abnormality. The error is based on simulated error reading using the approaches proposed in Tsai et al. (2019).

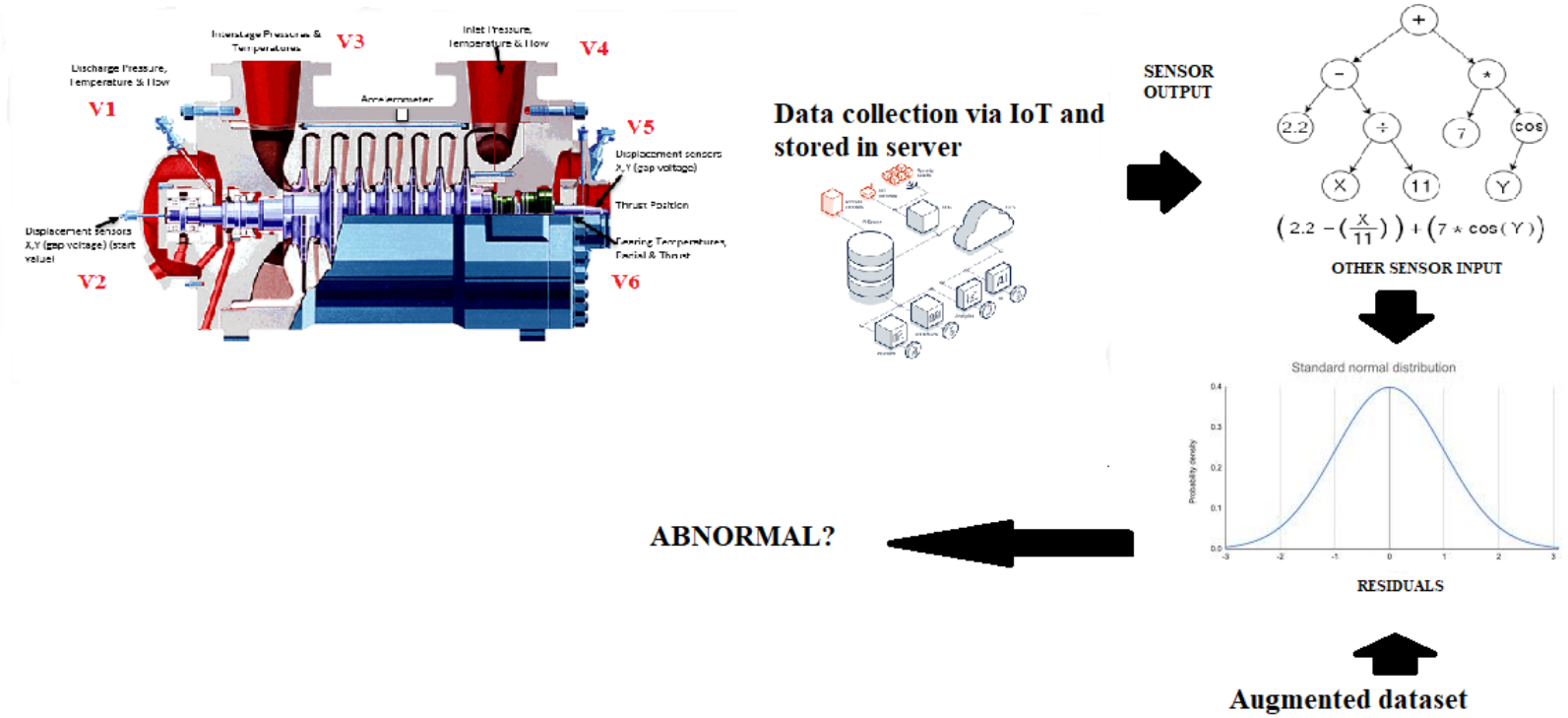

Figure 3.1: Overall concept of the proposed sensor fault detection (Wong et al. (2022)

3.1.1 Acquisition of Data

The data were collected using the data logging system that records information at ten-minute interval for more than 2 years. According to the literature, various research works in this field also implement high frequency data analytic to anticipate machine failure. However, such implementation is problematic due to the enormous computational cost. Furthermore, rather than using dynamic oscillations, this technique detects abnormalities using a static approach. As a result, lower frequency data is used in order to forecast such failures in this study.

These data were sent to and stored in a local database. The data were extracted from offshore equipment using OSIsoft PI system. OSIsoft is a software development and support company that specializes in software that captures, processes, analyses, and stores any kind of real-time data. In essence, the PI System is a set of software tools that are used for data collecting, historical preservation, discovering, analyzing, distributing, and displaying information. Enterprise infrastructure for the administration of real-time data and events is what it is touted as being. The terms PI System and PI Server are often used interchangeably, however they are not the same thing. The PI System refers to all OSIsoft software products, while the PI Server is the primary product of the PI System. The PI System is comprised of the PI Server and all OSIsoft software products.

3.1.2 Multistage Compressor System and Sensor System

Sensors are installed in every part of the compressor as it is deemed as critical machinery in process safety and production reliability. The explanation of the subsequent subsections are imperative to enable an in-depth understanding on the features as shown in the appendices. The machine health conditions can be monitored and evaluated using key parameters such as shaft and bearing dynamic vibrations, bearing metal temperatures etc. Machine thermodynamic performance can be monitored using process variables and theoretical modeling to trend machine degradation or efficiency. All these maintenance monitoring

parameters relies on sensor data to enable proactive intervention by the operator prior to failure due to the mechanical integrity or reliability of the units. Figure 3.2 shows placement of sensors at various critical sections of a compressor to acquire data for the remote monitoring and diagnostic system (RMDS). Figure 3.3 is picture of the compressor unit in offshore platform. The data /reading as shown in GUI are updated on 10 minutes interval. All 46 data tags extracted for this research is tabulated in Appendix A, each category and the intended functions is discussed in subsequent sections.

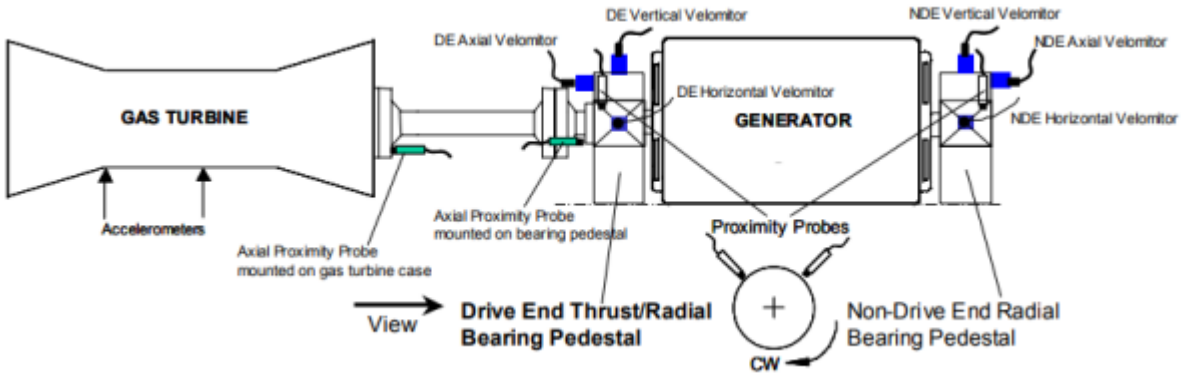

Figure 3.2: Sensors installation for the remote monitoring and diagnostic system of a compressor

Figure 3.3: Compressor in offshore compression platform

Quality data acquired from a centrifugal compressor is important factor that will determine a model performance, as small inaccuracies (in certain areas) can make a large difference between the measured versus actual conditions. The type of field data acquired are normally pressure, temperature, flow rate, vibration etc. Electronic pressure transmitters are utilized for control and safeguarding applications, they are used to measure and communicate a static or differential pressure value to a different location for monitoring or control purposes. The uncertainty of pressure transmitters is typically better than 0.1% of span if compared to reliability and robustness of pressure gauges and switches. The sensor module measures the pressure, converts the raw sensor measurement to a digital value and applies any corrections for variances between the calibration temperature and ambient temperature and for differential pressure transmitters for variances between the calibration and process static pressures. The sensor module passes its measurement and diagnostic information to the electronics module. The module then communicates this information to the control or monitoring system using either 4-20 mA analog signals or digital signals such as Foundation Fieldbus. Pressure transmitters are typically installed with instrument valves for verification and maintenance purposes. Figure 3.4 shows an example of instrument manifolds for

differential pressure transmitters.

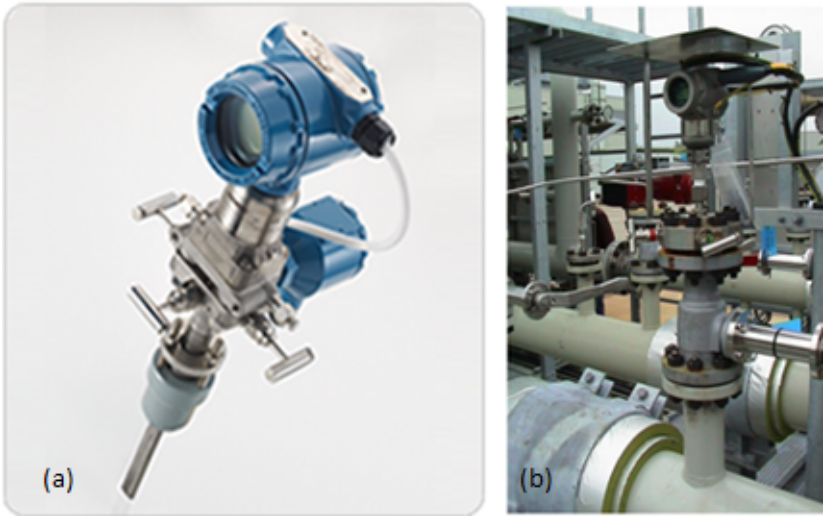

Figure 3.4: (a) Pressure transmitter, (b) instrument manifolds example

Temperature is another important data to calculate a compressor performance because the enthalpy of a gas is a much stronger function of temperature than pressure. However, it is more difficult to obtain accurate temperature measurements due to the slow response nature of temperature and the boundary layer effect in piping. Important considerations in the field are the accuracy/calibration of the temperature sensing device, its location, and installation. The two most common devices used to measure temperature are thermocouples and resistance temperature detectors (RTD), supported in a temperature probe. The functions of temperature probes are to provide mechanical support for temperature sensors (RTD and thermocouple), locate the sensor in close proximity to the thermowell tip, isolate the sensor from ambient moisture. Temperature probes are often spring loaded to ensure the probe tip is positioned at the bottom of the thermowell bore to ensure rapid responses to temperature changes. Temperature assemblies combine a temperature probe with a connection head as illustrated in Figure 3.5. The connection head may contain a temperature transmitter or a simple terminal strip for making connections between the temperature probe and end device (e.g., PLC input card) located elsewhere. Temperature transmitters

convert the raw temperature sensor output (e.g., resistance, voltage) in a temperature reading and output the reading via analog (e.g., 4-20 mA) or digital means (e.g., Foundation FieldBus). With suitable designs, temperature sensors may be direct wired to a PLC, a DCS or a flow computer.

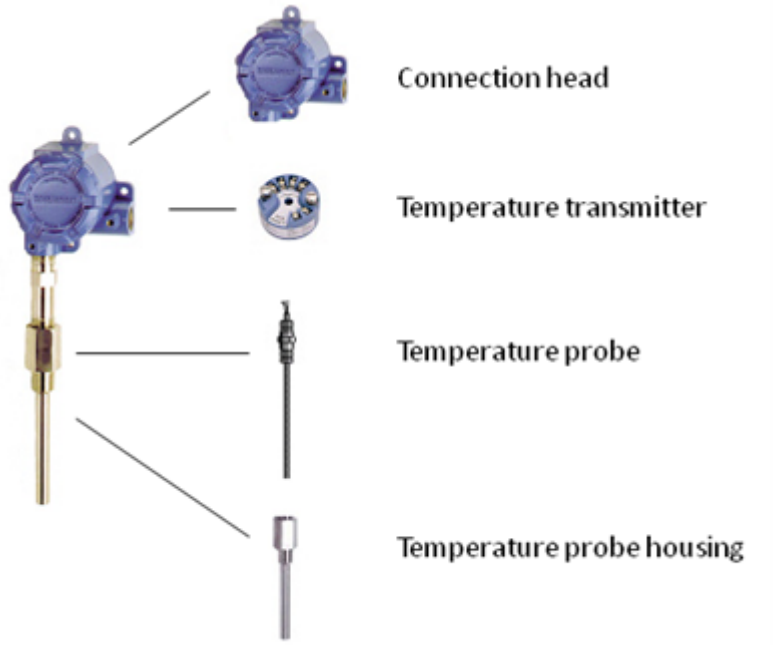

Figure 3.5: Temperature probe and transmitter

Other than the process measurements of pressure and temperature explained earlier, vibration is another critical machinery performance data. All rotating machinery have a portion of the "absorbed energy" converted to vibration. Vibration measurements may be expressed in terms of the type of acceleration or known as rate of change of velocity, the rate of change of displacement or known as velocity, displacement or the actual movement from a rest position, amplitude and frequency range of the sensor. The type of sensors may vary greatly in form and price; however, there are two basic classes, the seismic transducer and displacement transducer. Seismic transducers use inertially referenced measurement or a measurement relative to a point in space. The displacement transducer senses change in position or vibration relative to the mounting frame. Radial vibration or proximity probes and axial position or displacement probes are type

of displacement transducers that are more widely used. The probes are shown in Figure 3.6 and Figure 3.7. Transducers are sensors that convert physical behavior to an analog electrical signal. Transmitters convert transducer analog signals to analog electrical signals used in industry, 4-20ma.

Figure 3.6: Radial vibration probes

Figure 3.7: Axial proximity probe mounted on a thrust bearing

3.1.3 Process Flow Overview

All 46 tags of data extracted for this research can be divided into 8 categories and are discussed in the subsequent subsections. Figure 3.8 provides a high-level process flow of how the gas producing platform operate for better appreciation in the context of running a compressor. Starting from the wellhead or drilling platform, fluid from wells is sent to the production platform of which contains the water separation, condensate dehydration and gas dehydration facilities. A compression platform is connected to production platform, the compression facilities were installed to increase the gas reservoir pressure. Wet gas undergoes further water removal in the suction scrubber before it is compressed to export pressure through a single gas turbine driven centrifugal compressor. Wet gas is then routed from two production trains before the compressed gas is fan cooled and sent back to the glycol contactor tower on the production platform for dehydration. Majority of the dehydrated gas is sent to export gas line before being sent to shore for further gas treatment via trunklines. The remainder of the dehydrated gas is used as fuel gas, and for purging and blanketing purposes.

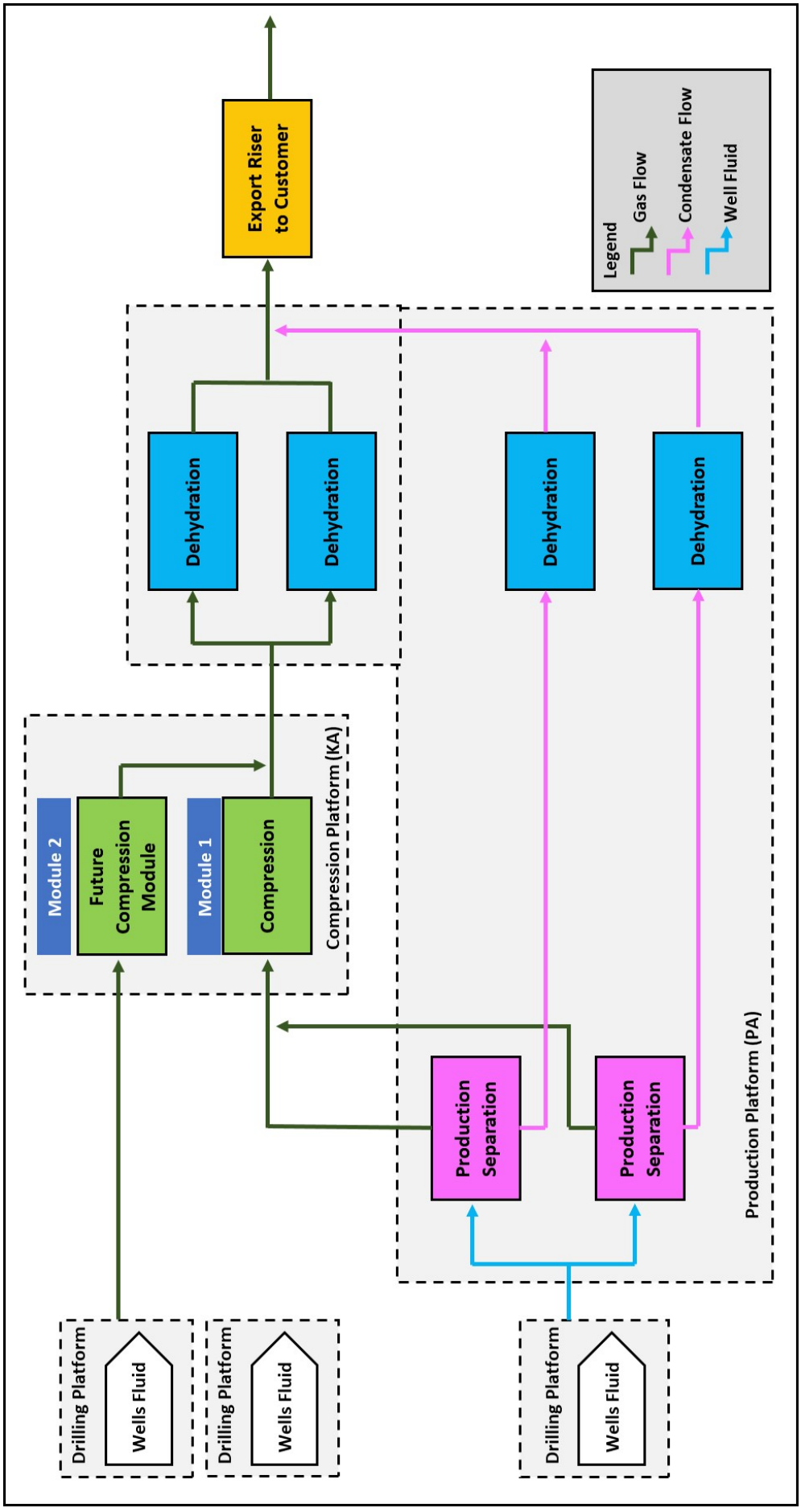

Figure 3.8: Process flow overview

3.1.4 Compressor and Condensate Export

As mentioned earlier regarding the suction scrubber, its function is to protect gas compressors from fine particulate and condensate. Without proper filtration, compressor performance is reduced and if this prolonged, internal parts might damage. Gas from the compression suction scrubber is routed to the compressor unit. The vessel internal components consist of a Schoepentoeter inlet device, a swirldeck and a mist mat, fitted to optimize gas or liquid separation and to coalesce and remove any liquid droplet entrained in the gas flow. Figure 3.9 shows the high and low liquid level is detected by level transmitter and suction scrubber is also fitted with level gauge which gives indication of liquid level in the vessel. These are important data to monitor performance as they send signal that initiate alarm and trigger unit pressurized shutdown when high level in vessel is detected. Table 3.1 shows the list of data features used in the models.

Figure 3.9: Level gauge, level transmitter and level bridle

Table 3.1: Data tags for Compressor Condensate Export

No.	Tag Description	Unit
1	Tilted Plate Coalescer V-270	%
2	Tilted Plate Coalescer V-280	%
3	Suction Scrubber Level	%
4	E-2410A/B Discharge Temperature	degC
5	E-2410C/D Discharge Temperature	degC

3.1.5 Compressor Thermodynamic Performance

Compressor thermodynamic analysis of a compressor is the most straight-forward way to determine its health. However, it requires specific instrumentation of sensors and transmitters to be installed in the unit, as shown in Figure 3.10. Compressor suction, discharge pressure and temperature are normally well maintained by continuously controlled machine speed. Pressures will vary with machine speed and process gas composition, the purpose is to trend and recognize significant changes in a stage performance resulting from fouling, mechanical issues or possible fouling of inter-cooler heat exchangers. Compressor fouling refers to the build-up of unwanted materials causing the surface of the compressor blades rough. Temperatures can vary with machine speed, stage differential pressure and process gas composition. Changes indicate efficiency losses that could be attributed to fouling, excessive wear or internal damage to the machine. The data features used as input into these models are listed in Table 3.2.

Table 3.2: Data tags for Compressor Thermodynamic Performance Monitoring

No.	Tag Description	Unit
1	Suction Pressure	barg
2	Discharge Pressure	barg
3	Suction Temperature	degC
4	Discharge Temperature	degC

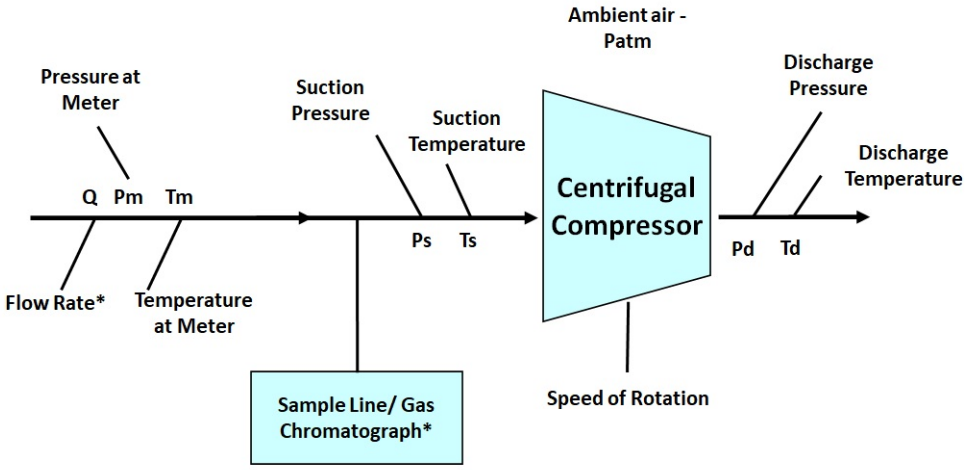

Figure 3.10: Typical instrumentation for compressor monitoring setup

3.1.6 Compressor Mechanical Performance

The primary mechanical protection on the main compressor train comes from axial thrust probes as shown in Figure 3.11. The thrust bearing maintains the position of the rotor relative to the case and prevents rotor to case contact. In the event of a thrust bearing failure, severe damage can result to both rotating and stationary machine components, hence the measurements of axial thrust is to protect against this damage. These are consider machinery protection system for critical machine trains, hence measurement required high reliability. Axial vibration is a good indication of machine surge, and axial rubs in machinery due to fouling or thrust bearing failure. Compressor surge is a condition when the amount of gas they are trying to compress is insufficient for the size of the compressor and the blades lose their ability to transfer energy from the shaft to the fluid, causing a reverse flow of the gas. Radial vibration is used as an indication for many machine malfunctions including unbalance, fouling, rubs, instability, seal problems etc. Typically two probes per bearing in XY configuration and these probes are voted 2 out of 2. The list of data features extracted to run in

these models is shown in Table 3.3.

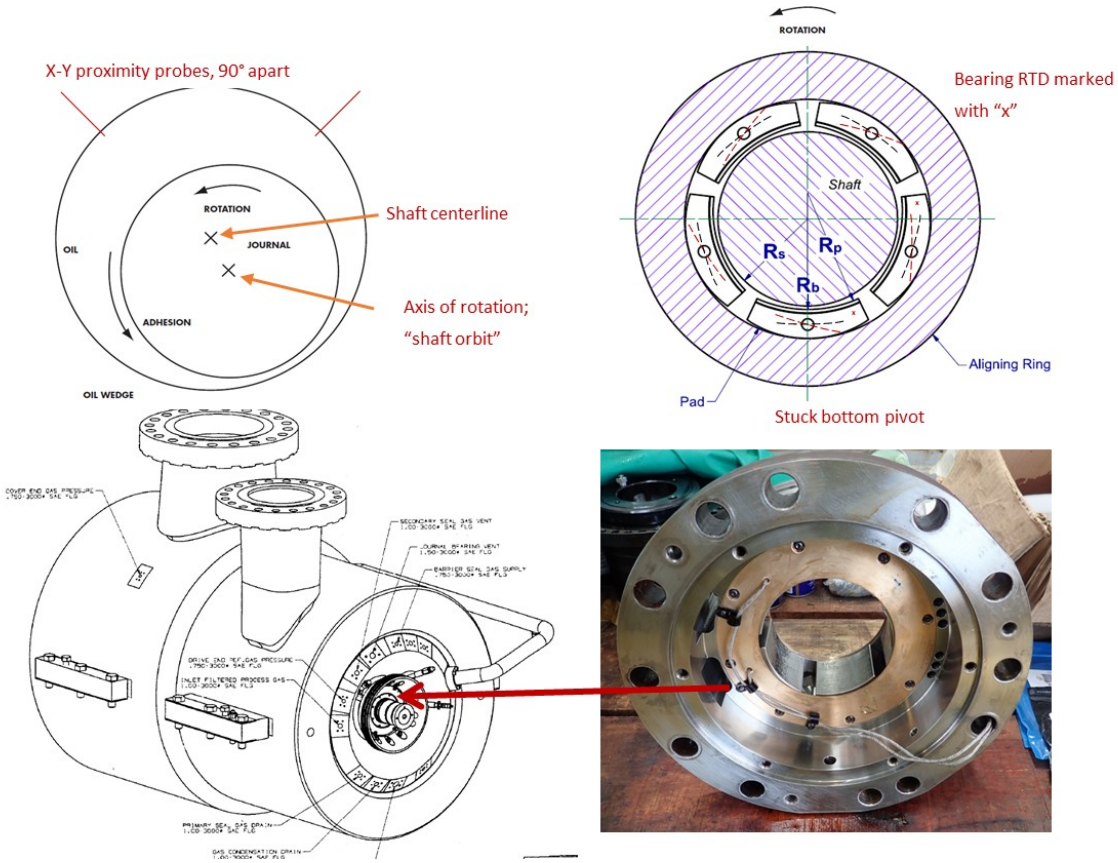

Figure 3.11: Axial thrust probes in tracing high radial vibration, rotor behavior and shaft orbit

Table 3.3: Data tags for Compressor Mechanical Performance Monitoring

No.	Tag Description	Unit
1	Vibration Drive End Radial X	µm
2	Vibration Drive End Radial Y	µm
3	Vibration Non-Drive End Radial X	µm
4	Vibration Non-Drive End Radial Y	µm
5	Axial Position A	mm
6	Axial Position B	mm
7	Drive End Radial Bearing Temperature	°C
8	Thrust Bearing Active Temperature 1	°C
9	Thrust Bearing Active Temperature 2	°C
10	Non-Drive End Radial Bearing Temperature	°C

3.1.7 Compressor Turbine Thermodynamic and Mechanical Performance

The primary function of the gas turbine engine in a compressor is to supply air in sufficient quantity to satisfy the requirements of the combustion burners. Vibration levels are monitored as they can be indicative of a change of balance or bearing wear. Figure 3.12 shows accelerometers are mounted on the casing of the gas generator and power turbine, it is important to ensure that the correct type of vibration sensors is fitted so that good machine protection is achieved and also using the data for trending purposes. Proximity probes are also fitted in the journal bearings because heavy casings may not be sensitive enough to detect rotor or bearing problems, especially when externally induced casing vibration is present which may mask vibration generated by the turbine. An increase of casing vibration may only be noted when a catastrophic failure has occurred. Proximity probes are selected because they give the actual relative motion between the rotor and bearing irrespective of the casing vibration. The challenges of proximity probes is often inaccessibility in the case of failure, and difficulties of checking calibration. Table 3.4 shows the list of features extracted from this monitoring category used as input in this research.

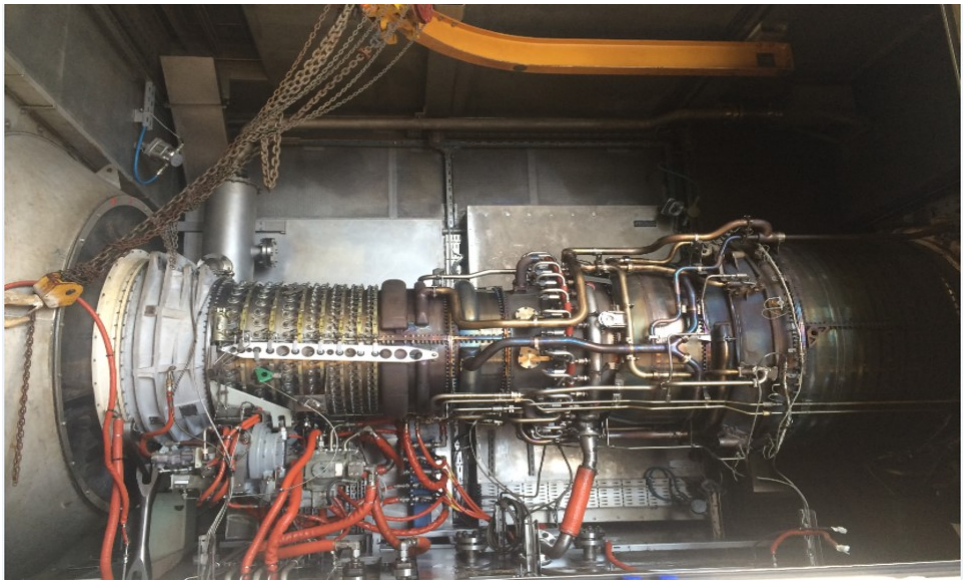

Figure 3.12: Turbine engine installed with proximity probes and accelerometer

Table 3.4: Data tags for Compressor Turbine Thermodynamic and Mechanical Performance Monitoring

No.	Tag Description	Unit
1	Compressor Discharge Air Pressure	barg
2	Gas Generator Bearing 1 Vibration	mm/s
3	Gas Generator Bearing 2 Vibration	mm/s
4	Power Turbine Bearing 4 Vibration	mm/s
5	Power Turbine Bearing 5 Vibration	mm/s

3.1.8 Gearbox Mechanical Performance

The purpose of a gearbox as shown in Figure 3.13 is to increase or reduce speed. As a result, torque output will be the inverse of the speed function. Torque is the turning force when load is applied at a distance away from the center of rotation. If the enclosed drive is a speed reducer (speed output is less than speed input), the torque output will increase; if the drive increases speed, the torque output will decrease. Monitoring gearbox performance is to determine potential bearing or rotor dynamic problems. Impact can be catastrophic when there is event of bearing damage leading to high gearbox vibration transferring to compressor

leading to seal failure on compressor. List of features from this category is shown in Table 3.5.

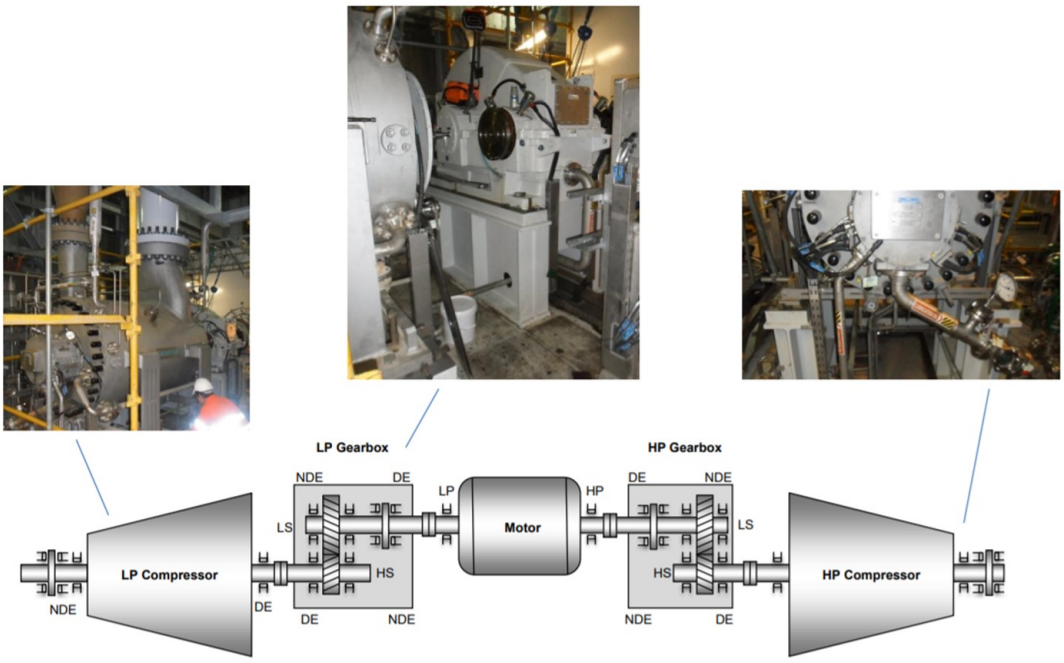

Figure 3.13: Gearbox location

Table 3.5: Data tags for Gearbox Mechanical Performance Monitoring

No.	Tag Description	Unit
1	High-Side Shaft Drive End Radial Vibration X	µm
2	High-Side Shaft Drive End Radial Vibration Y	µm
3	High-Side Shaft Non-Drive End Radial Vibration X	µm
4	High-Side Shaft non-Drive End Radial Vibration Y	µm
5	Low-Side Shaft Drive End Radial Vibration X	µm
6	Low-Side Shaft Drive End Radial Vibration Y	µm
7	Low-Side Shaft Non-Drive End Radial Vibration X	µm
8	Low-Side Shaft Non-Drive End Radial Vibration Y	µm
9	Low-Side Shaft Axial Position	mm
10	High-Side Shaft Non-Drive End Radial Bearing Temperature	°C
11	Low-Side Shaft Drive End Radial Bearing Temperature	°C
12	Low-Side Shaft Non-Drive End Radial Bearing Temperature	°C
13	Low-Side Shaft Thrust Bearing Temperature 1	°C
14	Low-Side Shaft Thrust Bearing Temperature 2	°C

3.1.9 Turbine Enclosure Monitoring

The gas turbine burner may be considered as a source of ignition, as it introduces a flame for controlled combustion, hence the gas turbine should be located outside the hazardous area. If the turbine is located within an enclosure it should have an adequate ventilation system. The enclosure ventilation fans provide cooling air flow through the enclosure to prevent damage to heat sensitive components. The operation of the ventilation fans also maintains the turbine enclosure at a slight positive pressure to prevent potential of hydrocarbon gases from the surrounding area entering the turbine enclosure. Enclosure high temperature will trip compressor unit as failure of early detection can be catastrophic as shown in Figure 3.14 and 3.15 below. The list of data features used as input into these models is listed in Table 3.6.

Figure 3.14: Enclosure temperature HH tripped on NPT signal failure caused by oil and debris ingressed into the probes causing erroneous signal

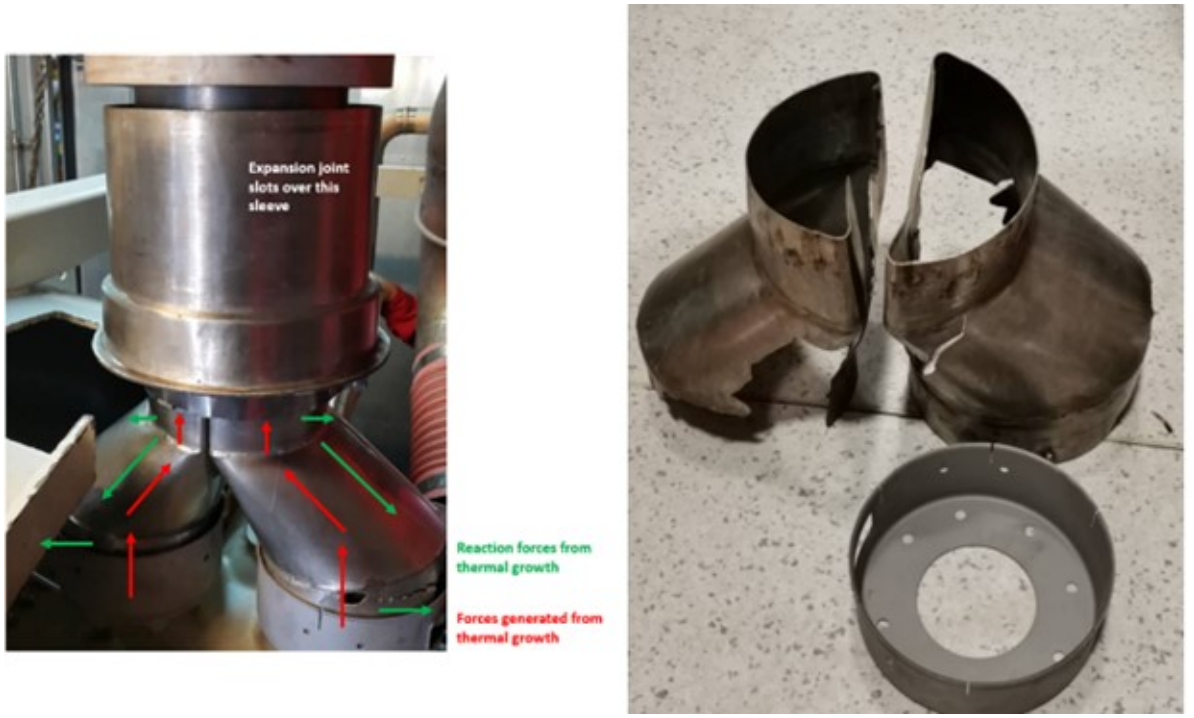

Figure 3.15: Enclosure high temperature due to hot air leaking inside the enclosure

Table 3.6: Data tags for Turbine Enclosure Monitoring

No.	Tag Description	Unit
1	Turbine Enclosure Temperature	°C
2	Turbine Enclosure Differential Pressure	mbar

3.1.10 Dry Gas Seal System

Compressors are typically equipped with dry gas seals to prevent gas from escaping between the stationary compressor body and the rotating shaft. Compressors are normally shut down when high seal leakage occurs, indicating a seal failure. Compressor can experience a catastrophic failure of the suction end dry gas seals (both primary and secondary) resulting in high pressure gas entering the gas turbine driver's enclosure through the common lube oil system. Figure 3.16 shows an event of gas and lube oil ignited, resulting in a fire that heavily damaged the turbine enclosure. Leaking bearing housing seals are sometimes given low priority, transmitter reliability is critical in providing data to decide if early intervention is needed. The monitoring scope includes trending of all the primary seal vent pressure, seal gas supply pressure and seal gas differential pressure (dp) if the reading is genuinely high or low. Table 3.7 shows the list of features extracted from this category used as input in this research.

Table 3.7: Data tags for Dry Gas Seal System Monitoring

No.	Tag Description	Unit
1	Primary Seal Gas Supply Pressure	barg
2	Separation Gas Supply Pressure	barg
3	Drive End Primary Vent Pressure	barg
4	Non-Drive End Primary Vent Pressure	barg
5	Seal Gas Filter dP	barg
6	Seal Gas dP	barg

Figure 3.16: Fire incident due to dry gas seal failure

3.1.11 Data Pre-processing

Data may be obtained automatically from a variety of sensors. Many different OSIsoft and third-party PI Interfaces are used to obtain the majority of the information. Users may then access this information using a standard set of tools (such as Microsoft Excel, a web browser, or the PI Process Book) and check for connections between the data points collected. All tags were analogue to simplify the model i.e., to avoid unnecessary increase of input dimensions if discrete tags were used. The coverage of the data tags includes mechanical, auxiliary, and overall process systems of the compressor. Moreover, the min-max approach was used to normalize the features such that their values fall inside the range [0,1].

The min-max formula is given by

$$x_i = \frac{\bar{x}_i - \bar{x}_{min}}{\bar{x}_{max} - \bar{x}_{min}}, \tag{3.1}$$

where x_i is the normalized feature data, $\bar{x}_i$ is the original feature data, $\bar{x}_{min} = \min(\bar{x}_i)$ and $\bar{x}_{max} = \max(\bar{x}_i)$. Furthermore, x denotes as the feature vector in time t of which the element is obtained by using Equation 3.1 and y as the output (RPM) at time t, and $(x, y) \in [0, 1]$. Here, $x \in R^D$ is the D-dimensional multivariate features representing a normalized sensor input.

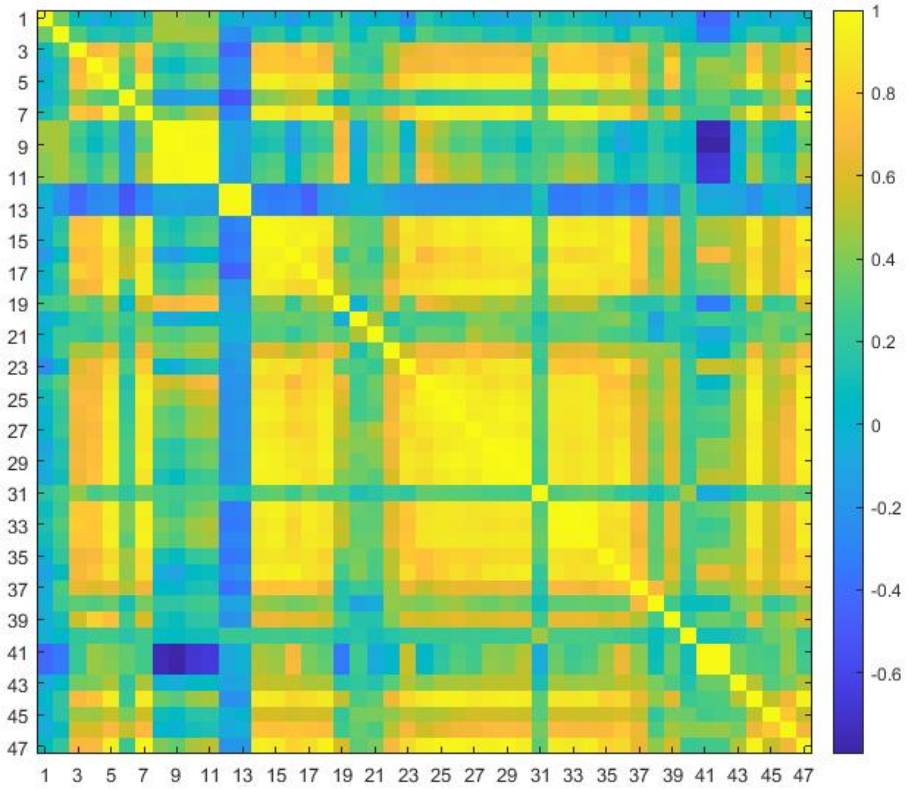

Figure 3.17: Correlation between the extracted features

As a preliminary investigation into the correlation, Fisher-Correlation was used to evaluate the correlation between the features. Feature 1-46 represent the various features in consideration while column 47 represents the normalized

RPM value which is the regression target. Evidently, the diagonal entries are 1.00 showing correlation of the feature to itself. The figure shows some highly correlated features to the normalized RPM value. The equation 3.2 express the correlation between the 2 variables and $\hat{B}$ and $\hat{A}$ are average of the features respectively. Part of the data are available at (Wong, 2022). Due to IP and policy issue, this data is available for thesis defense (from 1st July 2022 to November 2022 only).

$$r = \frac{\sum\sum(A_{mn} - \hat{A})(B_{mn} - \hat{B})}{\sqrt{\sum\sum(A_{mn} - \hat{A})^2 (B_{mn} - \hat{B})^2}} \tag{3.2}$$

During the data pre-processing stage, there are additional considerations and steps taken in order to improve the data quality. Industrial data often faced with data quality challenges and this is unavoidable. The poor quality observed in the data are such as a string text was found in a numerical data tag, the error is caused by communication failures. When these errors are detected, affected data is cleaned by back-filling with the most recent valid value. When timestamp is unable to obtain reading, the normal practice is to use the previous T-1 sensor data. There are four possibilities when handling data with zero reading, replace with minimum, maximum, average or zero. Since data is continuous, it is a common practice to replace with previous values by assuming data will not change within the time interval (Kang, 2013). Sensor data collected is at both transient and steady state, as seen in the data collected over more than 2 years. Time domain data is applied and used for static modeling. Low variance tags are removed as they are deemed inconsequential to sensor failures, they will not vary sufficiently in the training duration to influence the instrument operating conditions. Numerically, tags that behaved similar to a discrete data type are excluded in the training data. There is no outlier data in regression model. The extracted data was reviewed together with the operations and discipline

engineers to further detect any possible data error that could affect quality of the research and to collate information regarding health status and process condition of the instrument. During the review sessions, 14 tags were removed based on feedback from the engineers of which later the tags were confirmed for having high multicollinearity, including them will distort the model outcome. Besides cleaning bad quality data, historical trip events were also reviewed, events that are not trendable to be excluded from the training data. Examples of such events were manual shutdown for preventive maintenance work, nuisance instrument failure with instantaneous spikes, and compressor starting up and ramping down.

3.2 Multiple Linear Regression and Neural Networks Models

Figure 3.18 illustrates the entire structure of the thesis. As illustrated, the data collection phase was describe in detail in previous sections. In the phase 1, plant modelling was implemented such that sensors deviation from normal or predicted can be quantified. this is implemented using MLR, ANN, and GP. As explained in introduction section , GP serves as our main target of implementation while ML and ANN acts as comparison /benchmark approaches.

Establishing that ML concepts can be highly useful for modelling complex relationships and building models for machines, similar concepts can be used for sensor abnormality detection. In this case, the states of machines may be modelled using ML, even if the model is complicated. The deviation at this point is that sensor outputs or working conditions may be modelled by the output of other sensors. In essence, it is a cross-correlation concept for sensors. The initial intuitive decision is to choose the simplest model and proceed to a more complex model if modeling complexity is insufficient. Note that the compressor machine under examination is somewhat old and is best suited to ML modeling. Among the candidates, artificial neural network and multiple linear regression fit this

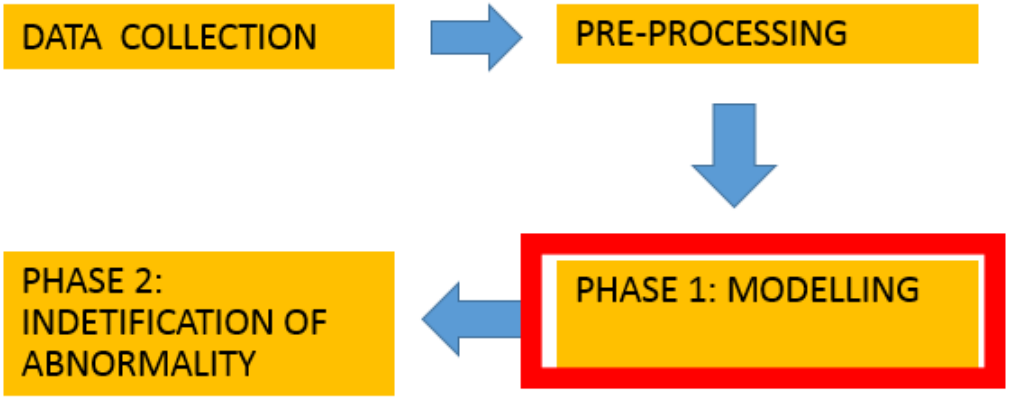

Figure 3.18: Block diagram : modelling of plant

description. Inevitably, neural networks are prone to vanishing gradients due to the application of the transfer function, such as the tangent sigmoid function. However, neural networks have an advantage of modelling more complex systems. The two main disadvantages of the neural networks are that these models are considered as a 'black box' and difficult to comprehend. Despite being efficient on record, they are normally not preferred in certain industries.

3.2.1 Multiple Linear Regression

Multiple linear regression is a multivariate version of the linear regression. Let $(y_n, x_{n,i})$ be the value and feature value pair for $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, I$. The multiple linear regression model for N observations and I features is given by

$$y_n = \sum_{i=1}^{I} x_{(n,i)}\beta_i, \tag{3.3}$$

where β_i are the weighted coefficients to the particular variable. This can be rewritten in a vector form as follows

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta}^T, \tag{3.4}$$

where $(\cdot)^T$ is the transpose operator.

The objective of the function is to minimize the least square error between $\boldsymbol{y}$ and $f(\boldsymbol{x}, \boldsymbol{\beta})$ is minimal. The weight for each variable is as solved using the least square method as follows

$$\beta_i = \frac{N\sum_{n=1}^{N} x_{(n,i)}y_n - \sum_{n=1}^{N} x_{(n,1)}\sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_{(n,i)}^2 - (\sum_{n=1}^{N} x_{(n,i)})^2}, \tag{3.5}$$

Note that the problem is almost always convex.

3.2.2 Curve Fitting Neural Networks

Neural networks generally work by cascading interconnected nodes linking the inputs to the output forming a series of networks. Each node represents a function of the weights and the values from the previous nodes. Hence, the output of a node is a function of $\boldsymbol{\beta}\boldsymbol{x}^T + \boldsymbol{\epsilon}$ (multi-regression), where $\boldsymbol{\beta}$ is the weight vector depicted by the edges and $\boldsymbol{\epsilon}$ is the bias value. The expression for the transfer function for each node is as expressed as

$$S\left(\sum \boldsymbol{\beta}\boldsymbol{x}^T + \boldsymbol{\epsilon}\right) = \frac{1}{1 + e^{-\sum \boldsymbol{\beta}\boldsymbol{x}^T + \boldsymbol{\epsilon}}}, \tag{3.6}$$

where $S(\cdot)$ is the sigmoid transfer function. For simplicity, from now on, the vector $\boldsymbol{\beta}$ denotes a solution of $1 \times n + 1$ where n is the number of variables ($\boldsymbol{\beta}$ include the bias term $\boldsymbol{\epsilon}$).

Neural networks have been deployed for various modelling works. Generally, there are various means of which includes gradient descend or stochastic methods. In each iteration, the weight vector, $\boldsymbol{\beta}$ is adjusted with the first order Jacobian.

Given in each iteration $(\boldsymbol{\beta} + \boldsymbol{\delta})$, $\boldsymbol{\delta}$ can be calculated such that $f(x_i, \boldsymbol{\beta} + \boldsymbol{\delta}) \approx f(x_i, \boldsymbol{\beta} + J_i\delta_i)$ for the ith instance (row), where $J_i = \partial f(x_i, \beta_i)/\partial\beta$ is the Jacobian matrix.

The sum of the squared error, $S(\boldsymbol{\beta} + \boldsymbol{\delta})$, is given by

$$S(\boldsymbol{\beta} + \boldsymbol{\delta}) \approx \sum_{i=1}^{m} [y_i - f(x_i, \boldsymbol{\beta}) - J_i\boldsymbol{\delta}]^2, \tag{3.7}$$

where $\boldsymbol{y}$ is the target vector. Derivation of $S(\boldsymbol{\beta} + \boldsymbol{\delta})$ w.r.t $\boldsymbol{\delta}$ gives

$$(J^T J)\boldsymbol{\delta} = J^T(\boldsymbol{y} - f(\boldsymbol{\beta})). \tag{3.8}$$

Adding a damping factor, λ to adjust the $\boldsymbol{\delta}$ yields

$$(J^T J + \lambda \boldsymbol{I}) = J^T(\boldsymbol{y} - f(\boldsymbol{\beta})). \tag{3.9}$$

Note that λ can be considered as a form of learning rate. Furthermore, the weight vector $\boldsymbol{\beta}$ can be solved using Gauss-Newton method.

3.3 Data Collection and Curve Fitting Processes

As explained earlier, the purpose is to first model sensor outputs and subsequently use model outputs to predict residuals. The equipment under consideration is a multistage centrifugal compressor used to provide pressure for gas transport, these machines are essential in gas production platforms. Data were collected from the data logging system, which captures the information in a 10-minute interval basis for more than 2 years. These data were streamed and stored in a local database. According to the literature, most of the research works in the similar areas have applied high frequency data analytic for predicting machine failure. This is ideal, but it is difficult to be implemented due to high computational complexity. Furthermore, this approach relies on a static approach and not on

the dynamic fluctuations to detect the abnormality. Therefore, this research investigates the lower frequency data for predicting such failure. The remaining 46 features are normalized such that the values are in the range of [0,1] using the min-max method given by

$$x_i = \frac{\bar{x}_i - \bar{x}_{min}}{\bar{x}_{max} - x_{min}}, \tag{3.10}$$

where x_i is the normalized feature data, $\bar{x}_i$ is the original feature data, $\bar{x}_{min} = \min\{\bar{x}_i\}$, and $\bar{x}_{max} = \max\{\bar{x}_i\}$.

x denotes as the feature in time t and y as the output (rpm) at time t, where $(x, y) \in [0, 1]$ is obtained by using Equation 3.10. Here, $x \in \mathbb{R}^D$ is the D-dimensional multivariate features representing a normalized sensor input. The detail of these sensor readings or features is as attached in the Appendix A.

3.3.1 Settings

The plant management systems were used to collect data for training and testing, which is a typical practice for oil and gas platforms in many regions. The data collected were at a 10-minute interval, which is regarded appropriate because increasing the frequency of data collection will dramatically increase data cost. More than 2 years' worth of data were collected for processing and training, which is long enough to generalize faults. Sensor readings of many types, such as temperature, vibration, pressure, and displacement sensors, are used to create data (please refer to Appendix A). Two models were applied, i.e., neural network and multiple regression models. In order to justify this application, the RPM shaft sensor output is predicted using the outputs of other sensors. The accuracy of the models is demonstrated by the R^2 values.

For the neural networks, a single hidden layer of 10 neurons was applied in the layer with tangent sigmoid as the processing transfer function and a learning rate of 0.1. The Levenberg-Marquardt algorithm was also applied to reduce computational training time as it enables gradient descent to be calculated without computing the Hessian matrix (2^{nd} order). Noted that the data sets were skewed

towards the normal condition for the output < 4000 rpm. The variances in data set are important such that the ML models may capture the different states of the compressor. The data are separated into 70%, 15%, and 15% for training, validation, and testing, respectively, for the case of neural network and 70% - 30% (training - testing) for multiple linear regression cases. We note that there is a fundamental difference in training process between neural network and multiple linear regression. It can be considered that the data split is the best approach, given not much discussion from the academic literature on this comparison, and for the sake of best practices for comparison.

The overview of the entire concept revolves around training machine learning models and comparing the residuals (machine learning model output and observed data) to identify the type of faults. As such, machine learning models need to be fairly accurate and computationally light weight to calculate the residuals. This is investigated this in the subsequent section. This approach applies "conventional" training procedures to train the models. The parameter setting and the model parameters description are justified in the model comparison section. Subsequently, the concept of residual comparison and by calculating the residuals (between the predicted and the independent test data) is tested. This thesis is demonstrated by observing the histogram of residuals, the type of faults can be identified (histogram analysis section).

3.3.2 Genetic Programming

Genetic Programming (GP) was originally designed to generate equations using evolutionary approach (stochastic approach). There are several variants in GP, such as those that utilize tree as equation encoding (tree-based GP) and graph (Cartesian GP). This section details the mutation mechanism of genetic programming. The idea behind the tree-based GP is to use encoded representations of generated equations to evolve over generations, which is similar to other evolutionary algorithm approaches. The methods are described in detail in Kai (2017).

In general, the tree-based GP applies four evolution operations: i) reproduction, ii) point mutation, iii) branch mutation, and iv) crossover. Figure 3.19 shows a single equation tree representing an equation. The depth of the equation tree is illustrated in the figure. The corresponding fitness of each depth configurations are shown in Table 4.12. As shown in the figure, the depth of the tree indicates the complexity of the equation. In this research report, the allowable maximum depth of 4, 5, 6, 7, 8, 9, 10 are evaluated. From the population, a set number of decision trees are chosen for a tournament, with the winners being promoted to the next generation. It's worth noting that the term "reproduction" refers to an exact clone of an existing candidate. There are no mutations added, and the solutions are simply copied from the populations. Branches are randomly picked from the population in branch mutation. A branch is a large portion of the tree that contains the points and function nodes (Figure.3.20). In crossover mutation, two trees selected as to produce two off springs with each appearance device are swapped between the two parents to obtain two child solution candidates. In point mutation, a single point of the tree is chosen for the operation Figure. 3.21, function notes are preserved, and only points are changed at random. This might be regarded as a tree's modest random mutation.

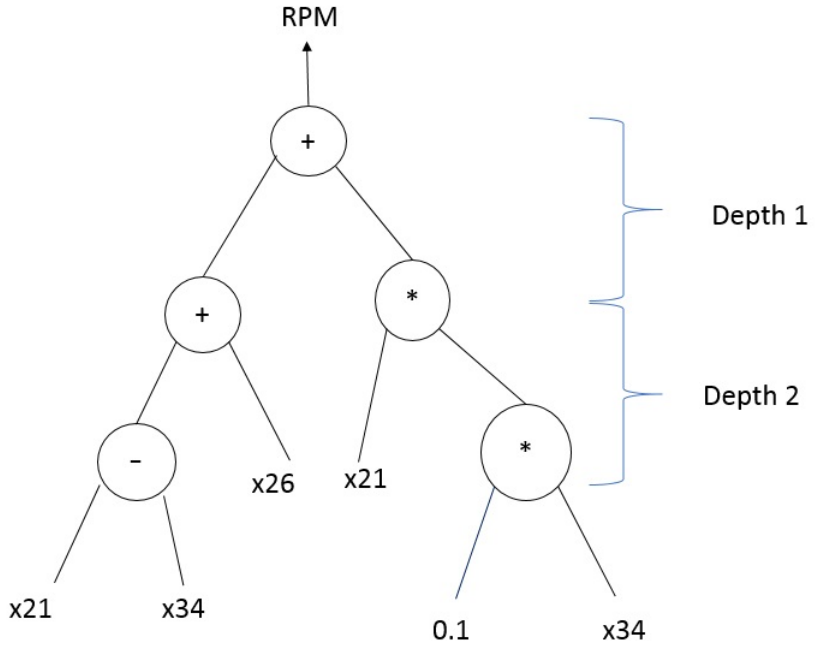

Figure 3.19: Structure of Equation Tree

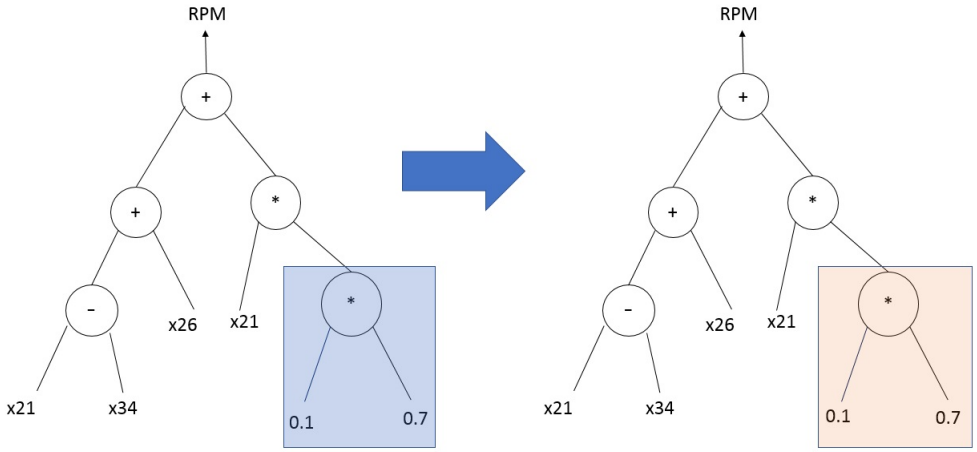

Figure 3.20: Branch mutation

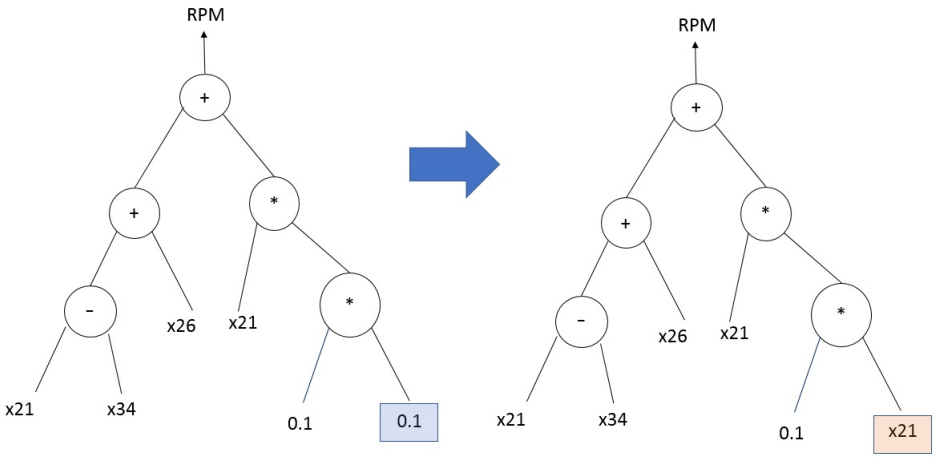

Figure 3.21: Point mutation

Table 3.8: Table of function list

Basic function			
$\alpha + \beta$	$\alpha - \beta$	$\alpha \times \beta$	$\alpha \div \beta$
Extended Functions			
$\alpha + \sqrt{\beta}$	$\alpha - \sqrt{\beta}$	$\alpha \times \sqrt{\beta}$	$\alpha \div \sqrt{\beta}$
$\alpha + \beta^2$	$\alpha - \beta^2$	$\alpha \times \beta^2$	$\alpha \div \beta^2$
$\alpha + \beta$	$\alpha - \beta$	$\alpha \times \beta$	$\alpha \div \beta$
$\alpha + cos(\beta)$	$\alpha - cos(\beta)$	$\alpha \times cos(\beta)$	$\alpha \div cos(\beta)$
$\alpha + sin(\beta)$	$\alpha - sin(\beta)$	$\alpha \times sin(\beta)$	$\alpha \div sin(\beta)$
$\alpha + tan(\beta)$	$\alpha - tan(\beta)$	$\alpha \times tan(\beta)$	$\alpha \div tan(\beta)$
$\alpha + lop1p(\beta)$	$\alpha - lop1p(\beta)$	$\alpha \times lop1p(\beta)$	$\alpha \div lop1p(\beta)$
$\alpha + sign(\beta)$	$\alpha - sign(\beta)$	$\alpha \times sign(\beta)$	$\alpha \div sign(\beta)$

3.3.3 Simulation of Errors

It is certainly not economical/feasible to perform invasive approach to acquire fault data. There is no available data on public domain for such purposes. In the approach to acquire fault data, reference was made to the proposal by (Tsai et al., 2019). Figure 3.22 shows the block diagram corresponding to this phase on investigation. This phase is carried subsequent to the sensor reading modelling phase. In the proposed approach, 3 types of faults were introduced : 1) complete/ constant faults 2) bias drift faults 3) degradation faults. The approaches to generate the fault readings are subsequently discussed.

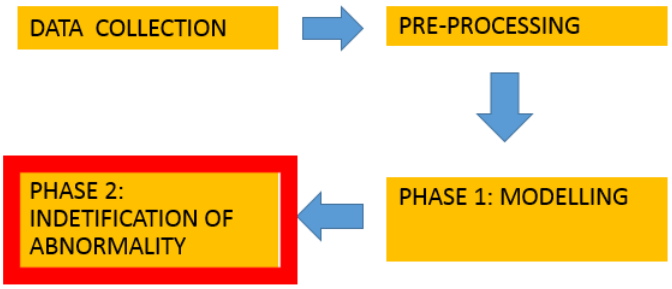

Figure 3.22: Block diagram : Abnormality detection

- Constant faults : pick successive samples , and replace them with constant value
- Bias drift faults: pick successive samples and replace each value v_i with $v_i + constantvalue$
- Degradation faults : pick successive samples start from i and replace value v_j with $v_j + degradation_rate * (j - i)$

  Noise faults : select a sensor measurement v_i and multiply a factor f determines the intensity of the noise fault a noise data $v_i' = v_i * f$ where $f = 1.2, 1.5, 2, 5, 10$

3.4 Summary of Chapter 3

In summary, the chapter outlines the entire methodology involved in achieving the objectives stated. The goal of the research is to be able to detect sensor faults over a period of time. 3 machine learning models were suggested which consist of Multi linear regression (MLR), Neural network (ANN) and Genetic Programming (GP). ANN are black box machine leaning models while MLR are not suitable for non-linear mapping. In critical equipment, a more transparent model such as mathematical model is preferred. This is the justification of applying GP for mathematical model generation over the other alternative. Both MLR and ANN may provide a comparative performance. Black box machine learning models can be dangerous in certain circumstances. Firstly, it lacks transparency. It can be difficult to understand how a black box model is making its predictions, which can make it difficult to detect and prevent bias, errors. The behavior of a black box model can be hard to predict, which can lead to unintended and potentially harmful outcomes, such as discrimination or misinformation. In some cases, the output of a black box model may be uncontrollable or difficult to manage, which can result in negative consequences such as economic loss or harm to individuals. Upon completion of model development, further evaluation on application to detect the sensor fault is done in the subsequent phase. Essentially the model will be used as a predictive tool by evaluating the differences between the observed and the predicted. These will be discussed in the subsequent chapter.

Chapter 4

Results and Analysis

As discussed in methodology chapter, this thesis consists of two phases. The first component includes the mathematical modelling of the data. The second involves implementation of the predictive modelling and using the model acquired for analyzing the error. With respect to presentation of the results, the first phase of the results is to evaluate the goodness of "fit of the model" acquired from the MLR/ANN and GP. The results were evaluated during the training and validation phases to identify associated patterns. The results for ANN/MLR were compared with the mathematical models generated by the GP algorithm. As discussed earlier, the GP mathematical models were preferred due to their "transparent" nature and ability to deal with non- linear mapping relatively well. Subsequently, the best GP model was selected for further evaluation of error/fault detection in the second phase. The two sections are named Phase 1 evaluation of model fitness and Phase 2 evaluation of the identification of faults from the histogram. For Phase 2, the error simulation approach was applied as it was not feasible to acquire sufficient amount of faulty sensor data reading.

4.1 Phase 1 Evaluation : Model Fitness

The Genetic Programming (GP) approach was run on various configurations and compared to the MLR and ANN. The results are discussed in this chapter and then compared with the MLR and ANN. The models were ran for 10 trials in which the results are presented in subsequent sections. Upon comparing with MLR and ANN, the mathematical models were applied with augmented dataset to simulate error. R^2 metric was applied as a goodness of fit evaluation for the models. R^2 is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s) in a regression analysis. It is a value between $-\infty$ and 1, where a value of 1 indicates that the model perfectly fits the data, and a value near 0 suggests that the model has limited predictive ability. The Mean Squared Error (MSE) is a measure of the average squared differences between the predicted values and actual values in a regression model. R-squared is a statistical measure that represents the proportion of variance in the dependent variable that is predictable from the independent variable. There is a negative correlation between MSE and R-squared, meaning that as R-squared increases, MSE decreases. In other words, a higher R-squared value indicates a better fit of the model to the data, which results in a lower MSE value.

$$R^2 = 1 - \frac{\sum_i^{i=N}(y_i - \hat{y}_i)^2}{\sum_i^{i=N}(y_i - \bar{y}_i)^2}, \tag{4.1}$$

$$MSE = \frac{1}{N}\sum_i^{i=N}(y_i - \hat{y}_i)^2 \tag{4.2}$$

Where N is the number of samples, y_i and $\hat{y}_i$ are the observed and predicted values respectively

In this section, the machine learning models for regression will be evaluated. Prior to comparing with the mathematical models (from GP), both ANN and MLR were evaluated. Generally, ANN are "black box models that perform well

with non linear modelling. MLR are essentially a weighted regression model that are relatively "transparent" but do not thrive on non linear mapping. Multi Linear regression (MLR) is a statistical method that models the relationship between a dependent variable and one or more independent variables as a linear equation. It assumes a linear relationship between the independent and dependent variables, which means that changes in the independent variables produce a constant change in the dependent variable. Non-linear mapping, on the other hand, refers to mathematical process that maps input variables to output variables in a way that is not proportional. Non-linear relationships are characterized by the output changing at different rates for different values of the input. Non-linear regression is a method used to model non-linear relationships between the independent and dependent variables. It involves transforming the independent variables and fitting a regression model to the transformed data. The objective of the model is to acquire a mathematical approach to minimize target values with the predicted values.

Table 4.2 and 4.4 show the evaluation and comparison of models acquired using MSE metrices a feature reduced model and full feature set while Table 4.3 and 4.5 show the models in equivalent R^2 metrices. The reduced features set were acquired by considering the highest absolute weights in MLR. In order to gauge the model performance, mean squared error (MSE) and R^2 were used as fitness of models. Table 4.2 and 4.3 show the comparison of the results of MSE and R^2 before feature reduction for multiple linear regression and neural network models, respectively. P-values (T-test) show the significance of the mean difference. In conclusion, with a threshold of 0.001 significance, the difference between neural networks and multiple linear regression cannot be dismissed as a random chance. In other words, neural network fitting produced slightly better in that sense. It can be seen from Table 4.2 that MSE of neural network model is slightly better than the multiple linear regression model in terms of the mean error. However, the p-value may be too insignificant to justify the conclusion since it could happen due

to random chance. Furthermore, since there is no validation in the multiple linear regression model, the p-value for this set (validation set) cannot be computed. In this paper, 1% significance value is applied for the mean difference between multiple linear regression and neural network models. In other words, any p-value lower than 0.01 does not show that the mean difference is significant, but the difference is only due to chances.

The coefficient values for both neural networks and multiple linear regression models were obtained by averaging the values over 10 trials. Furthermore, it is common to rank feature significance using regression coefficients. The five most significant features/sensors correlated to the shaft RPM are listed in Table 4.1. The most significant sensor correlating to the RPM sensor is sensor #44, i.e., the dry gas seal system monitoring (please refer to Appendix A). Table values show the mean value of the 10 trials with the associated standard deviation. The results shown in Table 4.1 further prompt subsequent analysis of applying models after feature reduction was performed. Once again, feature reduction is another common method to reduce complex models to a more compact model in line with the "Occams' Razor" principle which states a more compact and concise model is preferred over a bloated model. As such, the top five sensor readings were selected for subsequent analysis.

Table 4.1: Feature ranking, mean, and standard deviation

Sensor Index	13	14	29	41	44
Average	0.61738	0.35238	0.48969	0.62478	3.6632
Std. dev	0.15271	0.023830	0.011879	0.13427	0.21069
Ranking	3	5	4	2	1

This would further add credibility on the findings. Furthermore, in-depth analysis was performed by observing the changes of the coefficients in 10 trials. Once again, the observed and verified feature (or sensor) of #44 was consistently dominant throughout the 10 trials for the case of multiple linear regression demonstrated by the high weight coefficient. Coefficient values in multiple linear regression may occur as positive and negative values and the absolute value

of the feature are generally and indication of dominance and importance for the regression model. The positive and negative values indicate the inverse and correlation relationship of the features, respectively. When the results were interpreted against the actual reliability records of the compressor, seal and bearings failures were indeed the bad actors which escalated to multiple extended compressor shutdowns in the past.

As mentioned previously, the features set is then reduced to only the top five features as acquired from the regression analysis. Table 4.5 and 4.4 show the MSE and R^2 after feature reduction for both models, respectively. Both implementations, using neural network fitting and multiple regression, do not deteriorate the results, thereby indicating a progressive feature reduction. In conclusion, both implementations (reduced or full feature set) may be equally applicable for these purposes.

From the Tables 4.3 and 4.5, it can be observed that there were no significant difference between feature selection dataset. This shows that the additional features does not cause degradation to overall model performance. Similar observation can be seen from Table 4.2 and 4.4, it can also be concluded that ANN performed slightly better than MLR as shown from the average higher R^2 and lower MSE. The results were subsequently compared to the GP generated mathematical models.

Table 4.2: MSE of multiple linear regression and neural network models

MSE	Train		Validation		Testing	
	MLR	NN	MLR	NN	MLR	NN
Mean	$1.76 \cdot 10^{-4}$	$3.29 \cdot 10^{-5}$	N/A	$4.05 \cdot 10^{-5}$	$1.831 \cdot 10^{-4}$	$4.05 \cdot 10^{-5}$
Std. dev	$1.01 \cdot 10^{-5}$	$7.94 \cdot 10^{-6}$	N/A	$1.21 \cdot 10^{-5}$	$2.38 \cdot 10^{-5}$	$1.94 \cdot 10^{-5}$
P-value	$4.97 \cdot 10^{-7}$		N/A		$2.94 \cdot 10^{-6}$	

MLR: Multiple Linear Regression, NN: Neural Networks

Table 4.3: R^2 of multiple linear regression and neural network models

R^2	Train		Validation		Testing	
	MLR	NN	MLR	NN	MLR	NN
Mean	$9.96 \cdot 10^{-1}$	$9.99 \cdot 10^{-1}$	N/A	$9.99 \cdot 10^{-1}$	$9.96 \cdot 10^{-1}$	$9.99 \cdot 10^{-5}$
Std. dev	$2.05 \cdot 10^{-4}$	$1.42 \cdot 10^{-4}$	N/A	$2.45 \cdot 10^{-4}$	$4.84 \cdot 10^{-4}$	$3.95 \cdot 10^{-4}$
P-value	$6.72 \cdot 10^{-7}$		N/A		$4.44 \cdot 10^{-6}$	

MLR: Multiple Linear Regression, NN: Neural Networks

Table 4.4: MSE of multiple linear regression and neural network models (after feature selection)

MSE	Train		Validation		Testing	
	MLR	NN	MLR	NN	MLR	NN
Mean	$7.52 \cdot 10^{-5}$	$3.33 \cdot 10^{-4}$	N/A	$3.65 \cdot 10^{-4}$	$1.30 \cdot 10^{-3}$	$3.56 \cdot 10^{-4}$
Std. dev	$9.68 \cdot 10^{-7}$	$2.84 \cdot 10^{-5}$	N/A	$7.37 \cdot 10^{-5}$	$2.93 \cdot 10^{-5}$	$2.48 \cdot 10^{-5}$
P-value	$1.86 \cdot 10^{-7}$		N/A		$2.89 \cdot 10^{-7}$	

MLR: Multiple Linear Regression, NN: Neural Networks

Table 4.5: R^2 of multiple linear regression and neural network (after feature selection)

R^2	Train		Validation		Testing	
	MLR	NN	MLR	NN	MLR	NN
Mean	$9.74 \cdot 10^{-1}$	$9.92 \cdot 10^{-1}$	N/A	$9.93 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	$9.92 \cdot 10^{-1}$
Std. dev	$3.10 \cdot 10^{-4}$	$1.56 \cdot 10^{-1}$	N/A	$2.45 \cdot 10^{-3}$	$7.67 \cdot 10^{-4}$	$4.89 \cdot 10^{-4}$
P-value	$2.21 \cdot 10^{-7}$		N/A		$3.83 \cdot 10^{-7}$	

MLR: Multiple Linear Regression, NN: Neural Networks

4.1.1 Acquired Genetic Programming (GP) Models

For the sake of brevity, only the best mathematical solutions are presented in this section. It is noteworthy that rpm is the normalized value [0 1]. The fitness

Table 4.6: Comparison on various models and solutions generated

MSE						
Config.	GP* (BF)	GP* (EF)	GP+ (BF)	GP+ (EF)	ANN	MLR
Mean	1.09×10^{24}	3.35×10^{29}	2.67×10^{29}	2.81×10^{30}	4.04×10^{-5}	1.83×10^{-4}
Median	5.01×10^{-3}	4.83×10^{18}	5.74×10^{26}	4.08×10^{27}	4.81×10^{-5}	1.79×10^{-4}
Best	1.47×10^{-3}	4.70×10^{-4}	5.75×10^{-4}	4.53×10^{-3}	1.50×10^{-5}	1.43×10^{-4}
R^2						
Config.	GP* (BF)	GP* (EF)	GP+ (BF)	GP+ (EF)	ANN	MLR
Mean	6.58×10^{-1}	4.15×10^{-1}	3.46×10^{-1}	1.67×10^{-1}	9.99×10^{-1}	9.96×10^{-1}
Median	8.29×10^{-1}	-4.51×10^{-5}	-3.19×10^{-3}	-8.07×10^{-4}	9.90×10^{-2}	9.99×10^{-1}
Best	9.08×10^{-1}	9.91×10^{-1}	9.88×10^{-1}	8.97×10^{-1}	9.99×10^{-1}	9.97×10^{-1}

* max node = 5; + max node = 10; BF = Basic Function; EF = Extended Function

are from the validation data. Only solutions with $R^2 \geq 0.8$ from validation are selected for discussion. The fitness are shown in Table 4.7. Table 4.6 shows overall comparison with the equivalent MLR and ANN models. From the tables, the results show that the variance of the models are relatively high as compared to their MLR/ANN counterparts shown in Table 4.3 and 4.2. This shows that the solution space for GP is more complex and multimodal as compared to their ANN /MLR counterpart. Multimodal refers to an optimization scenario in which there are various "local minima" that would cause the optimization algorithm to get stuck in a non global minima. ANN uses backpropagation algorithm which is a deterministic approach. Nevertheless, the fitness search space is less multimodal in the situation of ANN as can be seen by the smaller variance of fitness within the population of generated solutions. From Table 4.7, it can be observed that the best model acquired from GP is as expressed in Equation 4.3.

Five constants were supplied which are $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Due to the computation considerations, only five constants were evaluated and only node depth of 10 was considered. Note that higher maximum node depth can be considered for future evaluations. In spite of these limitations, a solution was generated and applied for further implementation in sensor abnormality detection. The best candidate solutions from genetic programming was given by

$$rpm = x_5 - (0.9) \cdot (x_{22}) \cdot (x_{29})^2 \cdot (x_5)^2 (\dot{x_8})^2 \cdot \frac{sign{x_5}^{x_9}}{(x_{30})^{x_5}} + (x_8)^2 \tag{4.5}$$

$$rpm = x_{17} \cdot \frac{x_{25}}{\sqrt{(x_{37})}} + \sqrt{(x_{18})}(x_{17}) \tag{4.6}$$

$$rpm = -\left(x_{28} \times x_{42}^{x_{30}} \times x_{30}^2 \times x_7^{0.5}\right)^2 + x_7^{0.7}, \tag{4.3}$$

where x_i is the i-th feature or sensor of the compressor unit. The features and the corresponding description of the sensors can be found in Appendix A. This model yielded R^2 of 0.991 and MSE of 4.7×10^{-4}.

In order to better understand the models developed, keep in mind that shaft load is proportional to shaft RPM, and it also corresponds strongly with bearing temperatures. Rotor radial loads are proportionate to shaft RPM, and so directly effect the shaft's radial displacement. The square of rotor RPM is exactly proportional to any change in the pressure differential across the compressor (discharge suction). As a result, it has a direct influence on the rotor's axial movement (active or inactive thrust direction).

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \tag{4.4}$$

$$rpm = x_{22} \cdot x_{44} + x_7 \tag{4.16}$$

$$rpm = {x_1}^4 \cdot {x_9}^2 + x_5 + {x_6}^2 \cdot x_9 \tag{4.17}$$

$$rpm = x_7 + {x_{14}}^2 \cdot {x_{23}}^2 \cdot x_{38} \cdot \frac{x_{45}}{(0.5 \cdot x_{11})} \tag{4.18}$$

$$rpm = -x_7 \cdot 0.5 \cdot (x_{30})^{2^{x_{28}}} \cdot {x_{42}}^{2^{x_{30}}} + x_7^{0.7} \tag{4.7}$$

Table 4.7: Validation on selected models

Model	MSE	R^2
4.5	4.38 E-3	0.9021
4.6	1.40 E-3	0.9683
4.7	4.70 E-4	0.9900
4.8	4.79 E-4	0.8394
4.9	4.73 E-4	0.8409
4.10	3.634 E-4	0.8902
4.11	4.737 E-4	0.8988
4.12	5.203 E-4	0.8219
4.13	5.090 E-4	0.824
4.14	1.472 E-4	0.967
4.15	4.296 E-4	0.8234
4.16	5.15 E-4	0.8234
4.17	4.290 E-4	0.9025
4.18	4.975E-4	0.8333
4.19	7.411 E-4	0.8150
4.20	5.05 E-4	0.8286
4.21	4.24 E-4	0.870
4.22	4.24 E-4	0.9053
4.3	4.77 E-4	0.9910

$$rpm = \frac{x_7}{sign(x_{43})} - (x_{40}) * \cdot (x_3)^2 + (x_{23})^2 + 2 \cdot (x_{42})^2 + (x_{43})^2 \tag{4.8}$$

$$rpm = x_{31}^{x_7} (\dot{x_{28}})^2 + x_7 - (x_{40})^2 \cdot sign(x_7) - (x_{40})^2 + (x_{43})^2 \tag{4.9}$$

$$rpm = x_{27} + \sqrt{(}x_{29}) \tag{4.10}$$

$$rpm = x_{25} \cdot x_{29} \cdot x_{32} \cdot \frac{{x_8}^2}{(x_{28} \cdot x_4)} + x_5 \tag{4.11}$$

$$rpm = {x_{38}}^2 \cdot x_4 \cdot {x_{45}}^3 + x_7 \tag{4.12}$$

$$rpm = x_{26} \cdot {x_{38}}^2 + x_7 \tag{4.13}$$

$$rpm = 0.3 \cdot x_{29} + {x_1}^2 \cdot x_{11} \cdot x_{40} + x_1 \cdot x_{40} + x_{18} + x_{24} \tag{4.14}$$

$$rpm = {x_{23}}^3 + x_7 + x_1 8 \cdot {x_{38}}^2 \cdot \frac{{x_4}^2}{(0.7 \cdot x_{36})} \tag{4.15}$$

$$rpm = x_{15} \cdot x_{26} \cdot x_{38} \cdot x_{42} - x_{31} \cdot x_{33} \cdot \frac{x_{44}}{x_8} + x_{38} \cdot x_{40} + x_7 - x_{44} \cdot x_{45} \cdot \frac{x_6}{x_4} \quad (4.19)$$

$$rpm = x_{18} \cdot {x_{38}}^2 \cdot x_4 + x_7 \quad (4.20)$$

$$rpm = -0.3 \cdot x_1 \cdot x_{16} \cdot \frac{{x_{42}}^2}{{x_{30}}^2} - x_1 \cdot x_{43} \cdot x_6 + x_7 + x_{18} * x_{23} \cdot x_{30} \cdot x_{43} \cdot \frac{x_{45}}{x_{12}} \quad (4.21)$$

$$rpm = 0.9 \cdot x_{18} \cdot x_{33} \cdot x_{43} \cdot x_{45} + x_{10} \cdot x_6 - x_{12} \cdot x_5 + x_{14} \cdot x_{16} \cdot x_{38}^2 - x_{42} + x_5 \quad (4.22)$$

In view that GP develops new equations as regression models based on equation tree structure, there will unavoidably be numerous maximum /minimal solutions in the candidate solution space when applying the stochastic process described previously. This is known to be a high modality issue. Aside from the noticeable high modality, the population-based method to solution seeking would result in high computation.

As a preliminary step and for the sake of having a more holistic view on the effects of setting parameters on the targeted data application, the GP algorithm was tested on sample set using 70% - 30% training - testing split.

Previously for ANN training , the dataset split based on 70% (Training)-15% (Validation)-15% (Testing). Neural network training is based on 2 back propagation processes, involving the forward and backward propagation. However GP and MLR models are based on only 1 set of data which is training. Hence there is a difference in splitting the data set. There is no "apple to apple" comparison in splitting the data in ANN, GP and MLR. In any case, 70% (Training), 30% (Testing) would be considered a harder challenge as compared to 70% (Training),

15% (Validation), 15% (Testing) because the test independent data set will further legitimize the testing of the proposed algorithm. Therefore, we apply the harder test for the selected algorithm (GP) against the comparison algorithms.

Two types of function sets were applied as shown in Table 3.8, where $\hat{\alpha}$ and $\hat{\beta}$ are two inputs applied to the functions. The basic function set consists of four basic mathematical operations (addition, subtraction, multiplication, and division) while the extended function set consists of more complex mathematical operations, such as square and square root as well as absolute ($|\cdot|$) functions.

The fitness evaluated is MSE from independent test data. The large difference between the mean, median and best configuration demonstrates that there are substantial variations across configurations. The best configurations outperform the average configurations by a wide margin (refer to the difference between mean, median and best). This implies that in consecutive trials with varying generations, only the some acquire good fitness. This implies that the solution fitness surface has many local minima regions thereby causing solution to be stuck in local minima points. Local minima refers to a situation in optimization algorithms, where the optimization process finds a minimum value (a local optimum) that is not the global minimum, which is the true minimum value over the entire search space. This occurs when the optimization process is trapped in a region where the objective function has a low value, but this value is not the absolute lowest over the entire search space. This can result in sub-optimal solutions, and the optimization process can get stuck in a local minimum, unable to escape and find the global minimum.

From Table 4.8, there is no obvious difference in applying either the full set of math operators (extended function) or the basic function set. Observing that there is not much effect in changing the maximum mathematical length with regards to the fitness. This implies that it is not much of a link between fitness and maximum program length. It is agreed that the maximum node lengths of 5 and 10 can be investigated further. Note that the function nodes that the tree

is permitted to develop to optimize fitness are referred to as the program length. The absolute error (the difference between the modelled and observed values) is used to define the performance fitness.

Apart from the expected high variances in performance, there was no obvious pattern in determining the node depth. 10 trials were performed for every configuration. Among the 10 trials, at least 1 trial achieved MSE of 10^{-3} on the test data set (due to the random computation seed). Nevertheless, slightly better results were acquired by observing the median of the solution fitness. Therefore, this may not be highly substantial given that the mean performances were high in both configurations. Moreover, it can generally be observed that there was always an optimal solution generated in spite high average error recorded over a number of 10 trials. Several trials were required due to the nature of the problem being stuck in a local minima as the optimization progressed. Also, it can be seen that there was not substantial statistical significance to eliminate or to conclude any optimal node depth for further evaluation.

Table 4.8: Preliminary result to determine parameters

Extended Function Set			
Max. Node Depth	**Mean**	**Median**	**Best**
4	1.26×10^{27}	5.25×10^{-3}	5.13×10^{-3}
5	1.41×10^{28}	5.24×10^{-2}	2.38×10^{-3}
6	1.97×10^{30}	1.61×10^{27}	4.85×10^{-3}
7	5.39×10^{26}	8.28×10^{24}	3.63×10^{-2}
8	3.95×10^{30}	4.15×10^{24}	1.23×10^{-3}
9	3.99×10^{-3}	3.20×10^{-3}	2.71×10^{-3}
10	8.24×10^{25}	7.94×10^{-3}	4.87×10^{-3}
Basic Function Set			
Max. Node Depth	**Mean**	**Median**	**Best**
4	5.26×10^{-3}	4.77×10^{-3}	2.12×10^{-3}
5	5.66×10^{-1}	5.63×10^{-3}	3.20×10^{-3}
6	1.27×10^{23}	5.11×10^{-3}	1.26×10^{-3}
7	3.60×10^{28}	5.27×10^{-3}	1.32×10^{-3}
8	1.40×10^{-2}	4.62×10^{-3}	2.04×10^{-3}
9	1.33×10^{26}	5.87×10^{-3}	4.87×10^{-3}
10	4.95×10^{-3}	5.41×10^{-3}	2.18×10^{-3}

4.1.2 Evaluation on Selected Model (Equation 4.3)

In the previous subsections, GP was applied to train and subsequently, the best models were selected for further evaluation on the 20% independent dataset as explained earlier. As an evaluation to the selected best model based on validation data in the training phase, the corresponding regression plot (Figure 4.1) and corresponding histogram (Figure 4.2) shows the fitness of the acquired best model (refer to Equation 4.3). Regression plots are common ways to evaluate visually "goodness -of-fit". The x axis represents the output of the prediction model while the y axis represents the actual models based on the test data actual output. It is observed that values greater than 0.6 are quite linear with the expected output. However, regions less than 0.5 have lower correlation. This was investigated and the cause can be attributed to the available data. 95% of the training data are ranged at [0.6 1.00] (refer to the highlighted region). Further observation from Figure 4.2 also confirm validation of this statement by observing the density of the test data in the figure. The training phase tend to prioritize model development based on the minimization of the training dataset error. Nevertheless, this model (Equation 4.3) manage to score a "goodness of fit" of 0.9044, based on R^2 evaluation.

The data is independent test data reserved for test evaluation and has not been "seen" in the GP training phase. This model yields $R^2 = 0.9044$ which comparable to the R^2 evaluation in the validation phase.

4.1.3 Histogram of Residuals for Sensor Abnormality Detection

Figure 4.3- Figure 4.7 show the residuals between model output expressed in Equation 4.3 vs the augmented baseline data. Baseline data is the actual data collected as described in previous section. The residuals against the various augmented data is discussed in the subsequent chapters. Ideally, 100% of the residuals

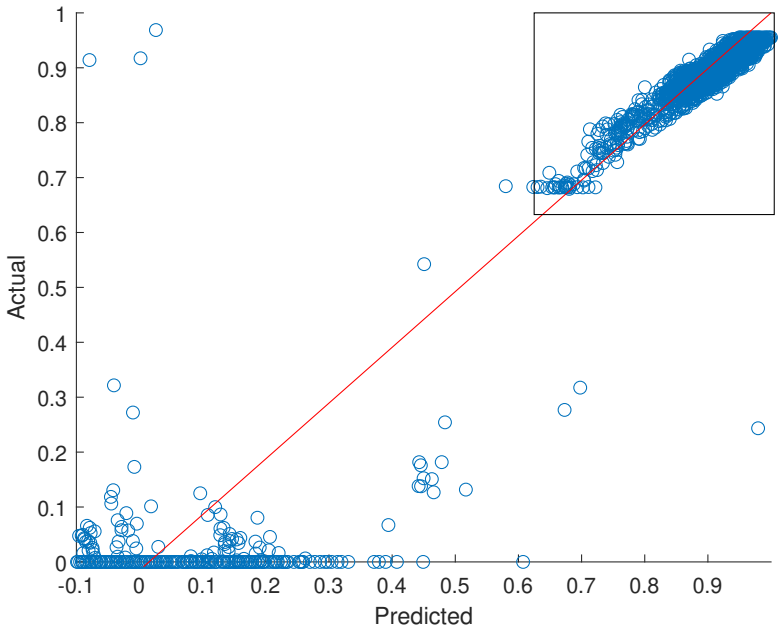

Figure 4.1: Regression plot predicted vs actual (independent test data)

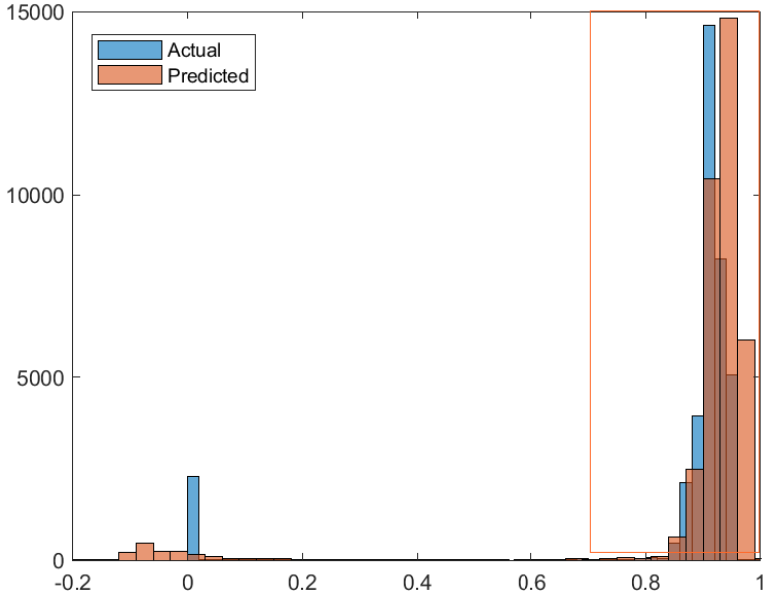

Figure 4.2: Histogram of expected values and predicted values

should fall into the center bin of 0 as shown in Figure 4.3. The slight deviation shows sheds light on the amount of inaccuracy of the acquired regression model. Nevertheless, the deviation from ideal case is practical as no machine learning model is expected to output 100% accuracy. The model expressed in Equation 4.3, which uses only four features, i.e., x_7, x_{42}, x_{30}, and x_{28}, can be considered as an extremely compact model compared to the MLR and ANN models, which utilize all the 46 features. Recall that the Occam's Razor principle in modelling which states that a better descriptor refers to a smaller and compact model. Degradation faults can be identified by the characteristic wider spread of the histogram as shown in Figure 4.6. Constant faults and bias drift fault would have multiple "peaks" in the histogram. Lastly, noise fault can be identified by a shifted residual mean as shown by Figure 4.7.

Theoretically, the abnormal data based on augmentation scheme proposed by Tsai et al. (2019) can be further classified. However, noted that a limited configuration was presented for degradation and bias drift. In fact, there could be other error configurations that can be introduced. This causes further classification of complexity and may need to include a machine learning model to achieve this purpose. Nevertheless, abnormality can be identified using algorithm without the deployment of any complex model. The sensor is considered normal if the majority of the obtained counts of residuals fall between -0.05 and +0.05. By calculating the percentage of residuals in this bin, a probabilistic model may be developed.

4.1.4 Concluding Remarks on Phase 1 Implementation

In this section, some discussion will be elaborated on the results acquired. In the previous section, the MLR, ANN and tree based genetic programming (GP) were applied first to model plant behavior and subsequently implementation in sensor fault detection. ANN and MLR are competitive solution in plant modelling approaches with MLR having more linear approximations as compared to

ANN. However, linear regressions posses better interpretability as compared to the embedded "black box" nature of ANN. On the other hand, GP posses "white box" interpretability and non-linear mapping capabilities. Summary of the characteristic of the 3 regression algorithms is expressed in Table 4.9 based on the strengths and weaknesses. The ranking shown with 1 being the highest and 3 being the least. Both MLR and GP are highly interpretable given that they are essentially equations. GP have higher transparency as they are essentially equation models generated. Non linear mapping refers to the capability of algorithm to map the output to the inputs in non-linear correlations. This can be shown from the R^2 values shown. A higher R^2 shows the capability of models to map inputs to outputs. Referring back to Figure 3.2, these features score high fisher correlation values (≥ 0.9). The advantages and justification of selecting GP was explained in earlier chapters.

Table 4.9: Characteristic of approaches for modelling

	ANN	MLR	GP
Interpretability	3	2	1
Non -Linear mapping	1	3	2

The implementation of MLR applies weightage to the individual features (x_1 to x_{46}). This provides opportunity to rank the features according to the magnitude of the weights. It is common for stakeholders to apply the weights obtained as features ranking procedure. The most important features for RPM shaft prediction includes x_{44}, x_{41}, x_{13}, x_{29}, x_{14} based on the absolute value of the weights. Table 4.1 shows further details in the feature group. Based on this table, the RPM shaft is mostly correlated to dry gas seal conditions, compressor and gear box monitoring. However, this is only the case when linear MLR is considered. In the non linear model approach, there may be other features selected when non linear models are considered. The case in point is the model selected in GP. In GP, non linear relationships may emerge from the inclusion of non linear functions such as square or root square. As such, the best selected models as

Table 4.10: Most contributive sensor group based on MLR Regression

Feature	Monitoring sensor group
x_{44}	Dry gas seal monitoring sensor
x_{41}	Dry gas seal monitoring sensor
x_{13}	Compressor mechanical performance monitoring sensor
x_{29}	Gear box mechanical monitoring sensor
x_{14}	Compressor mechanical performance monitoring sensor

Table 4.11: Features details as expressed in Equation 4.3

Feature	Monitoring sensor group
x_{42}	Dry gas seal monitoring sensor
x_{30}	Gear box mechanical monitoring sensor
x_{7}	Compressor thermodynamic performance monitoring
x_{28}	Gear box mechanical monitoring sensor

expressed in Equation 4.3 shows different selection of features. Without alluding to any feature ranking, Table 4.11 shows the details of the features selected for the best models. Conclusion that can be made from this analysis is that models may apply varying features as a results on non linear functions being involved. In this case, feature training using MLR weights may not reveal much correlation to selected features in non linear models.

Despite implementation with varying settings in GP, it is not conclusive that there is any configuration can produce better mathematical models. The results further solidify the notion that the fitness surface is indeed highly multimodal. Nevertheless, from the results acquired, the best model from the population does produce compatible results with the ANN/MLR models counterparts.

4.2 Phase 2 Evaluation: Identification of Faults from Histogram

Recalling the error simulation methodology as proposed by Tsai et al. (2019), the findings are listed in this section below. The histogram presents an obvi-

Table 4.12: Preliminary result to determine parameters

Extended Function Set			
Max. Node Depth	**Mean**	**Median**	**Best**
4	1.26×10^{27}	5.25×10^{-3}	5.13×10^{-3}
5	1.41×10^{28}	5.24×10^{-2}	2.38×10^{-3}
6	1.97×10^{30}	1.61×10^{27}	4.85×10^{-3}
7	5.39×10^{26}	8.28×10^{24}	3.63×10^{-2}
8	3.95×10^{30}	4.15×10^{24}	1.23×10^{-3}
9	3.99×10^{-3}	3.20×10^{-3}	2.71×10^{-3}
10	8.24×10^{25}	7.94×10^{-3}	4.87×10^{-3}
Basic Function Set			
Max. Node Depth	**Mean**	**Median**	**Best**
4	5.26×10^{-3}	4.77×10^{-3}	2.12×10^{-3}
5	5.66×10^{-1}	5.63×10^{-3}	3.20×10^{-3}
6	1.27×10^{23}	5.11×10^{-3}	1.26×10^{-3}
7	3.60×10^{28}	5.27×10^{-3}	1.32×10^{-3}
8	1.40×10^{-2}	4.62×10^{-3}	2.04×10^{-3}
9	1.33×10^{26}	5.87×10^{-3}	4.87×10^{-3}
10	4.95×10^{-3}	5.41×10^{-3}	2.18×10^{-3}

ous approach to identify the faults from non-fault sensor readings. Accumulated over time, most of the residual readings should fall within the histogram bin at mean=0. The slight deviation is due to model imperfection. Hence, formally identification can be obtained with any deviation from this. In the case of constant faults, histogram of residuals have multiple peaks as demonstrated in Figure 4.4. Similarly, bias drift faults generate similar type of histogram profile as demonstrated in Figure 4.5. The histogram of residuals apply "flatter" to the distribution to other bin values. However, it can be seen that there are substantial values at the 0 bin. The final histogram as shown in 4.7 shows the noise fault scenario. They can be formally distinguished from a single peak observed from histogram but not at mean value at 0. Observed from the equation that it is merely shifting the actual value to higher value. Therefore, the residual is expected to resemble non faulted residuals with shifted mean. By following the description above, these faults can be categorized into various faults just by observing the residuals (histogram of readings deviation).The summary of these observations correlating

to the type of faults can be seen from Table 4.13. By observing the characteristics of the fault from the histogram of residuals, stakeholders may be able to identify the faults. Despite the abnormalities being simulated, the nature of the faults (based on the description) is believed to be consistent with the observation listed in the table with different degradation levels based on the description, the amount of the histogram being "flatten" may vary. However, the general characteristic of the histogram correlating to the faults may be applied.

- Constant faults : pick successive samples, and replace them with constant value
- Bias drift faults: pick successive samples and replace each value v_i with $v_i + constantvalue$
- Degradation faults : pick successive samples start from i and replace value v_j with $v_j + degradation\_rate * (j - i)$

  Noise faults : select a sensor measurement v_i and multiply a factor f determines the intensity of the noise fault a noise data $v_i' = v_i * f$ where $f = 1.2, 1.5, 2, 5, 10$

Table 4.13 shows the summary of methods to identify fault types. The table is acquired by observing the histogram of residuals as shown in Tables 4.4-4.7. The table shows that differentiating between normal and abnormal sensor reading is relatively trivial. In the normal conditions, there should be normal curve with mean = 0 and should not be multiple peaks in the histogram. Constant faults and bias drift have almost similar profile thereby causing it difficult to distinguish between the both. Degradation faults causes the histogram to have a flatten profile thereby can be identified easily. Noise faults are due to the presence of noise thereby shifting the residual of mean from 0 Value.

This research only looked at one sensor anomaly. As the measurement is dependent on the correlation model with other sensors, it may be slightly influenced

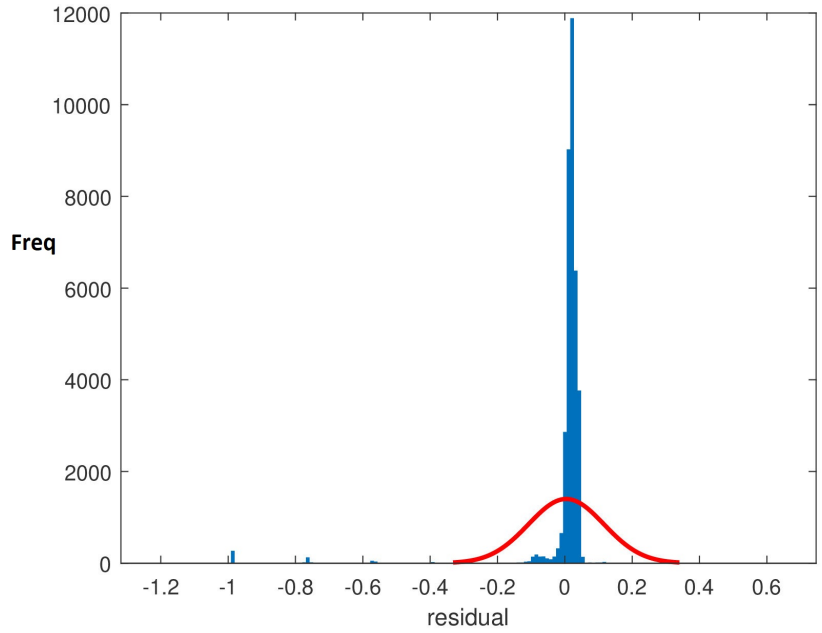

Figure 4.3: GP model vs baseline data

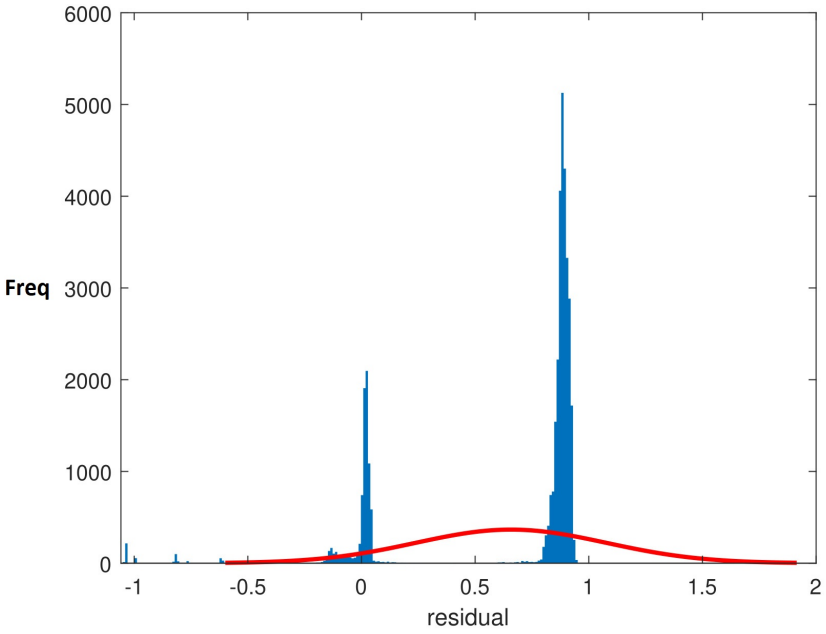

Figure 4.4: GP model vs constant faults

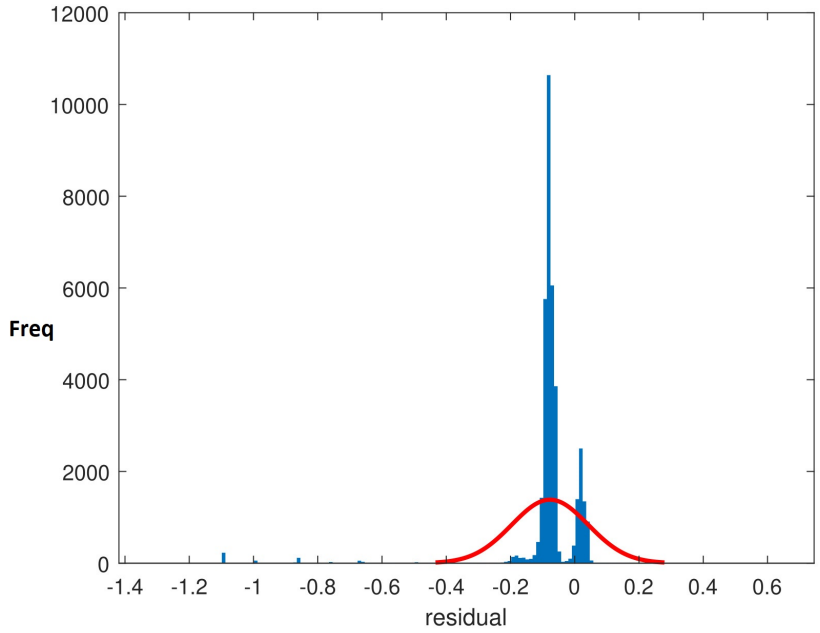

Figure 4.5: GP model vs bias drift fault

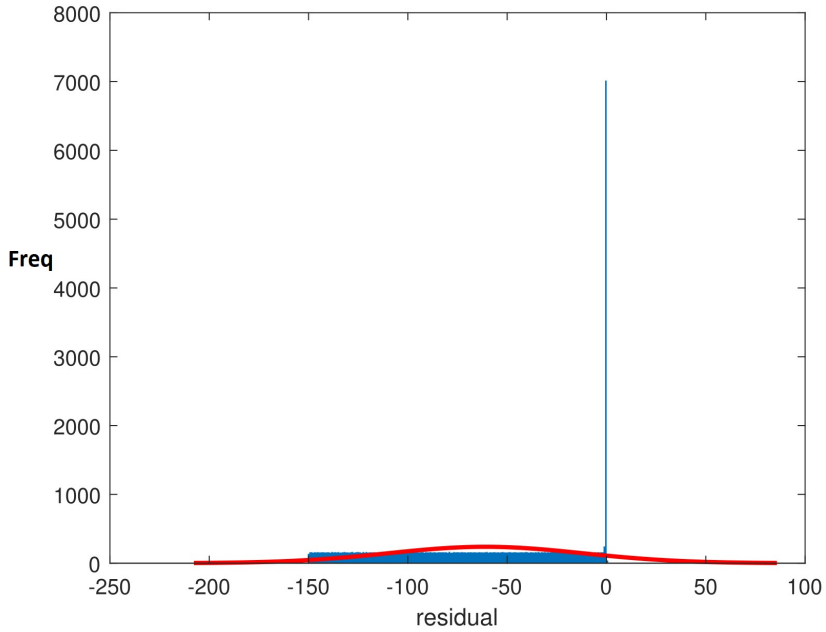

Figure 4.6: GP model vs degradation fault

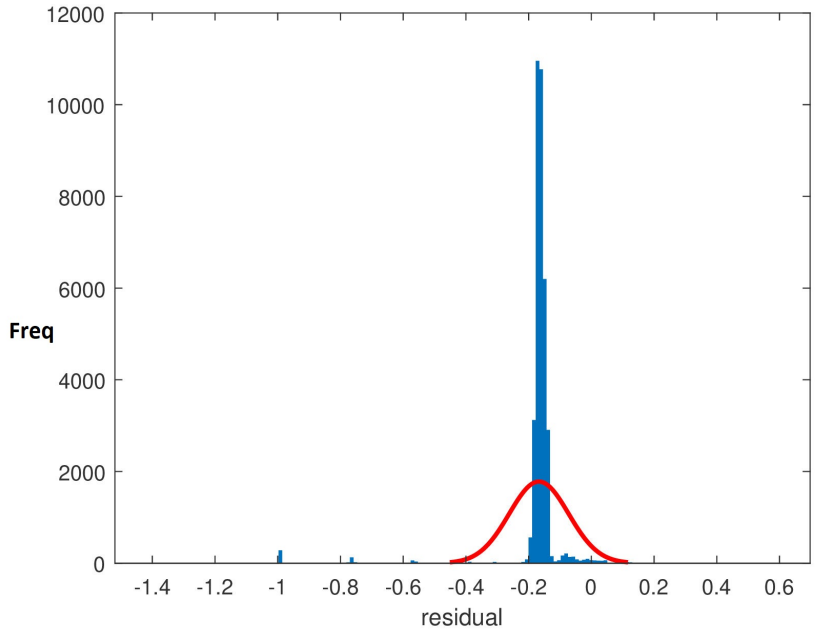

Figure 4.7: GP model vs noise fault

Table 4.13: Identification of faults from histogram

	Observations from residual histograms	Multiple peaks	shifted mean	"flatten" Histogram
1	Constant faults	yes	no	no
2	Bias drift	yes	no	no
3	degradation faults	no	no	yes
4	noise faults	no	yes	no
5	Normal conditions	no	no	no

if numerous sensors are abnormal. Given the model's small size (only four features), it has the benefit of not being as influenced as an ANN or MLR model that uses all 46 features. In other words, a model with a smaller feature set would be more stable.

4.3 Benchmarking and Comparison with Relevant Research Work

Comparison and bench marking will be discussed in this section. This section will evaluate and compare research works that deals with failure modelling. This will be approximate comparison since this is a specific implementation on shaft sensors in compressor. It is noteworthy that this thesis was a top-down study to predict abnormality using long term data. There is no exact bench marking as this is an applied an investigation as our current approach is specifically for compressor and niche research domain. Hence, some relevant/related comparison will be discussed in this section. As stated, the fault data is simulated as it is not practical to acquire real fault data. The real data were sectioned into 10 partitions and applied to generate fault data using the described approach. This approach is commonly known as cross validation. 8 sections were used for training and 2 partitions was used to simulate error fault. This was repeated 20 times. The recognition was implemented to differentiate normal and non-normal distribution. By analyzing residuals that fall into the bin 0.1 with mean 0 and a threshold of 0.95 distribution in this range, an accuracy of 100% was achieved. However, this is based on simulated faults.

Authors in Khalastchi et al. (2013) outline a technique employing a structural model for real-time identification and resolution of sensor malfunctions. Authors validate this approach by applying it to Mobile Robotic system, a lab robot, and FlightGear, a flight simulator, to experimentally assess its efficacy. The research 0.9% accuracy rate. In Guo et al. (2022), researchers implemented moving average

analysis for energy system sensors. Using quartile analysis and statistical analysis, a 83.96% recognition rate was acquired. In Tsai et al. (2019), authors attempted to recognize multiple sensor degradation. In earlier section, similar simulation of sensor fault were implemented in this thesis. A 96.00% accuracy rate was recorded using this approach on simulated fault data.

Comparing with Khalastchi et al. (2013), Guo et al. (2022) and Tsai et al. (2019), our current approach of using a mathematical model to model expected output and comparing the residuals seems to suggest that approach is relevant and comparable to other sensor diagnosis approach albeit it is noteworthy that application wise, the environment could be different.

4.4 Discussion

In this chapter, the results are presented in 2 phases. It needs to be reiterated that low frequency data (10 minutes interval) is implemented for this research project. Higher frequency data would be computationally expensive for data storage. The first phase deals with building a GP model that best describes prediction of RPM sensor values based on other sensor values. GP approaches enable the generation of candidate mathematical models which exhibits "white box" characteristic rather than embedded "black box" model. "Black box" models are defined as machine learning models that cannot be explained by stakeholders despite able to perform well. It was found that GP models generated exhibit high performance variances. The best model was selected and tested with independent data set yields performance of $R^2 = 0.9044$. Subsequently, this model was applied for a detection of sensor (Phase 2). We consider only a single RPM sensor at the moment. Histogram of residuals was calculated that reveals the abnormality. The normal operating mode would yield a histogram of residuals centered at 0. As stated earlier in the assumption is that the faulty sensor reading at this stage are simulated. Nevertheless, the results section does indicate the working concept proposed and fulfills the research objectives stated.

Chapter 5

Conclusion and Future Works

In the previous chapters, various aspects of the research project was deliberated and they intend to solve 2 research gaps. The first research gap deals with the inability of current strategies to make use of existing algorithms to detect abnormality from "data logs". The complexity lies in the frequency update rate of 10 minutes that makes usage of dynamic analysis (such as LSTM, RNN). As such, a solution was proposed to apply static approach to predict the residual (between observed and predicted). The second research gap lies in the perception that persist in the usage of machine learning models specifically those that appears to be "black box". For example, consider the Neural network which has been widely deployed for data driven models. The scholarly approach provides some investigation on enhancing the optimality and tuning of neural network. Despite ongoing for almost 2 decades, this approach still does not offer much improvement in debunking the interpretability of the working models, this may lead to doubt and resistance of usage. In order to cater to this specific research gap, a math based model was preferred over the black box models. The discussion on the applicability of non interpretable machine learning models is a well known among industrial application practitioners as deliberated in Rudin (2019). The deliberation presented in Rudin (2019) offers explanation on the difference between explaining black boxes and using inherently interpretable models, lists

several important arguments against using explainable black boxes in high-stakes decisions, discusses obstacles to interpretable machine learning, and offers a number of examples where black box models might be replaced by interpretable ones in the fields of criminal justice, healthcare, and computer vision. As such, in this proposed solution, a Genetic programming approach was implemented to obtain a generated mathematical models which is interpretable. The solution offers compactness in solution in line with the 'Occham's Razor' principle. This principle states that when presented with 2 models of system, the simper and more compact model should be adopted. Therefore, even in the presence of equivalent Neural network, a mathematical model should be considered. This is considered as a strong justification to the approach.

5.1 Achievement of Objectives

Recall in Chapter 1, 2 objectives were proposed:

1) Implement an approach to detect the faults in sensors by only using low frequency data logs.

2) Implement and modelling as much as possible using "White box" model approach (explainable AI approach).

In achieving objective 1, the proposed algorithm implements a short to long term data evaluation and evaluates the abnormality of the sensor. The residual histogram gives indication on the health of the sensor. The residual was attained by evaluating the difference between the observed vs the normal. The abnormal sensor data was simulated based on the methodology proposed by Tsai et al. (2019), It is assumed that the data for training is sensor normal function as no abnormality was reported in the span of data collection.

As for objective 2, the prediction model utilizes GP to generate a mathematical equation contrary to conventional "black box" models. The best equation

model was discussed.

5.2 Future Investigations

Despite the achievement of proving the concept of sensor fault using independent portion of data log (20%) and subsequently model agreement (evaluated using $R^2 = 0.9044$), there are still several improvements that can be considered for this research. Firstly, one of the primary improvements to be considered is related to the dataset itself. Although it is universally accepted that it is impossible to have perfectly balanced data set. In the current data, 95% of data output (RPM sensor) is ≥ 0.6. This is unavoidable as there is only very few cases of pressure drop (abnormality). Most of the time, the shaft RPM output should be at 8000 rpm. In this problem, there is only a few solutions available and they involve synthetic generation of data. One such solution is the SMOTE approach (synthetic Minority Oversampling Technique). The approach applies a form of inverse KNN (KTH Nearest Neighbor) to infer possible synthetic datasets. As the name implies, implementation of this method is not without reproach. The data are obviously synthetic and one could argue if such approach is feasible for such an important facility. The alternative to this is massive collection of data and manually selected such that the histogram of instances are almost constant. This would be not a economical approach given the research time allocated for this research. Hence, this could be a dilemma than can be resolved in the future.

Another important aspect that was considered is the simulation of sensor fault data. It was explained earlier, the sensor fault was simulated based on the methodology proposed by (Tsai et al., 2019). It is noteworthy that this is still a simulated approach since installation of faulty sensor would not be feasible. However, there is still room in the future research to investigate on the various and more realistic approach in simulating with sensor faults.

As a future path of this study, the residuals may enable further categorization into the various forms of sensor anomaly. However, given the varying degrees

of mistake, this work is difficult. It is sufficient to infer at this point that the suggested strategy for detecting anomaly utilizing enriched data set is effective. Most importantly, the regression mathematical model developed using GP is small and ideal for usage in Programmable Logic Controllers (PLCs) or offline periodic detection.

5.3 Concluding Remarks

Compressors play an important role in various plants systems especially in the context of oil and gas plants. While most the research are focused on detecting faults of compressors, another aspect that are often overlook are the health status of the compressors. This is more critical in compressor system as the sensors are operating in harsh environment. While the principles discussed in this thesis can be generalized to sensors, the discussion have been centered on compressors system due to the stakes of the sponsor for this study. With this established, further considerations have been made on the types and methods of detection. From the literature review, methods pertaining to detection of faults in equipment or sensor systems normally falls under 2 major category: dynamic or static. The first category focuses on the pattern of changes in the readings. This is certainly more complex without guaranteeing better results. The second category doesn't need to capture the changes of the data with respect to time. In this aspect, the 2nd method is more feasible for most of the applications involving compressor. In this case, data were acquired using a 10 minute logging system thereby effecting rendering dynamic approach non -feasible approach.

In the previous chapters of the thesis, motivations leading up to the results and discussion sections were presented. Implementing any algorithms that are "black box" in nature high risk and may not serve as a good "ambassador" to encourage adoption into the industry. By stating this as the primary motivations, what is desired is to have a mathematical model that represents the relationship between the sensor values.

As a concluding remark, the major objective of this thesis is to establish a method for generating mathematically based regression models that would effectively detect abnormalities utilizing data logs. It is worth mentioning that the fundamental benefit of using GP is the "white box" character of produced equations. This is impossible to do with the ANN. Furthermore, the gradient per permutation in GP is high so that the fitness surface becomes highly multimodal. As a result, comparing to ANN or MLR equivalents, it would need several trial runs with larger populations. Nevertheless, it can be stated that the GP produces the most "white box" model as compared to other approaches due to the mathematical model generation approach.

Bibliography

M. A. D. Alves, et al. (2021). 'A Modified Algorithm for Training and Optimize RBF Neural Networks Applied to Sensor Measurements Validation'. *IEEE Sensors Journal* **21**(17):18990–18999.

P. E. Bhaskaran, et al. (2020). 'Future prediction & estimation of faults occurrences in oil pipelines by using data clustering with time series forecasting'. *J. Loss Prev. Process Ind.* **66**:104203.

P. E. Bhaskaran, et al. (2021). 'IoT Based monitoring and control of fluid transportation using machine learning'. *Comput. Electr. Eng.* **89**:106899.

A. Bousdekis, et al. (2017). 'A Proactive Event-driven Decision Model for Joint Equipment Predictive Maintenance and Spare Parts Inventory Optimization'. *Procedia CIRP* **59**:184–189.

Y. S. Byun, et al. (2019). 'Sensor Fault Detection and Signal Restoration in Intelligent Vehicle'. *Math. Probl. Eng.* **19**(15):3306.

A. Cachada, et al. (2018). 'Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture'. In *Proc. 2018 IEEE 23rd Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, pp. 139–146.

M. Cavagliá, et al. (2020). 'Improving the background of gravitational-wave searches for core collapse supernovae: a machine learning approach.'. *Machine Learning: Science and Technology.* pp. 015005. 10.1088/2632–2153/ab527d.

J.-M. Cha, et al. (2017). 'Sensor drift detection in SNG plant using auto-associative kernel regression'. In *2017 IEEE International Systems Engineering Symposium (ISSE)*, pp. 1–4.

A. Gaddam, et al. (2020). 'Detecting Sensor Faults , Anomalies and Outliers in the Internet of Things : A Survey on the Challenges and Solutions'. *Sensors* **9**(3):1–15.

L. Galotto, et al. (2015). 'Data based tools for sensors continuous monitoring in industry applications'. pp. 600–605.

L. Galotto, et al. (2007). 'Sensor Compensation in Motor Drives using Kernel Regression'. pp. 229 – 234.

D. Goyal, et al. (2019). 'Non-contact sensor placement strategy for condition monitoring of rotating machine-elements'. *Eng. Sci. Technol. an Int. J.* **22**(2):489 – 501.

Y. Guo, et al. (2022). 'Sensor Fault Detection Combined Data Quality Optimization of Energy System for Energy Saving and Emission Reduction'. *Processes* **10**(2).

H. Hanachi, et al. (2017). 'Enhancement of prognostic models for short-term degradation of gas turbines'. In *Proc. 2017 IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, pp. 66–69.

P. Ibarguengoytia, et al. (2001). 'Real Time Intelligent Sensor Validation'. *Power Engineering Review, IEEE* **21**:63–64.

J. S. Rahhal and D. Abualnadi (2020). 'IOT Based Predictive Maintenance Using LSTM RNN Estimator'. In *Proc. 2020 Int. Conf. Elect. Commun. Comput. Eng. (ICECCE)*, pp. 1–5.

R. Jegadeeshwaran & V. Sugumaran (2015). 'Brake fault diagnosis using Clonal Selection Classification Algorithm (CSCA) - A statistical learning approach'. *Eng. Sci. Technol. an Int. J.* **18**(1):14 – 23.

H. Jiang, et al. (2019). 'Performance Prediction of the Centrifugal Compressor Based on a Limited Number of Sample Data'. *Math. Probl. Eng.* **2019**:13.

E. S. Johansen, et al. (2021). 'Cost Effective, Digital, Fail-Safe Production Tree and Wellhead Actuator System'. In *OTC Offshore Technology Conference*, vol. Day 2 Tue, August 17, 2021.

T. M. Jose (2018). 'A Novel Sensor Based Approach to Predictive Maintenance of Machines by Leveraging Heterogeneous Computing'. In *Proc. 2018 IEEE SENSORS*, pp. 1–4.

S. Kai (2017). 'Tensor flow enabled Genetic Programming'. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) Companion, ACM 2017* pp. 1872–1879.

H. Kang (2013). 'The prevention and handling of the missing data'. *Korean Journal of Anesthesiology* **64**:402– 406.

K. Kapitanova, et al. (2012). 'Being SMART about failures: assessing repairs in SMART homes'. pp. 51–60.

E. Khalastchi, et al. (2013). 'Sensor fault detection and diagnosis for autonomous systems'.

D. Li, et al. (2020). 'Recent advances in sensor fault diagnosis: A review'. *Sens. Actuators A Phys* **309**.

Y. Li, et al. (2017). 'An event-based analysis of condition-based maintenance decision-making in multistage production systems'. *International Journal of Production Research* **55**(16):4753–4764.

O. Loyola-González (2019). 'Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View'. *IEEE Access* **7**:154096–154113.

A. Naskos, et al. (2020). 'Event-Based Predictive Maintenance on Top of Sensor Data in a Real Industry 4.0 Case Study'. In P. Cellier & K. Driessens (eds.), *Machine Learning and Knowledge Discovery in Databases*, vol. 1168. Springer, Cham.

E. Priyanka, et al. (2020). 'Integrating IoT with LQR-PID controller for online surveillance and control of flow and pressure in fluid transportation system'. *J. Ind. Inf. Integr.* **17**:100127.

E. Priyanka & S. Thangavel (2020). 'Decision Making Based on Machine Learning Algorithm for Identifying Failure Rates in the Oil Transportation Pipeline'. In *Proc. 2020 Int. Conf. Decision Aid Sci. Appl. (DASA)*, pp. 914–919.

E. Priyanka, et al. (2021a). 'Review analysis on cloud computing based smart grid technology in the oil pipeline sensor network system'. *Petroleum Res.* **6**(1):77–90.

E. Priyanka, et al. (2021b). 'Miniaturized antenna design for communication establishment of peer-to-peer communication in the oil pipelines'. *Petroleum Res.* .

R. Kurz and K. Brun (2012). 'Upstream and Midstream Compression Applications - Part 1: Applications'. In *Proc. ASME Turbo Expo 2012*, pp. 1–11.

N. S. Rosli, et al. (2019). 'Predictive Maintenance of Air Booster Compressor (ABC) Motor Failure using Artificial Neural Network trained by Particle Swarm Optimization'. In *Proc. 2019 IEEE Student Conf. Res. Dev. (SCOReD)*, pp. 11–16.

N. S. B. Rosli, et al. (2018). 'Application of Principal Component Analysis vs. Multiple Linear Regression in Resolving Influential Factor Subject to Air Booster Compressor Motor Failure'. In *Proc. 2018 IEEE 4th Int. Symp. Robot. Manuf. Automat. (ROMA)*, pp. 1–5.

C. Rudin (2019). 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. *Nature Machine Intelligence* **1**:206–215.

N. Sairam & S. Mandal (2016). 'Thermocouple sensor fault detection using Auto-Associative Kernel Regression and Generalized Likelihood Ratio Test'. In *2016 International Conference on Computer, Electrical Communication Engineering (ICCECE)*, pp. 1–6.

N. Sakthivel, et al. (2014). 'Comparison of dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibration signals'. *Eng. Sci. Technol. an Int. J.* **17**(1):30 – 38.

B. Steurtewagen & D. Van den Poel (2019). 'Root Cause Analysis of Compressor Failure by Machine Learning'. In *Proc. 2019 Petroleum Chem. Ind. Conf. Eur. (PCIC EUROPE)*, pp. 1–5.

B. Sujeong & K. Duck-Young (2019). 'Abrupt variance and discernibility analyses of multi-sensor signals for fault pattern extraction'. *Computers & Industrial Engineering* **128**:999–1007.

S. Thangavel, et al. (2021). 'Dynamic Modeling and Control Analysis of Industrial Electromechanical Servo Positioning System Using Machine Learning Technique'. *J. Testing Eval.* **49**:2425–2440.

F. K. Tsai, et al. (2019). 'Sensor Abnormal Detection and Recovery Using Machine Learning for IoT Sensing Systems'. In *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 501–505.

Z. Wang, et al. (2017). 'Improved sensor fault detection, isolation, and mitigation using multiple observers approach'. *Syst. Sci. Control Eng.* **5**(1):70–96.

P. Wong (2022). 'Pauline compressor data corresponding to thesis'. `http:https://github.com/WKWONG123/paulinewongthesisdata`. [Online; accessed 19-July-2022].

P. Wong, et al. (2022). 'A minimalist approach for detecting sensor abnormality in oil and gas platforms'. *Petroleum Research* **7**(2):177–185.

X. Jia and Q. Cheng and Y. Hou (2018). 'Sensor Fault Detection Based on State Estimation Observer in Discrete Nonlinear Systems'. In *Proc. 2018 Int. Conf. Control Automat. Inf. Sci. (ICCAIS)*, pp. 542–547.

I. Yazar, et al. (2017). 'Comparison of various regression models for predicting compressor and turbine performance parameters'. *Energy* **140**(2):1398–1406.

Appendix A: List of Machine Sensors

- x_1: Compressor & Condensate Export Tilted Plate Coalescer V-270
- x_2: Compressor & Condensate Export Tilted Plate Coalescer V-280
- x_3: Compressor & Condensate Export Suction Scrubber Level
- x_4: Compressor Thermodynamic Performance Monitoring Suction Pressure ($barg$)
- x_5: Compressor Thermodynamic Performance Monitoring Discharge Pressure ($barg$)
- x_6: Compressor Thermodynamic Performance Monitoring Suction Temperature ($^{\circ}C$)
- x_7: Compressor Thermodynamic Performance Monitoring Discharge Temperature ($^{\circ}C$)
- x_8: Compressor Mechanical Performance Monitoring Vibration Drive End Radial X (μm)
- x_9: Compressor Mechanical Performance Monitoring Vibration Drive End Radial Y (μm)
- x_{10}: Compressor Mechanical Performance Monitoring Vibration Non-Drive End Radial X (μm)

- x_{11}: Compressor Mechanical Performance Monitoring Vibration Non-Drive End Radial Y (μm)
- x_{12}: Compressor Mechanical Performance Monitoring Axial Position A (mm)
- x_{13}: Compressor Mechanical Performance Monitoring Axial Position B (mm)
- x_{14}: Compressor Mechanical Performance Monitoring Drive End Radial Bearing Temperature ($^{\circ}C$)
- x_{15}: Compressor Mechanical Performance Monitoring Thrust Bearing Active Temperature 1 ($^{\circ}C$)
- x_{16}: Compressor Mechanical Performance Monitoring Thrust Bearing Active Temperature 2 ($^{\circ}C$)
- x_{17}: Compressor Mechanical Performance Monitoring Non-Drive End Radial Bearing Temperature ($^{\circ}C$)
- x_{18}: Compressor Turbine Thermodynamic Performance Monitoring Compressor Discharge Air Pressure ($barg$)
- x_{19}: Compressor Turbine Mechanical Performance Monitoring Gas Generator Bearing #1 Vibration (mm/s)
- x_{20}: Compressor Turbine Mechanical Performance Monitoring Gas Generator Bearing #2 Vibration (mm/s)
- x_{21}: Compressor Turbine Mechanical Performance Monitoring Power Turbine Bearing #4 Vibration (mm/s)
- x_{22}: Compressor Turbine Mechanical Performance Monitoring Power Turbine Bearing #5 Vibration (mm/s)

- x_{23}: Gearbox Mechanical Performance Monitoring High-Side Shaft Drive End Radial Vibration X (μm)
- x_{24}: Gearbox Mechanical Performance Monitoring High-Side Shaft Drive End Radial Vibration Y (μm)
- x_{25}: Gearbox Mechanical Performance Monitoring High-Side Shaft Non-Drive End Radial Vibration X (μm)
- x_{26}: Gearbox Mechanical Performance Monitoring High-Side Shaft Non-Drive End Radial Vibration Y (μm)
- x_{27}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Drive End Radial Vibration X (μm)
- x_{28}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Drive End Radial Vibration Y (μm)
- x_{29}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Non-Drive End Radial Vibration X (μm)
- x_{30}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Non-Drive End Radial Vibration Y (μm)
- x_{31}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Axial Position (mm)
- x_{32}: Gearbox Mechanical Performance Monitoring High-Side Shaft Non-Drive End Radial Bearing Temperature ($^{\circ}C$)
- x_{33}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Drive End Radial Bearing Temperature ($^{\circ}C$)
- x_{34}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Non-Drive End Radial Bearing Temperature ($^{\circ}C$)

- x_{35}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Thrust Bearing Temperature 1 ($^{\circ}C$)
- x_{36}: Gearbox Mechanical Performance Monitoring Low-Side Shaft Thrust Bearing Temperature 2 ($^{\circ}C$)
- x_{37}: Turbine Enclosure Monitoring Temperature ($^{\circ}C$)
- x_{38}: Turbine Enclosure Monitoring Differential Pressure ($mbar$)
- x_{39}: Dry Gas Seal System Monitoring Primary Seal Gas Supply Pressure ($barg$)
- x_{40}: Dry Gas Seal System Monitoring Separation Gas Supply Pressure ($barg$)
- x_{41}: Dry Gas Seal System Monitoring Drive End Primary Vent Pressure ($barg$)
- x_{42}: Dry Gas Seal System Monitoring Non-Drive End Primary Vent Pressure ($barg$)
- x_{43}: Dry Gas Seal System Monitoring Seal Gas Filter dP (bar)
- x_{44}: Dry Gas Seal System Monitoring Seal Gas dP ($barg$)
- x_{45}: Compressor & Condensate Export E-2410A/B Discharge Temperature ($^{\circ}C$)
- x_{46}: Compressor & Condensate Export E-2410A/B Discharge Temperature ($^{\circ}C$)