

University of Bahrain

College of Science

**Department of
Postgraduate Programs**



Realtime Data-driven Fault Detection and Diagnosis System for a Local Refinery

Thesis submitted in partial fulfilment of the requirements for the
degree of M.Sc. in Big Data Science and Analytics

Prepared by:

Abdulrahim Mohammed

20134188

Supervised by:

Dr. Mohamed Bin Shams

Associate Professor

University of Bahrain

Kingdom of Bahrain

March 2022



Thesis Discussion Committee Decision

Student Name: Abdulrahim Mohammed	Student ID: 20134188
Department: Choose an item.	College: Science
Program: M.Sc. in Big Data Science and Analytics	
Thesis Title: Real-time Data-driven Fault Detection and Diagnosis System for a Local Refinery	
Supervisor Name: Dr. Mohamed Bin Shams	Student GPA: 4.00

The Discussion Committee met on [Click here to enter a date.](#) In [Click here to enter text..](#) Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade [Choose an item.](#) in the thesis subject.

- ☐ Without making any changes to the thesis.
- ☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name

1. Dr. Mohamed Bin Shams (Advisor)
2. **Dr. Mustafa Hammad** (Internal Examiner)
3. Dr. El-Sayed M. El-Alfy (External Examiner)

Signature

[Click here to enter text.](#)

[Click here to enter text.](#)

[Click here to enter text.](#)

Abstract

Mechanical assets are vital for process industries business operations. Failure of these assets leads to business interruption and significant losses. Early detection of anomalies can help industries to formulate a maintenance plan and avoid unplanned downtimes. This also enhances assets' reliability, maintainability, and availability. The aim of this study is to leverage different data-driven multivariate statistical techniques to detect a heat exchanger leak at a local refinery in real-time. The data is collected from the refinery's historian for a heat exchanger in a distillation unit preheat train network. A set of offline models are developed using static Principal Component Analysis (PCA), dynamic PCA, static Projection to Latent Structures (PLS), and dynamic PLS to detect a heat exchanger leak that occurred in the past. The models then are deployed online to monitor the heat exchanger's performance and detect leaks in real-time. The results of the study demonstrated that static PCA, dynamic PCA, and dynamic PLS succeeded to detect the leak 50 days (with 95% confidence limit) and 19 days (with 99% confidence limit) before it was found in the facility by operators. The online models also, without generating any false alarms, performed well in monitoring the heat exchanger's performance in real-time.

Dedication

This thesis is dedicated to my parents for their endless love, encouragement, and sacrifices.

Acknowledgments

All praise and gratitude be to Allah (s.w.t), who has guided me to complete this project. This journey would have not been possible without the help and support of many people. I would like to express my sincere gratitude to my supervisor Dr. Mohamed Bin Shams for his guidance, support, and encouragement throughout this research. I am very grateful and lucky to have Dr. Mohamed as a supervisor for my BSc and MSc final projects. His expertise and inspiring personality always motivate me to give my best efforts. My sincere thanks also to Dr. Sawsan Hilal for being encouraging and always there for her students.

My warm and heartfelt thanks go to my supervisor and my colleagues at work, who have been very supportive and encouraging throughout this journey. A special thanks also to Aspentech for providing the software to use in this research.

My deepest gratitude goes to my parents for their prayers, love, and patience. Many thanks to my sisters for always being there for me. I would also like to thank my friends who have always encouraged me.

Table of Contents

Defense Committee’s Approval	A
Abstract.....	B
Dedication.....	C
Acknowledgments	D
Table of Contents.....	E
List of Tables	H
List of Figures.....	I
Chapter One Introduction to the Study	2
1.1. Introduction.....	2
1.2. Research Problem	3
1.3. Justification of the Problem.....	6
1.4. Research Questions.....	6
1.5. Research Objectives.....	7
1.6. Research Significance.....	7
1.7. Thesis Structure	8
Chapter Two Literature Review.....	10
2.1. Introduction.....	10
2.2. Heat Exchanger Tube Failure Analysis	10
2.3. Process Monitoring, Fault Detection, and Diagnosis	12
2.4. Process Fault Detection Approaches	14
2.4.1. Model-based Methods	15
2.4.2. Data-driven Methods	17

2.5. Oil and Gas Products Colors Prediction	29
2.6. Research Gaps	30
2.7. Conclusions.....	31
Chapter Three Methods and Procedures	34
3.1. Introduction.....	34
3.2. Process Description	34
3.3. Principal Component Analysis (PCA).....	36
3.3.1. PCA Algorithm.....	36
3.3.2. Dynamic PCA.....	40
3.4. Projection to Latent Structures (PLS).....	41
3.5. Research Framework	42
3.6. Variable Selection.....	43
3.7. Data Collection and Preparation.....	45
3.8. Offline Model Development.....	48
3.9. Online Model Development	49
3.10. Conclusions.....	50
Chapter Four Results and Discussion	52
4.1. Introduction.....	52
4.2. Results.....	52
4.2.1. Offline Models.....	52
4.2.2. Online Models	65
4.2.3. Economic Benefits.....	71
4.3. Discussion.....	73
4.4. Conclusions.....	77

Chapter Five Conclusions, Limitations, and Future Works.....	79
5.1. Introduction.....	79
5.2. Limitations	79
5.3. Future Works	80
5.4. Conclusions.....	81
References.....	83
Appendices.....	96
المخلص	97

List of Tables

Table (1): A Summary of Fault Detection Studies Approaches.....	25
Table (2): Variables Priority Levels	44
Table (3): Final Selected Variables	45
Table (4): Summary of Variables used for Offline Analysis	47
Table (5): Summary of Variables used for Online Analysis	47

List of Figures

Fig. 1: Shell and Tube Heat Exchanger	3
Fig. 2: Preheat Train Heat Exchangers Maintenance Records at a Local Refinery (Exx is the Exchanger number xx).	4
Fig. 3: Preheat Train Equipment Failure	5
Fig. 4: Process Monitoring Loop	13
Fig. 5: Shewhart Chart	14
Fig. 6: An overview of the process under study at the refinery	36
Fig. 7: Principal Component Analysis.....	38
Fig. 8: Hotelling's T^2 Chart	39
Fig. 9: Research Framework.....	43
Fig. 10: System Data Flow	49
Fig. 11: Static PCA Cumulative Q^2 - Offline 1	53
Fig. 12: Static PCA Hotelling's T^2 - Offline 1	55
Fig. 13: Static PCA Variables Contribution	55
Fig. 14: Static PCA Squared Prediction Error – Offline 1	56
Fig. 15: Dynamic PCA Cumulative Q^2 - Offline 1.....	56
Fig. 16: Dynamic PCA Hotelling's T^2 - Offline 1.....	58
Fig. 17: Dynamic PCA Variables Contribution	58
Fig. 18: Dynamic PCA Squared Prediction Error – Offline 1.....	59
Fig. 19: Static PLS Inputs Cumulative Q^2 - Offline 1	60
Fig. 20: Static PLS Outputs Cumulative Q^2 - Offline 1	60
Fig. 21: Static PLS Hotelling's T^2 - Offline 1	61
Fig. 22: Static PLS Squared Prediction Error – Offline 1	61
Fig. 23: Static PLS Color Prediction – Offline 1.....	62
Fig. 24: Dynamic PLS Inputs Cumulative Q^2 - Offline 1	63
Fig. 25: Dynamic PLS Outputs Cumulative Q^2 - Offline 1.....	63
Fig. 26: Dynamic PLS Hotelling's T^2 - Offline 1.....	64

Fig. 27: Dynamic PLS Squared Prediction Error – Offline 1	65
Fig. 28: Dynamic PLS Color Prediction – Offline 1	65
Fig. 29: Static PCA Cumulative Q^2 - Offline 2	66
Fig. 30: Static PCA Hotelling's T^2 - Offline 2	67
Fig. 31: Static PCA Squared Prediction Error – Offline 2	67
Fig. 32: Dynamic PCA Cumulative Q^2 - Offline 2.....	68
Fig. 33: Dynamic PCA Hotelling's T^2 - Offline 2.....	68
Fig. 34: Dynamic PCA Squared Prediction Error – Offline 2.....	69
Fig. 35: Static PCA Hotelling's T^2 - Online (real plant data).....	70
Fig. 36: Static PCA Squared Prediction Error – Online (real plant data).....	70
Fig. 37: Dynamic PCA Hotelling's T^2 - Online	71
Fig. 38: Dynamic PCA Squared Prediction Error – Online	71
Fig. 39: Reprocessing Cost for 50,000 barrels of Product	72
Fig. 40: Per Day Loss Profit Opportunities	73

Chapter One

Introduction to the Study

- 1.1. Introduction**
- 1.2. Research Problem**
- 1.3. Justification of the Problem**
- 1.4. Research Questions**
- 1.5. Research Objectives**
- 1.6. Research Significance**
- 1.7. Thesis Structure**

Chapter One

Introduction to the Study

1.1. Introduction

In the current volatile market, the integrity and availability of unit operations are crucial to the oil and gas industries. That is because a failure in such assets can lead to unplanned business interruption, and profit loss. According to Senseye (2021), on average, machines failure in the unit operations results in 32 hours of unplanned downtime per month for each oil and gas facility. The annual financial cost of this is estimated to be \$84 million per facility.

Industries usually rely on conventional reactive, preventive, or condition-based maintenance strategies to maintain these assets. These maintenance approaches, however, require a significant amount of investment but remain inefficient. Moreover, despite implementing different maintenance strategies, operation assets naturally deviate from their optimal states due to extended use, changes in process dynamics, or more random degradation.

Recently, there has been a growing interest in adopting data-driven fault detection approaches to identify anomalies in unit operations (AlGhazzawi & Lennox, 2008; Al Tamimi et al., 2019). These methods leverage the process and mechanical asset's data to detect faults at early stages. This can help the oil and gas industries to prepare a maintenance plan, minimize assets downtimes, and avoid process disruption.

1.2. Research Problem

Heat exchangers are unit operations widely used in process chemical and petrochemical industries to transfer heat between two fluids. There are several types of heat exchangers used in industries. One of the most common types is the shell and tube heat exchanger consisting of a bundle of tubes mounted inside a cylindrical shell. There are two streams passing through this heat exchanger, namely, a cold stream, and a hot stream. One of the streams flows over the tubes and the other fluid passes through the shell as shown in Fig. 1. While the two fluids flow over the heat exchanger, the heat is transferred from the hot stream to the cold stream (Serth, 2007).

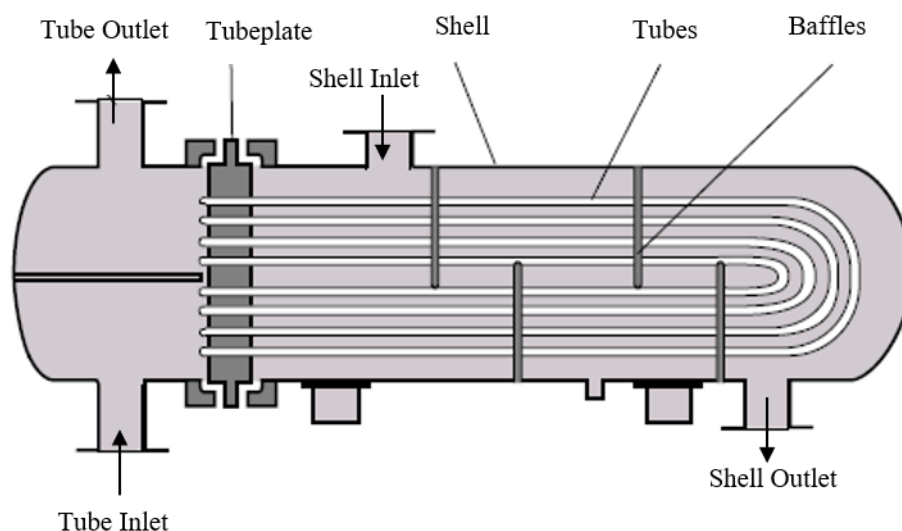


Fig. 1: Shell and Tube Heat Exchanger

Shell and tube heat exchangers are widely found in the downstream refining industries crude preheat train networks. This network is an essential system to minimize the amount of energy required to heat the crude to its boiling temperature for separation. In this system, the crude is preheated with products and other hot streams in a series of heat exchangers to increase the temperature from an ambient condition to a higher temperature. The crude is further heated to its distillation

temperature using heaters. The higher the crude is heated in the preheat train network, the lesser the fuel gas required to heat the crude in the heaters leading to energy saving.

There are many events or faults associated with the heat exchangers which have negative effects on the refinery process operations. One of the frequent faults is internal leak. There have been 45 internal leak incidents reported in the past few years for 17 heat exchangers in a preheat train network at a local refinery. Fig. 2 illustrates the number of internal leak events reported in refinery's internal records for each heat exchanger in the network. Furthermore, 90% of the equipment failure in the preheat train was identified to be due to internal leak incidents as demonstrated in Fig. 3.

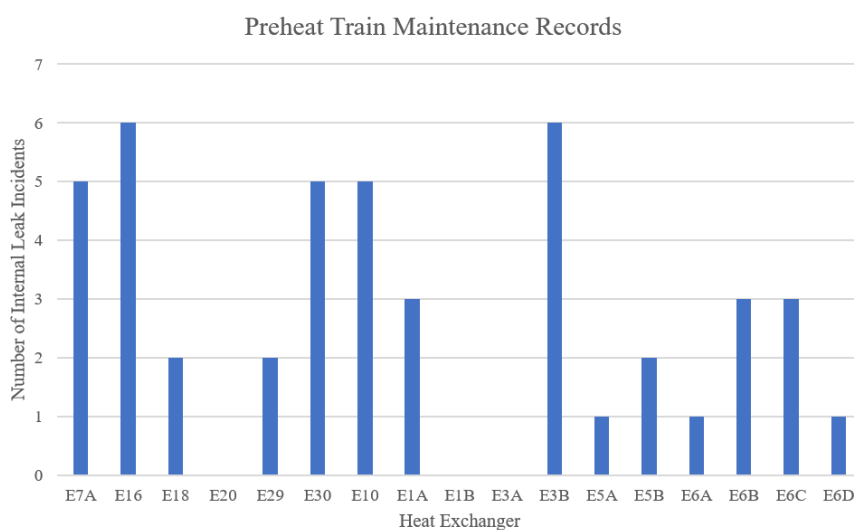


Fig. 2: Preheat Train Heat Exchangers Maintenance Records at a Local Refinery (Exx is the Exchanger number xx).

Preheat Train Equipment Failure

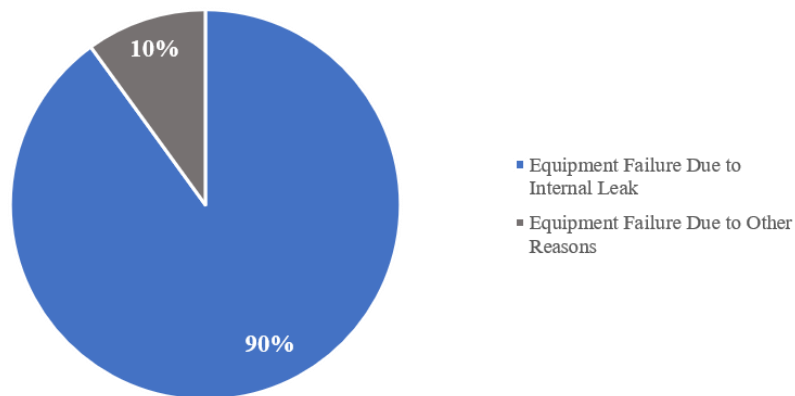


Fig. 3: Preheat Train Equipment Failure

The current deployed fault detection and diagnosis scheme for the heat exchangers' network depends on the operators' skills and experience. Experienced operator accomplishes this by visually analyzing many correlated univariate process trends e.g., temperature, pressure, and flows. Refineries also occasionally monitor the color of the product streams as changes in this parameter are an indication of a possible leak in the heat exchangers. However, apart from the high cost and maintenance expenses of color analyzers, these devices are prone to malfunction. Moreover, the observable color change can be considered as a late fault signature, since at this delayed stage the crude has already been carried over to the product.

Moreover, while several studies demonstrated the effectiveness of using statistical multivariate analysis in detecting faults in manufacturing processes, very limited studies have applied it to the refinery operations. Furthermore, none of the studies have investigated the use of statistical methods to detect heat exchanger internal leaks in real-time.

1.3. Justification of the Problem

The strong performance of the preheat train network is essential to the crude distillation units. While it plays a critical role in the overall process performance, the deterioration in its performance may be a source of considerable risk. The heat exchanger tube leak if not detected at an early stage, can result in loss of containment or carryover of crude to products. This can lead to numerous instances of Loss Profit Opportunities (LPO) which negatively contribute to the refinery's financial performance. Detecting these faults at the earliest helps refiners to take the necessary actions before affecting product quality.

Furthermore, when a heat exchanger in a preheat train network is isolated because of an internal leak, the unit feed is reduced for the entire downtime period, which implies a production reduction. Early detection of leaks allows an early maintenance schedule and hence minimizes the associated downtime intervals. This also helps oil and gas industries to avoid business interruptions and improve their financial performance.

The lack of studies applying multivariate statistical analysis in detecting faults in refineries heat exchangers is another motivating factor to adopt Principal Component Analysis (PCA) and Projection to Latent Structure (PLS) to detect leaks in shell and tube preheat exchangers at a local refinery.

1.4. Research Questions

This study aims to address the following research questions:

- What is the performance of the data-driven multivariate techniques such as static PCA, dynamic PCA, static PLS, and dynamic PLS in detecting heat exchanger leaks that occurred in the past?

- How soon a heat exchanger leak can be detected using multivariate statistical analysis?
- What are the main contributing factors to heat exchangers internal leaks in a crude preheat train network?
- How proficient are multivariate statistical approaches in monitoring heat exchangers and detecting leaks in real-time?
- What is the performance of static PLS and dynamic PLS in predicting oil and gas industries products colors?
- What are the economic benefits when a heat exchanger leak is detected early before the product quality is affected?

1.5. Research Objectives

The objectives of this project are to demonstrate the importance of process data analytics in detecting faults at refineries heat exchanger networks through:

- Developing a real-time fault detection and diagnosis system that leverages multivariate statistical techniques such as PCA, PLS, and their dynamic variants.
- Applying the developed system for early detection of anomalous conditions in a local refinery preheat train heat exchanger.

This project also aims to assess the performance of static and dynamic PLS in predicting oil and gas industries products colors.

1.6. Research Significance

The significance of the current research can be summarized as follows:

- The study leverages multivariate statistical data analysis to detect a common fault in a local refinery crude preheat train heat exchanger.

- The research findings will help to identify the contributing factors and patterns that lead to an internal leak in shell and tube heat exchangers.
- The study fills the gap in the literature in applying multivariate data analysis techniques to detect and diagnose faults in large scale oil and gas industrial settings.
- The research examines the performance of PCA, dynamic PCA, PLS, and dynamic PLS in detecting leaks in oil and gas industries heat exchangers.
- The study evaluates the performance of PLS, and dynamic PLS in predicting refineries products colors.
- The research demonstrates the financial benefits of early leak detection in oil and gas industries preheat train network heat exchangers.

1.7. Thesis Structure

This study is organized into the following five chapters: chapter one introduction, chapter two literature review, chapter three methods and procedures, chapter four results and discussion, and chapter five conclusions, limitations, and future works.

Chapter Two

Literature Review

2.1. Introduction

2.2. Heat Exchanger Tube Failure Analysis

2.3. Process Monitoring, Fault Detection, and Diagnosis

2.4. Process Fault Detection Approaches

2.5. Oil and Gas Industries Products Color Prediction

2.6. Research Gaps

2.7. Conclusions

Chapter Two

Literature Review

2.1. Introduction

This chapter provides a detail literature review of previous studies relevant to this research. First, process industries heat exchanger tube failure analysis studies are presented. Next, the concept, definitions, and objectives of process monitoring, fault detection, and diagnosis are explained. After that, different approaches applied in the literature to monitor and detect faults in various processes are described. The chapter then presents previous studies of predicting products color in oil and gas industries. Finally, the gaps found in the reviewed literature are highlighted.

2.2. Heat Exchanger Tube Failure Analysis

Several studies have investigated the failure of heat exchanger tubes in process industries. Many researchers identified corrosion as the main reason behind the failure of heat exchanger tubes in different applications. In a case study, Panahi et al. (2020) studied the failure of a shell and tube heat exchanger in a natural gas treating plant after 8 years of service. The heat exchanger in this application is used to transfer the heat from hot natural gas in the tubes to cold amine solution inside the shell. To understand the underlying reasons behind the tube failure, the researchers performed a microscopic examination and analyzed the chemical composition of the tube materials. Moreover, they studied the microstructure of the tubes, performed X-ray diffraction analysis from the

surface scales formed on the failed tube, and analyzed the Activated Methyl Diethanol Amine (AMDEA) solution in the shell. The case study revealed that heat exchanger tube leakage was mainly due to crevice corrosion accompanied by cracking. The authors recommended the use of more resistive alloys, tight control of corrosive agents, and proper tube expansion to avoid future tube failures.

Ranjbar and Abasi (2013) in a study assessed the failure and leakage of crude oil preheater tubes in Mono Ethylene Glycol (MEG) and water solution. In this heat exchanger that consists of fire tubes and oil tubes mounted inside a shell full of MEG and water mixture solution, the heat is first transferred from the fire tubes to the MEG solution in the shell and then transferred from the solution to the crude oil. The reports indicated that the tubes suffered from external corrosion leading to leakage. The researchers performed chemical and metallurgical tests to examine the failure of the tubes. The corrosion rate was also evaluated through immersion and polarization tests where the pH of the solution was continuously monitored. The results of this study demonstrated that various acids have been formed due to the interaction of MEG solution with fire tubes. Consequently, the pH of the solution has decreased accelerating the corrosion rate of crude tubes. The corrosion created cavities on the surface of the tubes leading to failure and leakage. Finally, the researchers reduced the corrosion rate using a pH stabilizer (pH=10.5).

Mazlan et al. (2018) also carried out a failure analysis on heat exchanger tubes subjected to Naphthenic Acid Corrosion (NAC) in a refinery vacuum distillation plant. The tests included visual examination of the failed tube and micro-hardness test of the leakage zone. Moreover, scanning electron microscopy and surface microstructure X-ray diffraction was performed to identify and analyze the cracking boundaries

on the surface of the tubes. Finally, the alloy structure was studied via electron dispersive X-ray. The research findings revealed that the formation of chromium carbide was accompanied by corrosion. The authors concluded that the material used in the heat exchanger tube was not suitable for this application due to the absence of elements to protect the tubes against NAC attack.

Usman and Khan (2008) conducted another failure analysis on heat exchanger tubes in an ammonia plant. The heat exchanger was used to produce steam at 306 °C using the reformer outlet gas heat inside the shell. The analysis performed to identify the cause of heat exchanger tubes failure included visual inspection, chemical studies, metallographic analysis, and fractographic analysis. Unlike previous studies that identified corrosion as the main cause behind the leakage of the heat exchanger tubes, the results of this research illustrated that the failure occurred because of thermal fatigue that resulted in cracking of the tubes.

2.3. Process Monitoring, Fault Detection, and Diagnosis

Industries usually use process control systems e.g., the ubiquitous Proportional, Integral, Derivative (PID) controllers or the Advanced Process Control (APC) technologies to maintain safe operation, increase throughput and yield as well as to boost energy efficiency. Process control systems achieve these objectives by compensating for several types of disturbances. However, there are deviations which cannot be addressed effectively by these control systems. These deviations are referred to as faults (Chiang et al., 2001). Isermann and Ballé (1997) define fault as “an unpermitted deviation of at least one characteristic property or parameter of the system from the acceptable/usual/standard

condition” (p. 710). The occurrence of faults in process and manufacturing industries is manifested as process parameter changes, disturbance parameter changes, control valve or sensor problems (Kesavan & Lee, 1997).

The objective of process monitoring activity is to ensure a smooth operation while dealing with abnormal conditions. In other words, process monitoring is associated with detecting, identifying, diagnosing, and removing faults that occur in a process as demonstrated in Fig. 4. Fault detection is to determine if a fault has occurred in a process. Early fault detection provides an early warning to the operation personnel to rectify the situation and hence avoid any serious consequences of it in the future. Fault identification is about finding information related to the fault that has taken place (Jeong et al., 2019). Fault diagnosis, however, is to identify type, size, magnitude, time, and location of the detected anomaly. Finally, process recovery is to eliminate the fault by taking the necessary corrective actions. In a process monitoring loop, the fault detection process is repeated. When a fault is detected, fault identification, fault diagnosis, and process recovery procedures are deployed sequentially in a closed loop (Chiang et al., 2001).

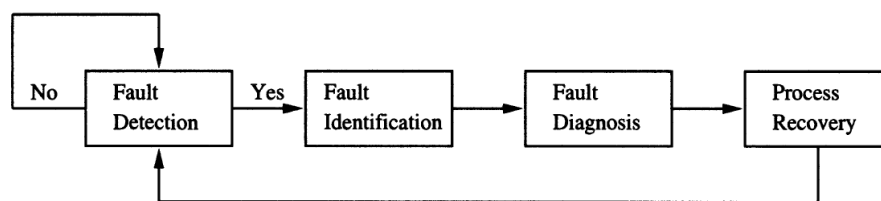


Fig. 4: Process Monitoring Loop

Effective implementation of process monitoring procedures helps to reduce unscheduled downtime, improve safety, and decrease operational cost. However, due to high dimensionality and complexity of modern processes, the faults cannot be efficiently addressed through univariate

control charts. Univariate monitoring is to monitor individual process variables to detect faults using Shewhart chart shown in Fig. 5. Any value less than or higher than the lower or upper control limits, respectively, is considered as fault. In this sequential hypothesis testing approach to monitor a process variable, the challenge is to determine the lower and upper control limit values such that the probabilities of type I and II errors are minimized. Therefore, several better multivariate methods for process monitoring and fault detection have been developed and applied (Chiang et al., 2001).

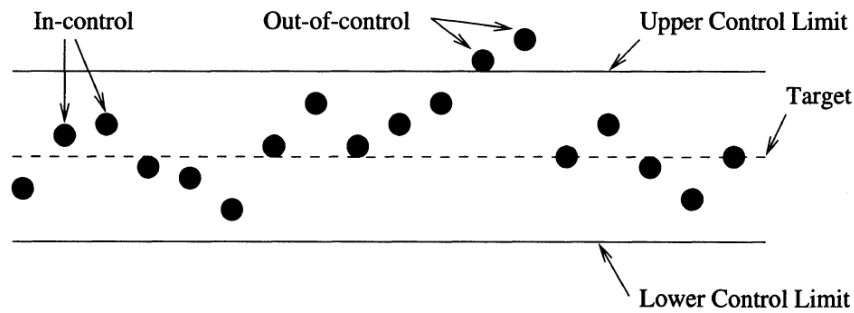


Fig. 5: Shewhart Chart

2.4. Process Fault Detection Approaches

There has been an increasing interest in applying fault detection and diagnosis systems to enhance safety, reliability, efficiency, and profits in the process and manufacturing industries. However, there is no single approach for process monitoring or fault detection and diagnosis. Different studies have applied different methods to monitor and detect faults in industrial processes. These methods have shown different capacities in dealing with different types of faults. This section discusses in detail various methods that have been used in previous process monitoring and fault detection studies.

2.4.1. Model-based Methods

Model-based fault detection approaches are built using process input and output signals and implementing mathematical dynamic process models (Isermann, 2005). Numerous studies have used model-based methods to detect faults in different applications. Weyer et al. (2000) applied a grey-box model-based approach to detect faults in heat exchangers. This approach requires to build models for both heat exchanger and the fault. First principles model of the process was developed using heat exchanger energy balance equations. Moreover, a grey-box model of the deterioration of the heat transfer surface was adopted to identify fault in the heat exchanger. The researchers utilized a recursive identification technique with forgetting factor to trace heat transfer coefficients. At normal state the heat transfer coefficient is constant or slightly declining. However, during abnormal situations due to material breakage the coefficients increase dramatically, and the fault is detected through a Cumulative Sum (CUSUM) test. The simulation results illustrated the effective performance of the model to detect faults in different scenarios.

Detecting faults through residuals generation is a common approach deployed in model-based methods. A residual is a signal that has a value of zero in the absence of faults, and non-zero upon the occurrence of a fault (Svard, 2009). Kinnaert et al. (2000) designed a model-based fault detection system for a gas-liquid separation unit. The system in this study is designed to detect faults in sensors, actuators, and process components in the unit. Initially, a first principle dynamic model of the plant is developed using material balance equations. Residual generators are then designed to detect faults using CUSUM test. These signals have been generated using mathematical models of the unit as well as sensors and

actuators measurements. Next, a decision module is developed to determine if the residuals are considerably varying from zero and to identify the faulty components. Finally, the results of the study have been validated using simulation. While, CUSUM test assumes a known fault magnitude, the study findings demonstrated that the applied approach shows a reasonable performance in detecting faults with a magnitude twice nominal fault magnitude. In another model-based fault detection study, Sun et al. (2002) used least square algorithm with time varying factor to detect leak in steam-water boiler systems. The residuals were estimated comparing actual measurements and the predicted measurements using overall boiler material balance relationship. However, the residual signals were observed to be noisy and difficult to use for leak detection. Therefore, the residual was processed with a time-varying factor in a least square algorithm to improve the leak detection accuracy. Furthermore, due to existing disturbances and noisy measurements in the plane, to reduce the number of false alarms and avoid missing faults the leak detection alarm threshold was set using normal operating condition of the boiler. The model was tested on actual industrial data to detect three leaks happened earlier. The results showed that the model was able to detect the leak 16, 27, and 13 days before it was discovered by operators for three different cases.

Several other studies have used parity equations as an approach to generate residuals and detect faults in systems. In this method the residuals are generated based on the deviation of the process with the measured process model. There are several ways to develop parity equations including state space models and transfer functions (Fagarasan et al., 2006). Krishnan and Pappa (2005) used parity equations to detect sensors, actuators, and process faults for a shell and tube heat exchanger in an experimental setup. The residuals are generated by comparing actual

process model with the estimated process model that were developed using transfer functions. The residuals in this study were designed in such a way that each residual detects distinct faults. The model performed well when it was tested against sensor, actuator, and process faults. Moreover, Fagarasan et al. (2006) utilized parity equations to detect faults in a thermal process. Parity equations were developed by comparing the process model with the nominal process operating conditions using transfer functions. Like Krishnan and Pappa (2005), in this study also the residuals were developed to identify specific faults in the process. The thresholds, however, were determined experimentally to cover the noise effects. Finally, the performance of all the 7 residuals on detecting sensors and actuators faults were evaluated. It was found that detecting faults rely on the performance of the most sensitive residuals.

Other model-based approaches for fault detection includes observer-based method. In this approach the residual is generated by comparing the estimates from the observer with actual process measurements. The generated residual carries a value of zero in fault-free systems and a non-zero value when a fault occurs. However, due to measurement noises, model uncertainties, and process disturbances, the residuals deviate from zero in the absence of faults. The residual evaluation procedure then is carried out to trigger alarms in the presence of faults only (Khan, 2013). This approach has been implemented by many researchers over the past few decades to detect faults particularly in nonlinear systems (Hammouri et al., 1999; Yan & Edwards, 2008; Li et al., 2016; Jeong et al., 2019).

2.4.2. Data-driven Methods

Multivariate statistical and Machine Learning (ML) approaches have been used widely in the past few decades for fault detection and diagnosis in various industrial processes including chemical industries (Chiang &

Colegrove, 2007), oil and gas industries (AlGhazzawi & Lennox, 2008), steel industries (Zhang & Dudzic, 2006), nuclear power plant (Yu et al., 2020), smelting processes (Hu et al., 2014), and food industries (Yu & MacGregor, 2003). Unlike model-based fault detection approaches that rely significantly on mathematical and physical relationship of the process, data-driven techniques utilize process historical data and require minimum priori knowledge about the system (Bounoua & Bakdi, 2021).

2.4.2.1. Multivariate Statistical Techniques

Multivariate statistical methods are effective tools for process monitoring and fault detection due to their capabilities to extract information from a massive, correlated dataset by transforming it into a smaller set of uncorrelated parameters. Unlike model-based approaches, these methods are capable of handling measurement noise and extract original information effectively (Venkatasubramanian et al., 2003). Principal Component Analysis (PCA) (Bounoua et al., 2019), Partial Least Square or Projection to Latent Structures (PLS) (Godoy et al., 2013), Support Vector Machines (SVM) (de Souza et al., 2014), Fisher Discriminant Analysis (FDA) (Chiang et al., 2004), Independent Component Analysis (ICA) (Wang Kai et al., 2011), and Slow Feature Analysis (SFA) (Zhang et al., 2015) are common multivariate statistical techniques used for fault detection. However, PCA and PLS and their variants are the leading approaches that have been used widely for process monitoring and fault detection (Nawaz et al., 2021).

PCA is a dimensionality reduction technique that generates a lower dimensional set of uncorrelated variables by capturing the variability and correlation structure of the original dataset. For majority of the applications the significant variation in the dataset is captured with two or three dimensions, also known as latent variables. Moreover, this

technique helps to visualize the variation in the data using single plots. This simplifies and enhances process monitoring, fault detection, and fault diagnosis procedures and allows plant operators and engineers to interpret important patterns in the operated process. Furthermore, the knowledge extracted by PCA facilitates the identification of the variables responsible for the occurred faults (Chiang et al., 2001).

Numerous studies applied PCA for process monitoring, fault detection, and fault diagnosis in the past few decades. A few studies have used PCA to detect faults in oil and gas industries. AlGhazzawi and Lennox (2008) used PCA to monitor and detect abnormal conditions in a complex refining fractionation process by providing early indication in case of deviating from optimal products specifications. As the plant was controlled by a Model Predictive Control (MPC) system, the secondary objective of this study was to monitor the performance of APC controller as well. Following the careful study of the process, conducting interviews with the plant operators, and surveying the existing variables, a subset of the available measurements has been used to build the PCA model. Next, the normal operation range has been identified by excluding anomalous data, such as valve saturation periods, or when maintenance activities were performed on plant equipment. The data then was collected, in one-hour frequency, from plant historian for the normal operation horizon (8 months). A static PCA model for the entire process was developed using seven principal components. Moreover, multi-block PCA was developed to facilitate the root cause analysis of the faults. Hotelling's T^2 and Squared Prediction Error (SPE) charts were used to detect anomalous conditions. Furthermore, to address the static PCA limitation of not considering the time-series nature of data, a Recursive Principal Component Analysis (RPCA) model was also implemented. The study findings showed that unlike static PCA, RPCA was effective in detecting

faults in the process studied in this research. While the whole analysis was done in an offline setting, the study recommended to implement PCA in real-time to detect refinery process faults. Al Tamimi et al. (2019) also employed PCA to monitor and troubleshoot a hydrogen plant in a local refinery. The researchers collected data for 61 different variables around the plant for a period of 3 months in 1-minute intervals. Hotelling's T^2 and SPE charts have been used in this study to detect faults in the process. Moreover, the variables contributing to each fault have been identified in this research. Finally, the study demonstrated the economic benefits of applying such fault detection techniques by comparing the operation cost and production rate in the absence and presence of this tool. Again, an offline approach has been adopted in this work.

Unlike oil and gas industries, PCA has found many applications in chemical industries. However, due to the nonlinearity of chemical processes, most of the studies adopted different versions of PCA to improve its performance to monitor and detect faults in complicated nonlinear systems. Geng and Zhu (2005) for instance, proposed a Multiscale adaptive Nonlinear PCA (MS-NLPCA) developed using input training neural network. The model was used to monitor an ethylene cracking furnace and detect abnormal conditions in Coil Outlet Temperatures (COTs). The results of the study demonstrated that MS-NLPCA has shown a better performance in comparison with traditional linear PCA by providing earlier indication of faults and generating less false alarms.

Dynamic Principal Component Analysis (DPCA) (Ku et al., 1995), Kernel Principal Component Analysis (KPCA) (Deng et al., 2013), and Dynamic Kernel Principal Component Analysis (DKPCA) (Choi & Lee, 2004) have also found a lot of applications to monitor and detect faults in nonlinear chemical processes. DPCA was first presented by Ku et al.

(1995) to monitor and detect disturbances in nonlinear Tennessee Eastman process simulation. The algorithm was designed to account for autocorrelation in time-series data by adding variables with time lag to the original set of variables in the PCA model. Study findings illustrated that DPCA performed better than static PCA in detecting small disturbances in the process. KPCA finds principal components in extended feature space instead of input space (Muller et al., 2001). Choi and Lee (2004) proposed a DKPCA and applied this model on a nonlinear wastewater treatment plant for process monitoring and fault detection. KPCA was transformed to DKPCA by adding time-lagged input vectors. A comparison study of PCA, DPCA, KPCA, and DKPCA was conducted. Similar to other PCA studies, Hotelling's T^2 and Q statistics (in the feature space) were used to detect sensor faults and anomalous process conditions. The results showed that DPCA and DKPCA reduced the false alarms rate from 17.4% to 5.1% comparing to the other two monitoring methods.

PLS is another dimensionality reduction method that predicts outputs by maximizing the variance between the dependent and independent matrices. A common application of PLS is predicting process product quality data using other process variables. In addition to online product quality prediction, PLS can be used in process control algorithms as well as for process monitoring, fault detection, identification, and diagnosis. Each PCA fault detection and diagnosis method can be applied to PLS (e.g., T^2 and Q statistics), however, PLS has the advantage of monitoring the variations in the inputs that are related to the product quality parameters (Chiang et al., 2001).

Several studies adopted PLS to monitor products quality variables and detect faults in chemical systems (Kourti & MacGregor, 1995; Godoy et al., 2013; Jia & Mao, 2013; Wang & Yin, 2015). Xu et al.

(2018) also applied dynamic PLS to monitor gasoline and diesel endpoints in an oil and gas fractionation process. The data was collected for 28 variables around the column to predict the two quality variables and detect faults in the process. The study outcomes demonstrated that both PLS and dynamic PLS performed well in predicting the endpoints of gasoline and diesel, however, dynamic PLS achieved a higher prediction accuracy. Furthermore, all process anomalies were effectively detected using T^2 and Q statistics. In another study, Marjanovic et al. (2006) used a multiway PLS to accurately identify the endpoint of a batch process reaction that converts raw materials to final products. The multiway PLS that have been used prior to this research by other researchers (Wise & Gallagher, 1996; Lennox et al., 2001) on other nonlinear processes, transforms the three-dimensional data in hand into a two-dimensional matrix. The findings indicated that the technique adopted in this study has accurately estimated the reaction conditions and effectively predicted the cycle time of each batch.

2.4.2.2. Machine Learning Techniques

ML approaches have also been very popular to detect faults in nonlinear systems. Artificial Neural Network (ANN) (Patton et al., 1994; Jack & Nandi, 2002; Heo & Lee, 2018; Iqbal et al., 2019) and Long Short-Term Memory (LSTM) recurrent neural network (Zhao et al., 2018; Gravanis et al., 2022; Wang et al., 2022) are the most common methods used in the literature.

Different studies applied various ML algorithms to detect faults in Tennessee Eastman process. Tennessee Eastman is a benchmark chemical process simulator that mimic a real chemical plant that comprises of a reactor, a stripper, a separator, a condenser, and a compressor. The

process yields two main products by processing several inputs (Gravanis et al., 2022).

Heo and Lee (2018) applied ANN with different number of hidden layers (0, 1, 2, 3, and 4 layers) to detect 20 different faults in the process simulation. The results reported in the study demonstrated that ANN showed an outstanding performance with an average accuracy of 97.73% fault detection rate and 0.25% false alarm rate. Except for faults 8, 12, 14, and 17, ANN also outperformed DPCA and modified PLS that reported an average accuracy of 89.13% and 96.32% respectively in a study carried out by Yin et al. (2012) on the same process.

In a different study, Wu and Zhao (2018) used Deep Convolutional Neural Network (DCNN) for fault detection in the same process. To use DCNN for detecting 20 different abnormal conditions in the process, the data has been transformed into a two-dimensional matrix and divided into train and test. The train data used to build the model and test data to evaluate its performance. 12 different models with various architectures have been developed and tested. The researchers excluded faults 9, 15, and 16 without explaining the reason behind it. The findings showed that DCNN with an average fault detection accuracy of 97% and false alarm rate of 0.4% performed well. The study also demonstrated that the model can accurately detect the alarms 2-3 hours after the fault is introduced.

Gravanis et al. (2022) in another study adopted two different recurrent neural network algorithms namely, LSTM and Time Delay Neural Network (TDNN) with different feature reduction techniques to detect faults in the Tennessee Eastman process simulation. The study outcomes illustrated that LSTM with KPCA as feature reduction with an average accuracy of 98.5% for 20 different faults performed better than TDNN with KPCA as feature reduction with an average accuracy of 98.1%.

From the comparison of the results of several fault detection studies on the Tennessee Eastman process simulation, it is observed that ML algorithms provided a higher accuracy comparing to the multivariate statistical techniques, however, the latter provided more meaningful information about the fault such as location, time, magnitude, and variables contributing to the fault which are very useful in the process of fault diagnosis and elimination.

2.4.2.3. Hybrid Techniques

Many studies have tried to leverage the accuracy of ML algorithms and interpretability of multivariate statistical techniques by applying hybrid models that contain both approaches. Chen and Liao (2002) combined neural network and PCA (NNPCA) to monitor and detect faults in Tennessee Eastman process simulation. The study used neural network to model the nonlinear dynamic process and applied PCA to produce interpretable monitoring charts from the residuals generated by comparing actual measurements with neural network prediction. The hybrid model has been applied on 15 different faults and compared with PCA and DPCA models implemented on the same faults. The results demonstrated that for all 15 different faults NNPCA performed better than individual PCA and DPCA in terms of fault detection rate.

In another study, Zhu (2007) adopted a model that integrates NN and PLS to monitor two product quality variables in a factory. NN is incorporated in the PLS model to enhance the accuracy while monitoring nonlinear systems. The researcher concluded that the model yielded a high accuracy without reporting the model's performance in terms of fault detection rate or false alarms generation. Table (1) provides a summary of different process monitoring and fault detection approaches have been applied in the reviewed literature.

Table (1): A Summary of Fault Detection Studies Approaches

Studies		Model-based			Data-driven					
Authors	Context	Parity Equations	Observer-based	Others	Principal Component Analysis (PCA)	Projection to Latent Structures (PLS)	Support Vector Machines (SVM)	Artificial Neural Network (ANN)	Long Short-Term Memory (LSTM)	Convolutional Neural Network (CNN)
Patton et al. (1994)	Laboratory 3-tanks system, experimental environment							✓		
Kourti & MacGregor (1995)	Industrial polymerization process					✓				
Ku et al. (1995)	Tennessee Eastman process, Simulation				✓					
wise & Gallagher (1996)	Slurry-Fed Ceramic Melter					✓				
Weyer et al. (2000)	Heat exchanger			✓						
Kinnaert et al. (2000)	Gas-liquid separation unit, simulation			✓						
Lennox et al. (2001)	Fed-batch fermentation, Austria					✓				
Sun et al. (2002)	Boiler steam-water, Canada			✓						
Chen and Liao (2002)	Tennessee Eastman process, Simulation				✓			✓		
Jack & Nandi	Rotating						✓	✓		

Studies			Model-based			Data-driven					
Authors	Context		Parity Equations	Observer-based	Others	Principal Component Analysis (PCA)	Projection to Latent Structures (PLS)	Support Vector Machines (SVM)	Artificial Neural Network (ANN)	Long Short-Term Memory (LSTM)	Convolutional Neural Network (CNN)
(2002)	machinery, experimental environment										
Choi & Lee (2004)	Wastewater treatment plant, Simulation					✓					
Krishnan & Pappa (2005)	Heat exchanger, experimental environment		✓								
Geng & Zhu (2005)	Petrochemical, china					✓					
Fagarasan et al. (2006)	Thermal process, experimental environment		✓								
Marjanovic et al. (2006)	Chemical batch process, UK						✓				
Zhu (2007)	Product factory, simulation						✓		✓		
Yan & Edwards (2008)	Aircraft system, simulation			✓							
AlGhazzawi & Lennox (2008)	Oil and Gas, Saudi Arabia					✓					
Yin et al. (2012)	Tennessee Eastman process, simulation					✓	✓				
Godoy et al. (2013)	Chemical process, Simulation						✓				

Studies			Model-based			Data-driven					
Authors	Context		Parity Equations	Observer-based	Others	Principal Component Analysis (PCA)	Projection to Latent Structures (PLS)	Support Vector Machines (SVM)	Artificial Neural Network (ANN)	Long Short-Term Memory (LSTM)	Convolutional Neural Network (CNN)
Deng et al. (2013)	Tennessee Eastman process, simulation					✓					
Jia & Mao (2013)	Solvent extraction process, simulation						✓				
de Souza et al. (2014)	Reaction system, simulation							✓			
Wang & Yin (2015)	Tennessee Eastman process, simulation						✓				
Li et al. (2016)	Numerical example			✓							
Xu et al. (2018)	Oil and Gas, China						✓				
Heo & Lee (2018)	Tennessee Eastman process, simulation								✓		
Zhao et al. (2018)	Tennessee Eastman process, simulation									✓	
Wu and Zhao (2018)	Tennessee Eastman process, simulation										✓
Jeong et al. (2019)	Numerical simulation			✓							

Studies			Model-based			Data-driven					
Authors	Context		Parity Equations	Observer-based	Others	Principal Component Analysis (PCA)	Projection to Latent Structures (PLS)	Support Vector Machines (SVM)	Artificial Neural Network (ANN)	Long Short-Term Memory (LSTM)	Convolutional Neural Network (CNN)
Bounoua et al. (2019)	Tennessee Eastman process, simulation					✓					
Al Tamimi et al. (2019)	Oil and Gas, Bahrain					✓					
Iqbal et al. (2019)	Computer-based manufacturing system, experimental environment								✓		
Gravanis et al. (2022)	Tennessee Eastman process, simulation									✓	
Wang et al. (2022)	Pressurized water reactor, simulation									✓	

2.4.2.4. Variable Selection

Due to the high number of variables available in process and manufacturing industries, selecting appropriate variables that yield accurate results is crucial in data-driven based process monitoring. Many studies considered a large number of variables in their analysis and then applied feature selection methods to filter important variables. Several studies used PCA as a feature selection technique for fault detection applications (Malhi & Gao, 2004; Song et al., 2010; Yoo, 2019). In this

approach the variables associated with the highest variance due to a fault are identified by evaluating their directionality (Malhi & Gao, 2004). Other methods of feature selection methods include nonlinear SVM (Onel et al., 2019) as well as sequential backward and sequential forward (Rauber et al., 2020) that were applied to select appropriate variables for fault detection in Tennessee Eastman process. All studies demonstrated that feature selection have improved the performance of the models in detecting faults. Gravanis et al. (2021), however, leveraged Non-negative Matrix Factorization (NMF), PCA, and KPCA to reduce the dimensionality by introducing new features. The results illustrated that PCA and KPCA performed better than NMF. The researchers, however, did not present results without applying feature reduction methods.

Other studies used their domain expertise to only include variables that are relevant to the objective. AlGhazzawi and Lennox (2008) for instance, after comprehending process operation, conducting interviews with the operation personnel, and reviewing existing measurements, prioritized variables as ‘critical’, ‘important’, or ‘unimportant’ according to their effect on the plant operation. Finally, they selected the variables with ‘critical’ priority only in the process monitoring and fault detection analysis.

2.5. Oil and Gas Products Colors Prediction

Monitoring and maintaining final products quality specifications are crucial for oil and gas industries. These specifications include color, flash point, density, viscosity, etc. Accurate prediction of these properties can help industries to improve their profits and performance.

Limited studies, however, have adopted data-driven approaches to predict oil and gas industries products colors. Leam et al. (2020) built a

model to correlate the Saybolt color with chemical and physical properties of crude oil such as density and viscosity using different multiple linear regression techniques. The coefficient of determination, R^2 , was used to evaluate the performance of the models comparing to the results from laboratory. The models with a prediction error of less than 5% for 75% of the samples showed a good performance in predicting the color.

In another study, Barra et al. (2022) adopted PLS to predict quality properties of diesel fuel from Fourier Transform Medium Infrared Spectroscopy (FT-MIR) data. The properties predicted include color, density, viscosity, flash point, etc. The model with a R^2 of 0.96 have succeeded to predict diesel color with high accuracy.

2.6. Research Gaps

Numerous studies adopted different approaches to monitor and detect faults in different processes, however, there are limited applications for the oil and gas industries (AlGhazzawi & Lennox, 2008; Al Tamimi et al., 2019). Moreover, none of the studies in the reviewed literature used fault detection techniques to identify abnormal conditions in an industrial heat exchanger network. Although many studies stated that fault detection is associated with financial benefits for process and manufacturing industries, very limited studies (Al Tamimi et al., 2019) demonstrated the economic merits of applying such techniques.

A clear gap in the literature can be found within the context of a ‘real-time’ application of data-driven fault detection in oil refineries, as most of published literature is based on an offline analysis. The applications of online process monitoring, and fault detection are limited to chemical industries (Marjanovic et al., 2006; Chiang & Colegrove,

2007) and experimental environments (Person & Tovornik, 2005; Krishnan & Pappa, 2005). Though it is recommended by AlGhazzawi and Lennox (2008) to adopt process monitoring and fault detection techniques in oil and gas industries on a real-time basis, the latter has not been explored in the previous studies. Finally, even though many fault detection techniques can provide diagnostic information about the anomalies such as fault magnitude and detection time, very few researchers (Sun et al., 2002; Choi & Lee, 2004) presented this data in their studies.

2.7. Conclusions

In summary, many studies demonstrated that corrosion is the dominant reason behind heat exchanger tubes failure and leakage. However, these studies illustrated that corrosion is formed due to different causes and minimizing the factors that accelerate the corrosion rate is pivotal to avoid future tube failures. Process monitoring and fault detection systems help industries to identify anomalous conditions that adversely impact the process performance. These anomalies are difficult to be detected while monitoring individual variables or using process control systems.

There are two main approaches applied in the literature for monitoring and fault detection in process and manufacturing industries: model-based methods and data-driven methods. The former utilizes mathematical and physical relationship of the process to form a fault detection model. Most of the model-based methods generate residuals that have a value of zero at normal conditions and non-zero in the presence of a fault. Residuals can be generated using different methods.

Parity equations and observer-based methods are the most popular approaches used in the reviewed literature.

Although many studies demonstrated the effectiveness of model-based fault detection techniques, there are limited applications on actual manufacturing and process industries setups. Moreover, accurate model development using physical relationship is intricate and requires a lot of priori knowledge about the process. Furthermore, the residuals used to detect faults are usually noisy and generate a lot of false alarms.

Data-driven methods on the other hand leverage the process data to develop monitoring and fault detection models. Multivariate statistics and ML are the main data-driven fault detection techniques adopted in the reviewed literature. PCA and PLS are the most popular multivariate statistical techniques, however, due to nonlinearity of many processes, modified nonlinear PCA and PLS are also widely employed. On the other hand, ANN and LSTM are the most common ML approaches. Although, ML techniques succeeded to detect faults at a higher accuracy comparing to multivariate statistical techniques, the latter provided more useful information such as time, location, magnitude, and contributing variables that is critical for fault identification and diagnosis.

Several studies also developed hybrid models to benefit from both accuracy of ML and interpretability of multivariate statistical models. Lastly, due to the large number of variables in process and manufacturing industries, the variable selection approaches have been discussed. While some studies considered many variables and then applied feature selection methods such as PCA and SVM to reduce the number of variables, others used process and operational expertise to select variables that are related to the objective only.

Chapter Three

Methods and Procedures

- 3.1. Introduction**
- 3.2. Process Description**
- 3.3. Principal Component Analysis (PCA)**
- 3.4. Projection to Latent Structures (PLS)**
- 3.5. Research Framework**
- 3.6. Variable Selection**
- 3.7. Data Collection and Preparation**
- 3.8. Offline Model Development**
- 3.9. Online Model Development**
- 3.10. Conclusions**

Chapter Three

Methods and Procedures

3.1. Introduction

The purpose of this chapter is to present the methods adopted in this study to detect internal leaks in a shell and tube heat exchanger using data-driven multivariate techniques. First, the process around the heat exchanger is briefly described. Next, the algorithms implemented are explained. After that, the framework adopted in this research is introduced and individual stages including variable selection, data collection/preparation, and offline and online model development are detailed.

3.2. Process Description

A data-driven fault detection scheme is proposed in this study to identify internal leak in a heat exchanger network at a local refinery. The crude preheat train network consists of 17 different shell and tube heat exchangers. This network is set up to achieve maximum crude temperature before entering the fired heaters. The more the crude is heated in the preheat train system, the less the fuel gas is required in the heaters, resulting in energy saving.

Crude passes through the tube side of the heat exchangers in the preheat train network and is heated with Preflash, Atmospheric, and Vacuum distillation columns products before entering the heaters as shown in Fig. 6. Crude first exchanges heat with Preflash Medium Straight Run (MSR) product in the heat exchangers E7 and E30. The

heated crude then passes through heat exchanger E10 tubes where it is further heated with Preflash Heavy Straight Run (HSR) product that flows through the shell side of the heat exchanger. Crude oil then passes through the tube side of other heat exchangers in the preheat train network and exchanges heat with other product streams. MSR and HSR products however are further cooled down by passing across cooling water heat exchangers before going for treatment and storage. The heat exchanger E10 that has failed several times in the past few years due to internal leakage has been selected in this study for proposed leak detection scheme.

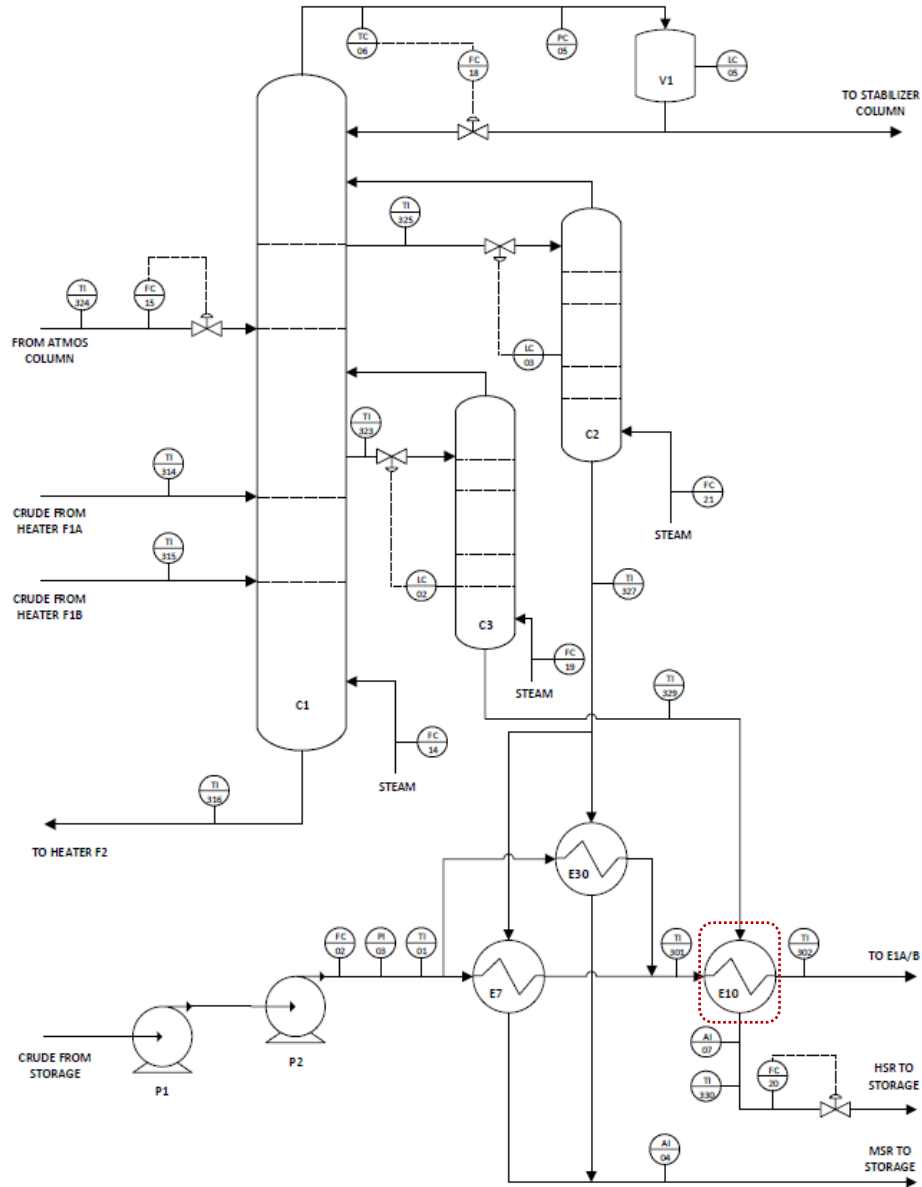


Fig. 6: An overview of the process under study at the refinery

3.3. Principal Component Analysis (PCA)

This section provides an overview of PCA. First, PCA algorithm is explained, and the statistics and plots used for fault detection are described. Next, dynamic PCA algorithm is detailed.

3.3.1. PCA Algorithm

PCA is an unsupervised, linear, dimensionality reduction method that utilizes the correlation among the features to provide a more simplified representation of the process variations. It converts the original

variables to a smaller set of uncorrelated variables while maintaining maximum variation in the dataset. Consider an auto scaled data matrix $\mathbf{X}$ that has zero mean and unit variance with k variables and n observations, the covariance matrix $\mathbf{S}$ is computed from:

$$\mathbf{S} = \frac{\mathbf{X}^T \mathbf{X}}{n - 1} \quad (1)$$

where $\mathbf{S}$ is a $k \times k$ matrix that captures the correlation among the variables, the larger the element of the covariance matrix, the higher the correlation and redundancy between the corresponding variables. Next, the eigenvalues and eigenvectors are estimated via eigenvalue decomposition of the covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \quad (2)$$

where $\mathbf{V}$ contains the loading vectors, and $\mathbf{\Lambda}$ is a diagonal matrix containing non-negative eigenvalues in decreasing order ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq 0$). Since the loading vectors corresponding to larger eigenvalues captures the significant variations in the dataset, the dimensionality of the dataset can be reduced by forming a loading matrix $\mathbf{P}$ that contains loading vectors associated to r largest eigenvalues. The observations in the original data matrix $\mathbf{X}$ then can be projected into a lower dimensional space in the form of score matrix $\mathbf{T}$ (latent variables):

$$\mathbf{T} = \mathbf{X} \mathbf{P} \quad (3)$$

The score matrix $\mathbf{T}$ then can be used to estimate the values in k dimensional space:

$$\hat{\mathbf{X}} = \mathbf{T} \mathbf{P}^T \quad (4)$$

where $\hat{\mathbf{X}}$ is a matrix containing the estimated observations, $\mathbf{T}$ is the score matrix that shows the coordinates along the hyperplane where the observations are projected, and $\mathbf{P}$ contains the loading vectors that indicate the orientation or direction of the hyperplane as shown in Fig. 7.

Finally, the residual matrix $\mathbf{E}$ can be computed by subtracting the estimated observations $\hat{\mathbf{X}}$ from the original observations $\mathbf{X}$:

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad (5)$$

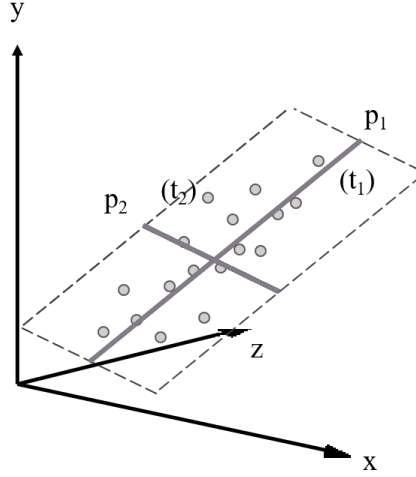


Fig. 7: Principal Component Analysis

The coefficient of determination for individual variables $R_{X,k}^2$ that shows the contribution of each variable to the variation explained by the model for the training dataset can be estimated as:

$$R_{X,k}^2 = \frac{SS(\mathbf{x}_k) - SS(\mathbf{e}_k)}{SS(\mathbf{x}_k)} \quad (6)$$

where $SS(\mathbf{x}_k)$ is the sum of squares of variable k observations in training data matrix $\mathbf{X}$, and $SS(\mathbf{e}_k)$ is the sum of squares of residuals associated with variable k in the training residual matrix $\mathbf{E}$. The overall model coefficient of determination R_X^2 which is an indication of the variation covered by the model for the entire training dataset can be determined as:

$$R_X^2 = \frac{SS(\mathbf{X}) - SS(\mathbf{E})}{SS(\mathbf{X})} \quad (7)$$

where $SS(\mathbf{X})$ is the sum of squares of all observations in training data matrix $\mathbf{X}$, and $SS(\mathbf{E})$ is the sum of squares of all residuals in the training residual matrix $\mathbf{E}$. Furthermore, (6) and (7) can be applied to the test

dataset to estimate $Q_{X,k}^2$ and Q_X^2 , the coefficients of determination for individual variables and entire data in the test dataset respectively. Both R^2 and Q^2 coefficients of determinations carry a value between 0 and 1; the closer the value to 1 the higher the variability explained by the model. However, the coefficient of determination for the test dataset Q^2 is smaller comparing to training dataset coefficient of determination R^2 ($Q^2 < R^2 \leq 1$) (Prosensus, 2014).

Hotelling's T^2 statistic is a well-known scalar that summarizes all scores for each observation vector. T^2 statistic has a positive value and shows the distance of the observations' projection from the center of the identified hyperplane. If there are A scores for each observation vector, then the T^2 value for i^{th} observation can be estimated as:

$$T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2 \quad (8)$$

where s_a^2 is the variance of each score vector $\mathbf{t}_a$ in the $\mathbf{T}$ score matrix.

Hotelling's T^2 summarizes all scores in a single plot and is widely used for fault detection purposes. Values that exceed the threshold which is calculated as weighted F -distribution at a certain level of significance are considered as faults as shown in Fig. 8.

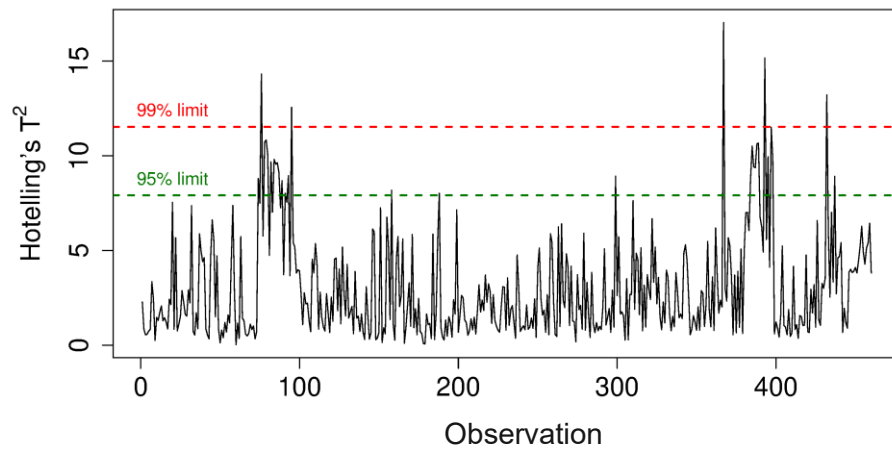


Fig. 8: Hotelling's T^2 Chart

Squared Prediction Error (SPE) is another plot that summarizes the residuals or the distance between the actual observations and the projected observations in a single plot. For a residual matrix $\mathbf{E}$ of size $n \times k$, the SPE for i^{th} observation can be computed as:

$$SPE_i = \sum_{a=1}^{a=k} e_{ia}^2 \quad (9)$$

This plot highlights unusual observations where actual measurements deviate from the observations projected by the model and is used along with T^2 statistic to detect faults in processes. Furthermore, contribution plots can be used to understand the variables contributing to anomalies in T^2 statistic and SPE charts. The contribution of each variable is calculated by comparing normal operation data with anomalous data points in each plot (Chiang et al., 2001; Dunn, 2010; ProSensus, 2014).

3.3.2. Dynamic PCA

Dynamic PCA was first proposed by Ku et al. (1995) to address PCA limitations of building linear static models and not considering the autocorrelative structure of time-series datasets. Dynamic PCA is identical to static PCA described in the previous section, except that time-lagged duplicate of each variable is added to the original dataset. For example, consider a dataset consisting of three variables $u(k)$, $z(k)$, and $p(k)$, the data matrix $\mathbf{X}$ in static PCA is given as,

$$\mathbf{X} = [u(k), z(k), p(k)] \quad (10)$$

Dynamic PCA accounts for autocorrelation or time dependency in the time-series data by appending the data matrix $\mathbf{X}$ with time-shifted variables,

$$\mathbf{X} = [u(k), z(k), p(k), u(k-l), z(k-l), p(k-l)] \quad (11)$$

where $u(k)$, $z(k)$, and $p(k)$ are the variables at time k , and $u(k-l)$, $z(k-l)$, and $p(k-l)$ are the one-time-lagged observations of the three variables.

The same procedures of finding the latent variables and calculating the monitoring statistics for static PCA are used in dynamic PCA.

3.4. Projection to Latent Structures (PLS)

PLS is a supervised dimensionality reduction technique that predicts the outputs by maximizing the covariance between the independent variables, X matrix and dependent variables, Y matrix. In other words, PLS finds the scores and loading vectors that are correlated to the predicted matrix Y and best explain the variance in the predictor matrix X . For process monitoring and fault detection, the predictor matrix contains process variables, and the predicted matrix includes quality or other variables that will be monitored.

As illustrated for PCA, the latent variables for matrix X and matrix Y that best summarize the two matrices can be estimated as:

$$T = XP \quad (12)$$

$$U = YC \quad (13)$$

where T is a matrix containing latent variables for the independent matrix X , U is a matrix comprising the latent variables for the dependent matrix Y , P is the loading matrix for X , and C is the loading matrix for Y . However, t and u scores are calculated to be maximally correlated:

$$U = T + Error \quad (14)$$

Therefore, by substituting (14) into (13) the predicted (dependent) matrix Y can be computed as:

$$Y = TC^T + Error \quad (15)$$

Hence, the score matrix T that explains the variation in the independent matrix X , and is highly correlated with the dependent matrix Y , can be used to estimate both independent and dependent variables:

$$\hat{X} = TP^T \quad (16)$$

$$\hat{Y} = TC^T \quad (17)$$

The residual matrices for independent and dependent variables can be estimated as:

$$E_X = X - \hat{X} \quad (18)$$

$$E_Y = Y - \hat{Y} \quad (19)$$

The coefficients of determinations R^2 and Q^2 , Hotelling's T^2 statistic, and SPE can be determined as demonstrated in (6)-(9) (Chiang et al., 2001; Procensus, 2014; Wang, 2022). Moreover, as demonstrated in the previous section for dynamic PCA, dynamic PLS model can also be constructed by adding time-lagged variables to independent and dependent matrices (Dong & Qin, 2015).

3.5. Research Framework

In this study different multivariate statistical analysis methods have been used to detect leak in a shell and tube heat exchanger. After a detailed review of the related theory and literature, the following research framework has been developed to detect internal leak in the heat exchanger on a real-time basis.

First, the most related variables have been identified and selected to be used in data collection and modelling stages. To validate the effectiveness of the developed models, different offline models were first developed using operation historical data to detect an internal leak that occurred a few years ago. The data was collected and prepared for the selected variables. Four different offline models were then developed using PCA, dynamic PCA, PLS, and dynamic PLS multivariate statistical methods. However, due to the changes in the operational conditions in the past few years, the models could not be deployed online. Therefore, another set of offline models were built using historical data over the past

four months. As depicted in Fig. 9, the data was first collected and prepared for the selected variables, the prepared dataset was then used to build the offline models and the associated statistics, the developed offline models were then deployed online to detect the heat exchanger faults in real-time.

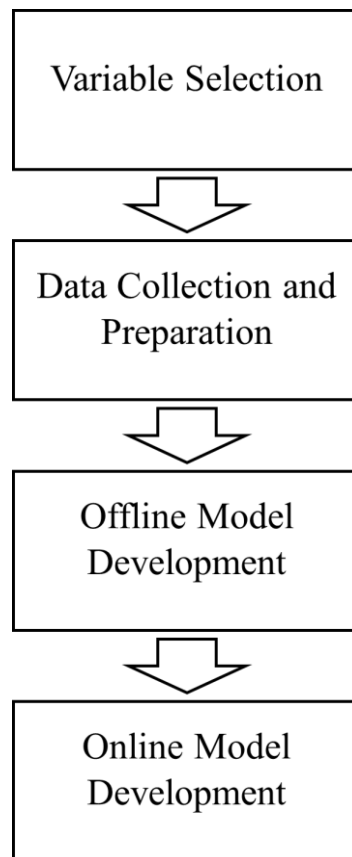


Fig. 9: Research Framework

3.6. Variable Selection

Following a detailed review of the Process Flow Diagrams (PFDs), Piping and Instrumentation Diagrams (P&IDs), and process description of the plant under study, all relevant measurements around the heat exchanger have been identified. Next, plant operators, process specialists, and process engineers were interviewed to identify the variables that are

expected to contribute the most to the heat exchanger internal leak. After understanding the process and conducting detailed interviews with plant personnel, the variables have been prioritized based on their criticalities and contributions to heat exchanger leak. Table (2) illustrates the priority level assigned to each variable considered in this study and shown in Fig. 6.

Table (2): Variables Priority Levels

Variable	Description	Priority	Variable	Description	Priority
TI01	Crude temperature	Unimportant	TI330	HSR temperature after E10	Important
TI301	Crude temperature before E10	Important	FC02	Crude flowrate	Unimportant
TI302	Crude temperature after E10	Important	FC14	Preflush steam temperature	Unimportant
TI314	Preflush feed temperature	Unimportant	FC18	Preflush reflux flow	Unimportant
TI315	Preflush feed temperature	Unimportant	FC19	HSR stripper steam temperature	Unimportant
TI316	Preflush bottom temperature	Unimportant	FC20	HSR flowrate after E10	Critical
TI323	HSR stripper feed temperature	Unimportant	PI03	Crude pressure	Unimportant
TI327	MSR temperature	Unimportant	LC02	HSR stripper level	Unimportant
TI329	HSR temperature before E10	Important	AI07	HSR color after E10	Critical

Critical and important variables then have only been considered for further modelling and analysis. Next, two new variables were created using existing variables. Crude and HSR streams temperatures before and after the heat exchangers were used to create crude and HSR temperature difference variables across the heat exchanger:

$$\Delta T1 = TI_{302} - TI_{301} \quad (20)$$

$$\Delta T2 = TI_{329} - TI_{330} \quad (21)$$

where TI_{301} and TI_{302} are crude temperature readings before and after heat exchanger E10, respectively, $\Delta T1$ is the temperature difference across the tube side of the heat exchanger and subscripts indicate the tag numbers. Similarly, TI_{329} and TI_{330} are HSR temperature readings before and after heat exchanger E10 respectively, and $\Delta T2$ is the temperature difference across the shell side of the heat exchanger. The two newly created variables were expected to contribute more in detecting the anomalies in the heat exchanger as they indicate the change in crude and HSR temperatures across the heat exchanger.

Therefore, the existing four temperature variables (TI_{301} , TI_{302} , TI_{329} , and TI_{330}) were replaced with the two variables developed in (20) and (21). Table (3) summarizes the final variables selected in this study to detect faults in Crude-HSR heat exchanger E10. The variables selected include crude temperature difference across the heat exchanger, HSR temperature difference across the heat exchanger, HSR color, and HSR product flow valve opening percentage.

Table (3): Final Selected Variables

#	Variable	Description	Unit of Measurement
1	$\Delta T1$	Crude temperature difference	Degree Fahrenheit (°F)
2	$\Delta T2$	HSR temperature difference	Degree Fahrenheit (°F)
3	AI7	HSR color	Saybolt
4	FC20	HSR product flow valve opening	Percentage (%)

3.7. Data Collection and Preparation

The objective of this study is to build a model that can differentiate between the heat exchanger's normal and abnormal conditions and detect

internal leak at an early stage. Therefore, to validate the effectiveness of the model in detecting internal leak, it needed to be trained and tested on a dataset that contains both normal operation data and data when a leak occurred. Thus, all maintenance records for the heat exchanger have been studied to identify the periods when the heat exchanger failed. The records demonstrated that two internal leak incidents have been reported for the heat exchanger E10 during the past few years. The data was collected for the selected variables for a one-year period prior to the leak (1st January 2015 to 9th December 2015) in 10-minute intervals from the refinery's historian software. The interval was selected carefully to capture the typical variations in the dataset and cover a large period. Next, the quality of data was checked by visually reviewing the datasets during both periods. The quality of data for the latest incident was poor due to high number of missing values and outliers. The latter is a common characteristic of industrial operation data. The other dataset, however, was found to be of good quality and has been used in this study.

The process of data cleansing and pre-processing was then carried out for the selected dataset. The records with missing and null values were first excluded from the dataset. Next, the normal operation range was specified for each variable after consultation with subject matter experts from the refinery. The outliers were then removed to avoid their impact in building the models. The resulting cleaned dataset included a total of 48,630 records. The variables, however, were found to be of different ranges as demonstrated in Table (4) and were normalized, i.e., subtracting the mean and dividing by the standard deviation for each variable.

Table (4): Summary of Variables used for Offline Analysis

#	Variable	Mean	Standard Deviation	Minimum	Maximum	Unit of Measurement
1	$\Delta T1$	63.48	2.75	54.36	75.59	°F
2	$\Delta T2$	313.83	10.07	284.17	354.40	°F
3	AI7	28.21	1.02	23.87	30.46	Saybolt
4	FC20	68.64	4.22	53.59	75.67	%

Due to the changes in the process operation over the past years, the developed offline models could not be used for online detection. Therefore, another dataset was collected from the refinery's historian for the previous four months (1st October 2021 to 28th January 2022) to build the online leak detection models. The data was collected in 10-minute intervals for the selected variables. Like the previous dataset, the data was first cleaned by removing outliers and missing values. Table (5) shows different statistical characteristics of the collected dataset. Finally, the cleaned dataset consisting of 17,227 records was normalized. The normalized dataset was used to build the leak detection models using static PCA and PLS. However, to build dynamic models adopting dynamic PCA and dynamic PLS, a 1-hour time lagged duplicate of each variable was added to the dataset.

Table (5): Summary of Variables used for Online Analysis

#	Variable	Mean	Standard Deviation	Minimum	Maximum	Unit of Measurement
1	$\Delta T1$	57.81	5.63	13.84	152.22	°F
2	$\Delta T2$	310.49	20.20	41.44	449.86	°F
3	AI7	29.91	0.36	5.78	30.29	Saybolt
4	FC20	12.98	3.61	0	71.57	%

3.8. Offline Model Development

As described in the previous sections of this chapter, two sets of offline models were developed. The first set were developed using a historical dataset that contained an internal leak incident in addition to normal operation data. These models were developed to assess models' performance in detecting heat exchanger internal leak occurred in the past. Another set of offline models were built using the data for the past four months for online deployment.

The pre-processed, normalized datasets were used to build offline leak detection models using four different algorithms. PCA, dynamic PCA, PLS, and dynamic PLS. PCA and dynamic PCA were adopted to detect heat exchanger E10 leak through Hotelling's T^2 and SPE charts. While analyzing different variables, it was observed that color analyzers can provide unreliable readings when used for long periods of time. Therefore, PLS and dynamic PLS algorithms were also implemented to predict HSR color and detect heat exchanger faults using Hotelling's T^2 and SPE charts. Models with different number of principal components have been developed for all four multivariate models. The performance of each model was evaluated using 10-fold cross validation. The models with the best performances have been selected for online deployment.

After building the models in MATLAB®, and to give the project a practical dimension, it was decided to use Aspen ProMV (version 12), a commercial data analytic platform from AspenTech to build the offline models in this research. The software provides monitoring charts with two different threshold limits (95% and 99% confidence limits). The values above these limits are considered anomalous and were used for fault detection. Different model interrogating plots such as the score plots, the loading plots, the biplots, and Variable Importance Plot (VIP)

have been used to evaluate the identified models. The software also identifies the main contributing variables to those anomalous values.

3.9. Online Model Development

Aspen ProMV online platform has been used to develop online models. The communication was first established between the refinery's historian and the online platform. Fig. 10 illustrates the data flow from the field instruments to the Aspen ProMV online software. The offline models were then imported into the online platform. Next, the variables were mapped with data historian tags.

The monitoring charts (Hotelling's T^2 and SPE) for fault detection were then customized. Two alarm levels with different confidence limits were created for each monitoring chart. 99%, and 99.9% confidence limits were selected as High alarm level (alarm with high severity), and High-High alarm level (alarm with very high severity), respectively for the T^2 statistic chart. For SPE chart, however, the High alarm level was selected at 97% confidence limit and the High-High alarm was set at 99.5% confidence limit. Following consultation with process specialists and engineers and to reduce the number of false alerts, a period of 2 hours above the threshold limits was specified for alarm generation. In other words, the online model was set to generate an alarm only if the values of the monitoring charts (Hotelling's T^2 and SPE) stay above the limits for 2 hours or more. Finally, the models were deployed online for real-time heat exchanger fault detection and identification. However, due to software limitations dynamic PLS could not be implemented online.

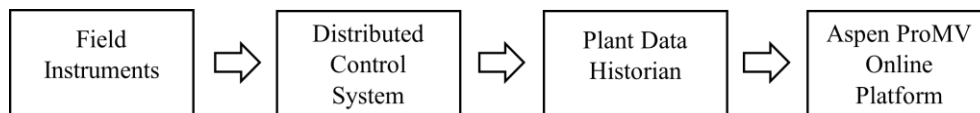


Fig. 10: System Data Flow

3.10. Conclusions

This chapter discussed the research methodology applied to detect internal leak in a selected heat exchanger. The formulation of research framework was explained, the process of variable selection and data collection and preparation was detailed, and finally the workflow to build the offline and online models were presented.

Chapter Four

Results and Discussion

4.1. Introduction

4.2. Results

4.3. Discussion

4.4. Conclusions

Chapter Four

Results and Discussion

4.1. Introduction

This chapter presents and discusses the findings of this research. It is organized into four sections. In the following section the results of both offline and online heat exchanger leak detection models are presented. Section two discusses the study findings in detail. Finally, the recommendations for future works and the conclusion from the overall findings are provided.

4.2. Results

In this section, first, the results of offline models developed on a historical dataset are presented. After that, the outcomes of the online leak detection models are provided.

4.2.1. Offline Models

As stated in the previous chapter, to assess the performance of the models in detecting heat exchanger leaks, four different models were trained and tested offline on a historical dataset that contained an internal leak incident. The developed models included static PCA, dynamic PCA, static PLS, and dynamic PLS.

4.2.1.1. Static PCA

Static PCA models with different principal components were developed and tested using 10-fold cross validation. Fig. 11 shows the coefficient of determination on the test dataset (cumulative Q^2) for static

PCA models with different principal components. The cumulative Q^2 increases from 0.47 for PCA model with a single principal component to 0.87 for PCA model with two principal components. However, after that the cumulative Q^2 increases in a lower rate to reach 0.94 and 0.96 for static PCA models with three and four principal components respectively.

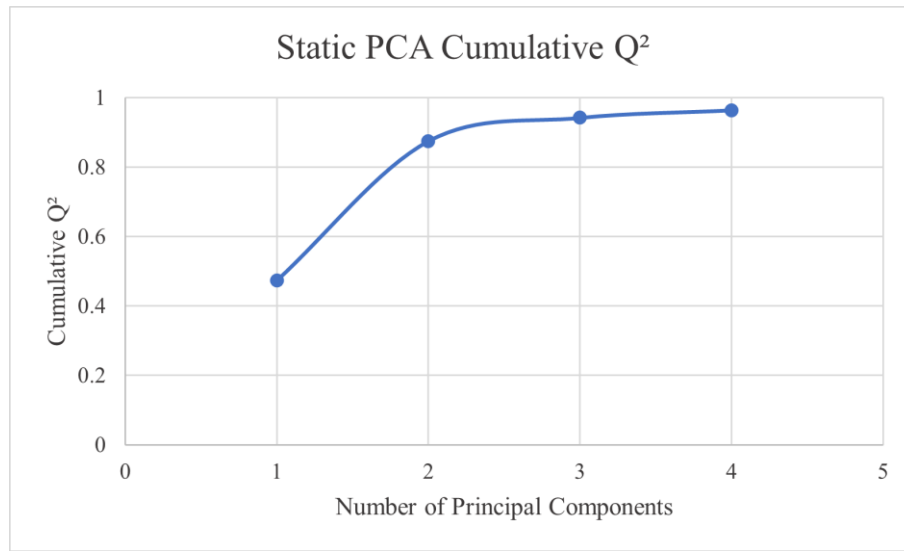


Fig. 11: Static PCA Cumulative Q^2 - Offline 1

The performance of static PCA model with two principal components in detecting heat exchanger's leak was evaluated using Hotelling's T^2 and SPE charts. As shown in Fig. 12, the T^2 statistic is below the 95% confidence limit from January to October 2015 except once (25th May 2015) when it exceeds the 99% limit. However, after that the T^2 statistic starts increasing gradually to exceed the 95% confidence limit on 20th and 28th October 2015 before it exceeds the 99% confidence limit on 20th November 2015. Hotelling's T^2 violation of confidence limit implies that the process variables have exceeded their normal variation limits that have been considered in the identified multivariate PCA model.

After reviewing the process variables, the alert on 25th May 2015 was found to be due to a process upset. However, there was no major

process upsets during October and November 2015. Further investigation illustrated that HSR color (AI7), and flow rate valve opening (FC20) variables are the main contributors to the violation of the Hotelling's T^2 95% and 99% confidence limits during October and November 2015. Fig. 13 demonstrates that during this period HSR color (AI7) and flow rate valve opening (FC20) are lower comparing to normal operating period. The process experts confirmed that this is attributed to an internal heat exchanger's leak. When the leak occurs, the crude that mixes with HSR product causing the HSR product color to change and simultaneously increases its flow rate. Due to the increased flow rate, the product flow valve is requested by the PID controller to close in order to maintain the HSR product flow rate at its specified set point. The Hotelling's T^2 statistic generated alerts with 95% confidence limit on 20th October 2015 and alerts with 99% confidence limit on 20th November 2015, while the leak was found on 9th December 2015 by plant operators at the refinery. In other words, the model provided alerts with 95% and 99% confidence limits, 50 and 19 days in advance respectively.

SPE chart that shows the deviation between the identified model and actual process, however, exceeds the 95% and 99% confidence limits several times as depicted in Fig. 14. Following the review of the dataset, the violation of the confidence limits was found to be due to process upsets in the form of spikes. Although this chart does not provide early warnings on the heat exchanger leak, it shows the overall status of the model.

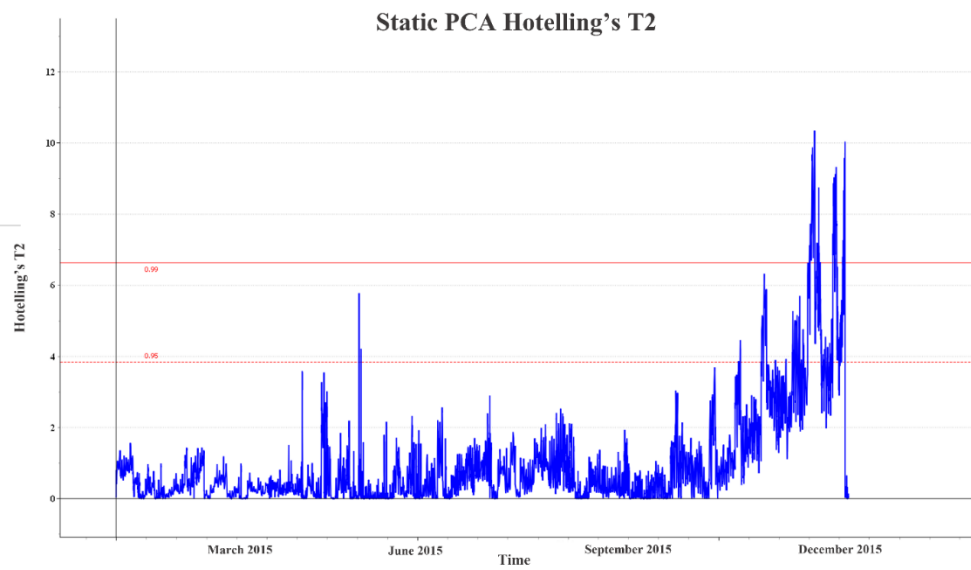


Fig. 12: Static PCA Hotelling's T^2 - Offline 1

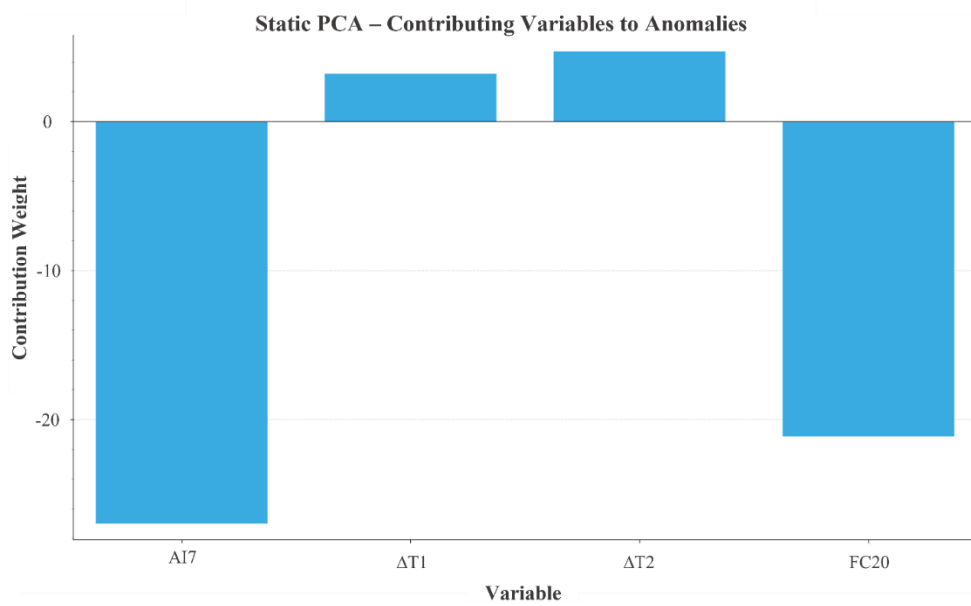


Fig. 13: Static PCA Variables Contribution

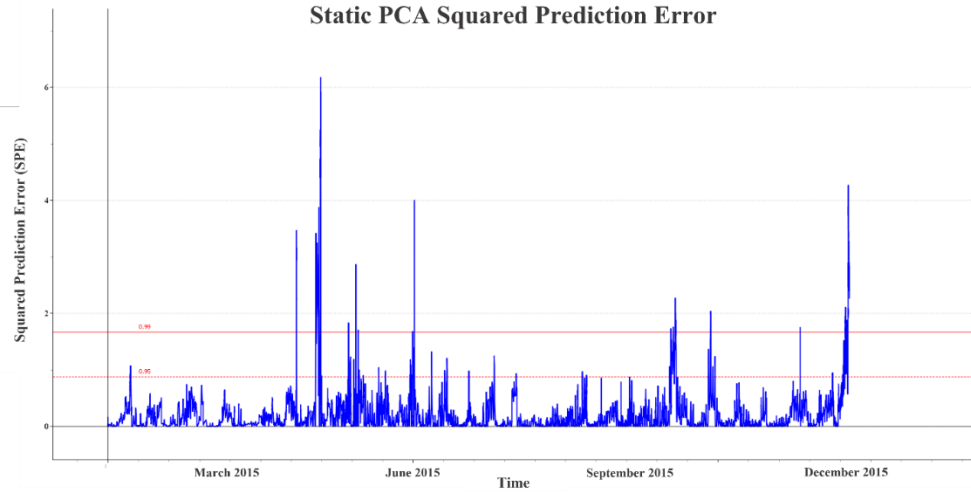


Fig. 14: Static PCA Squared Prediction Error – Offline 1

4.2.1.2. Dynamic PCA

Like static PCA, the performance of several dynamic PCA models with different number of principal components was evaluated in terms of cumulative Q^2 . 10-fold cross validation was used to train and test each model. While dynamic PCA model with a single principal component attained a cumulative Q^2 of 0.47, this value has increased to 0.89 for a dynamic PCA model with two principal components. After that, the cumulative Q^2 increases in a slower manner to reach 0.95, 0.97, 0.98, and 0.98 for dynamic PCA models with three, four, five, and six principal components respectively as shown in Fig. 15.

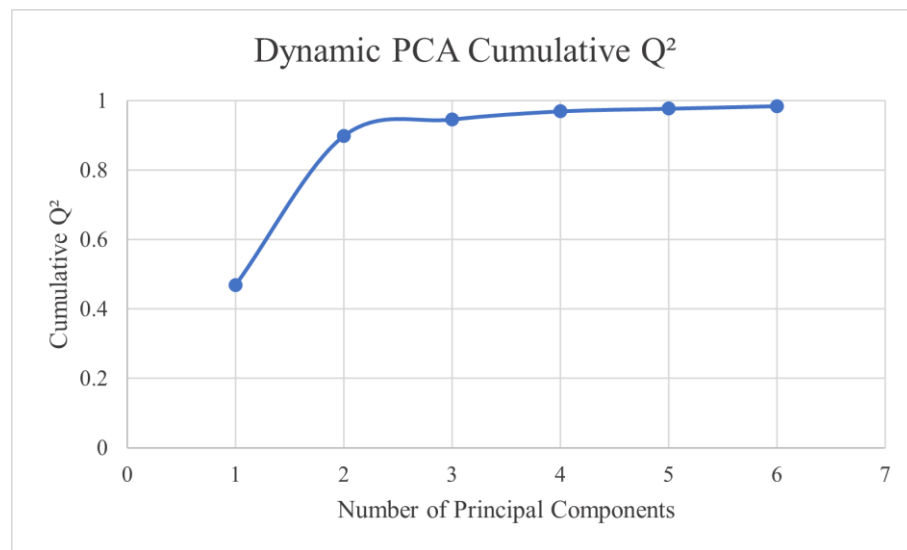


Fig. 15: Dynamic PCA Cumulative Q^2 - Offline 1

Hotelling's T^2 and SPE charts were then used to assess the performance of dynamic PCA model with two principal components in detecting heat exchanger internal leak. As illustrated in Fig. 16, the T^2 statistic maintains a value below the 95% confidence limit from January till October 2015, after that however, the value of T^2 statistic starts increasing to exceed the 95% first on 20th October 2015 then on 30th October 2015, and 13th November 2015 before it violates the 99% confidence limit on 20th November 2015. After that, the value of the Hotelling's T^2 statistic mostly remains above the 95% confidence limit until the leak was found by plant operators on 9th December 2015 when the heat exchanger was isolated.

As demonstrated in Fig. 17, HSR product color (AI7), HSR product color with 1-hour time lagged data (Lagged AI7), HSR product flow valve opening (FC20), and HSR flow valve opening with 1-hour time lagged data (Lagged FC20) have changed the most comparing to normal operation data and have been identified as the main contributors to identify the leak.

SPE on the other hand as depicted in Fig. 18 exceeds the confidence limits several times because of process upsets, and unlike T^2 statistic, the SPE chart does not provide early leak detection for the heat exchanger.

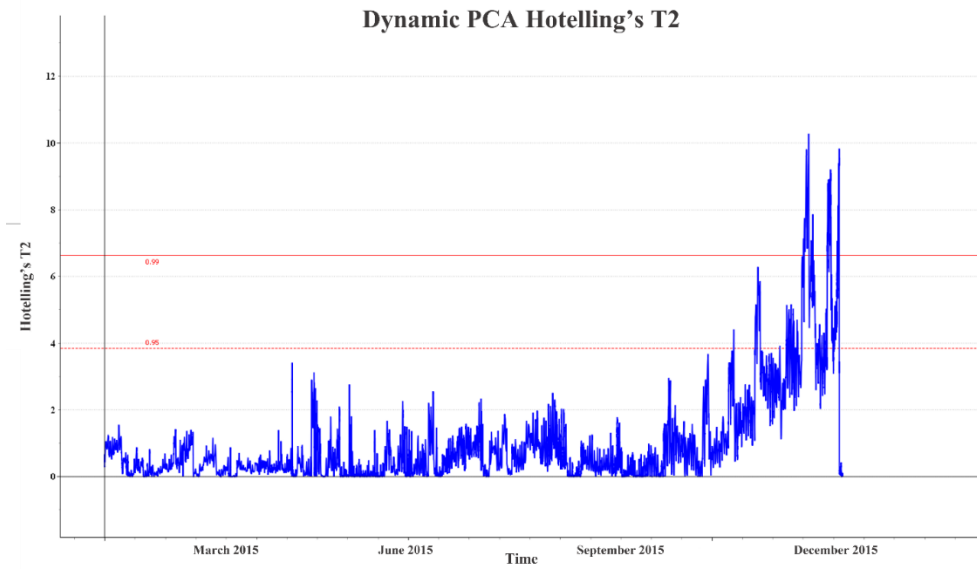


Fig. 16: Dynamic PCA Hotelling's T² - Offline 1

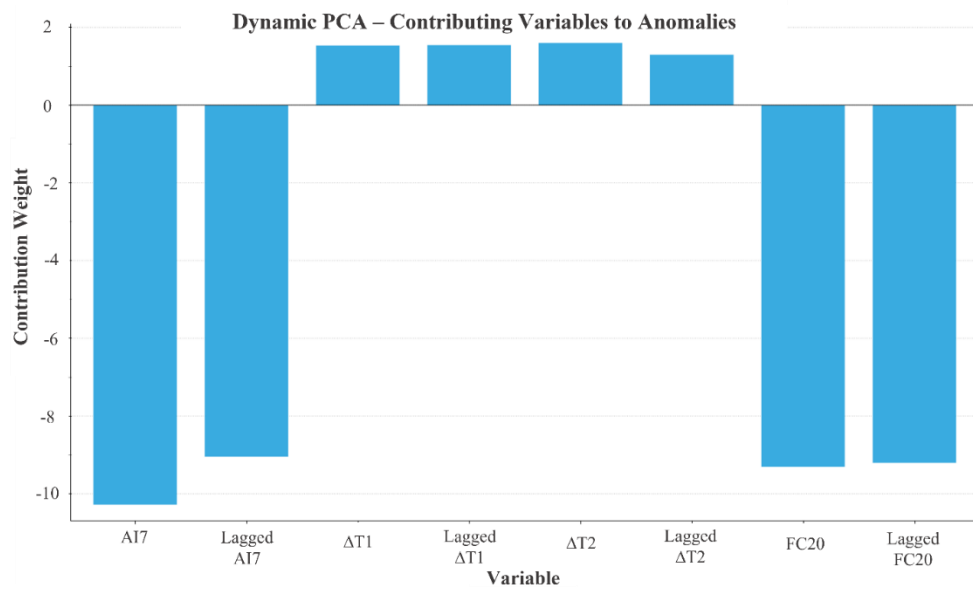


Fig. 17: Dynamic PCA Variables Contribution

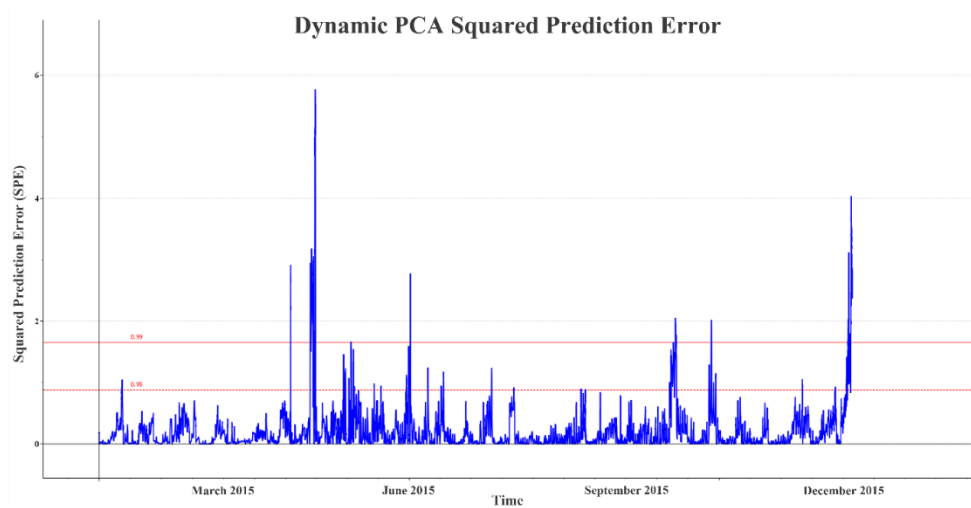


Fig. 18: Dynamic PCA Squared Prediction Error – Offline 1

4.2.1.3. Static PLS

Static PLS models with different number of principal components were developed to detect the heat exchanger leak and predict HSR product color. These models consisted of three input variables (ΔT_1 , ΔT_2 , and FC20) and a single output variable (AI7). The results show that for input variables the cumulative Q^2 starts at 0.31 for PLS model with a single component and increases to 0.59 and 0.95 for models with two and three components respectively as demonstrated in Fig. 19. However, the value of cumulative Q^2 for the output variable remains at 0.41 for models with one, two, and three principal components as shown in Fig. 20.

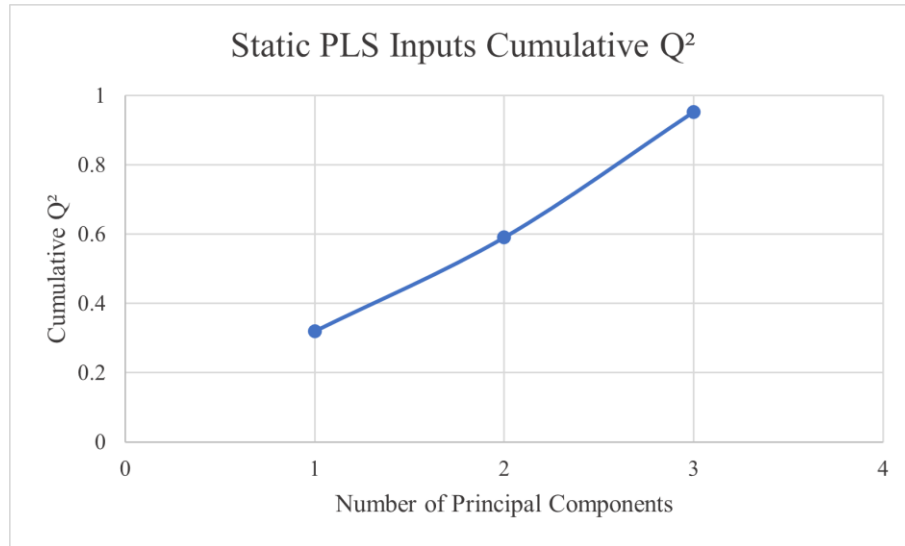


Fig. 19: Static PLS Inputs Cumulative Q^2 - Offline 1

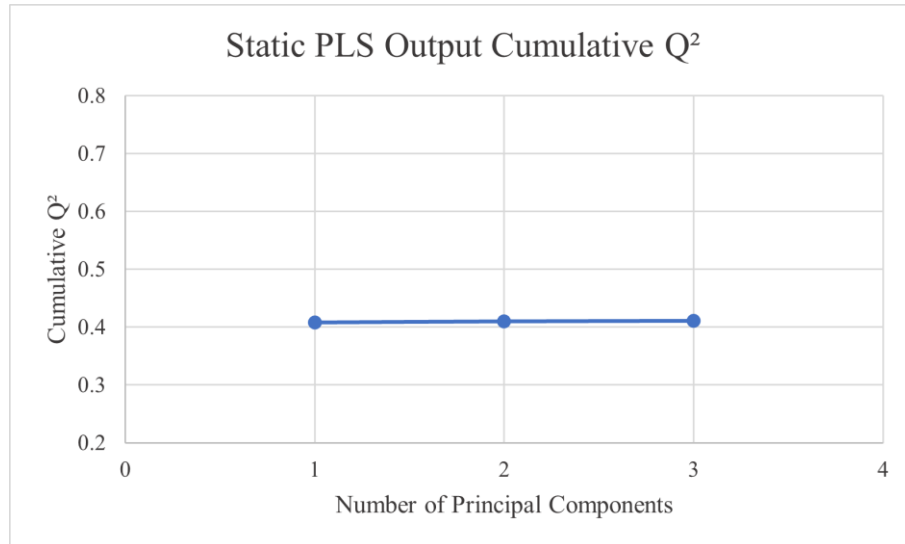


Fig. 20: Static PLS Outputs Cumulative Q^2 - Offline 1

Low cumulative Q^2 for the input variables with two principal components resulted in static PLS's poor performance in detecting heat exchanger leak where Hotelling's T^2 and SPE charts are unable to provide early indications as illustrated in Fig. 21 and Fig. 22 respectively. Furthermore, static PLS with output Q^2 of 0.41 and test Root Mean Squared Error (RMSE) of 0.79 performed poorly in predicting the HSR product color as depicted in Fig. 23.

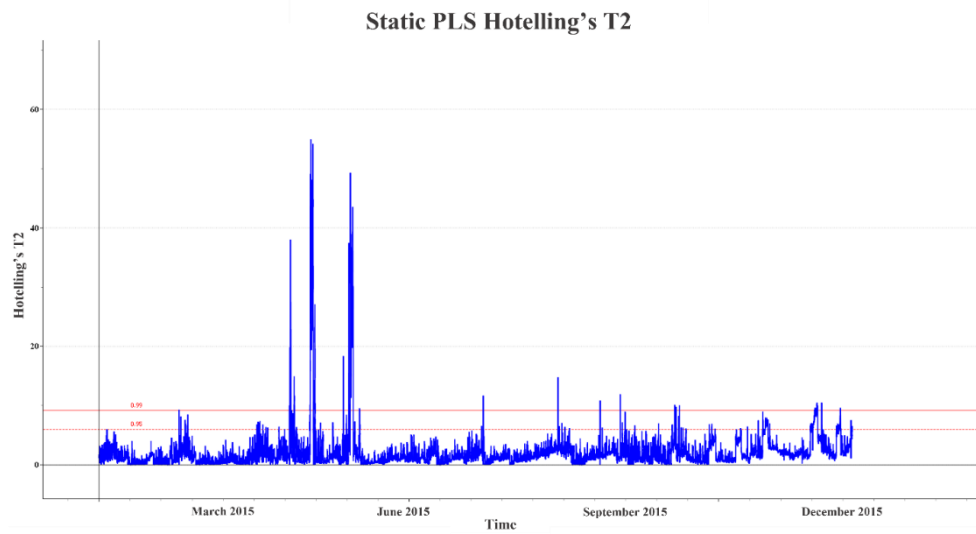


Fig. 21: Static PLS Hotelling's T^2 - Offline 1

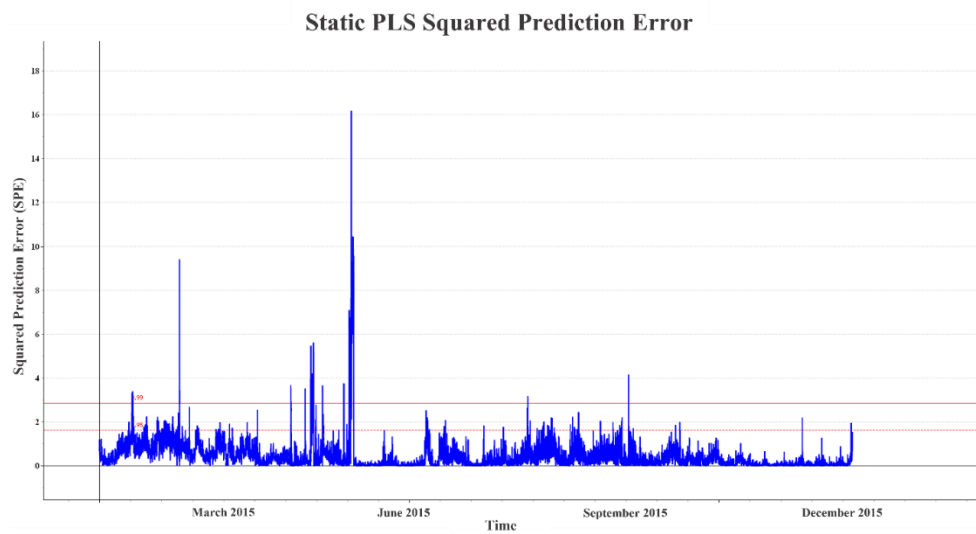


Fig. 22: Static PLS Squared Prediction Error – Offline 1

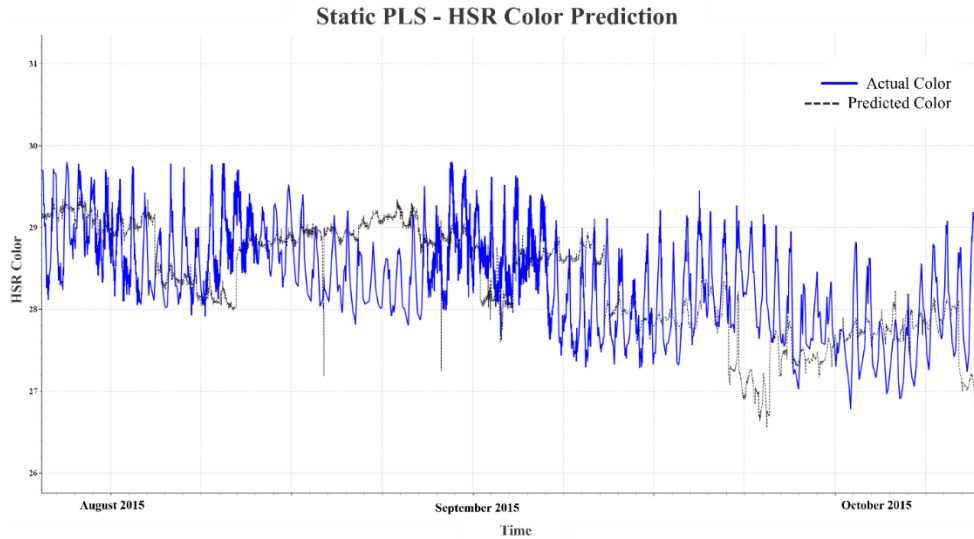


Fig. 23: Static PLS Color Prediction – Offline 1

4.2.1.4. Dynamic PLS

To build dynamic PLS models, variables with 1-hour time lag were added to the input variables used for static PLS to detect the heat exchanger leak and predict HSR product color. The performance of several dynamic PLS models were assessed using different number of principal components. As demonstrated in Fig. 24, the value of cumulative Q^2 for the input variables increases from 0.58 for dynamic PLS model with a single component to 0.84, and 0.98 for dynamic PLS model with two and three components respectively. After that however, the cumulative Q^2 remains at 0.99 for models with four, five, and six components. Furthermore, the output variable cumulative Q^2 starts at a high value of 0.96 for dynamic PLS model with one component and increases to 0.98 for a model with two components. Following that however, the value of cumulative Q^2 stayed at 0.98 for models with additional principal components as shown in Fig. 25.

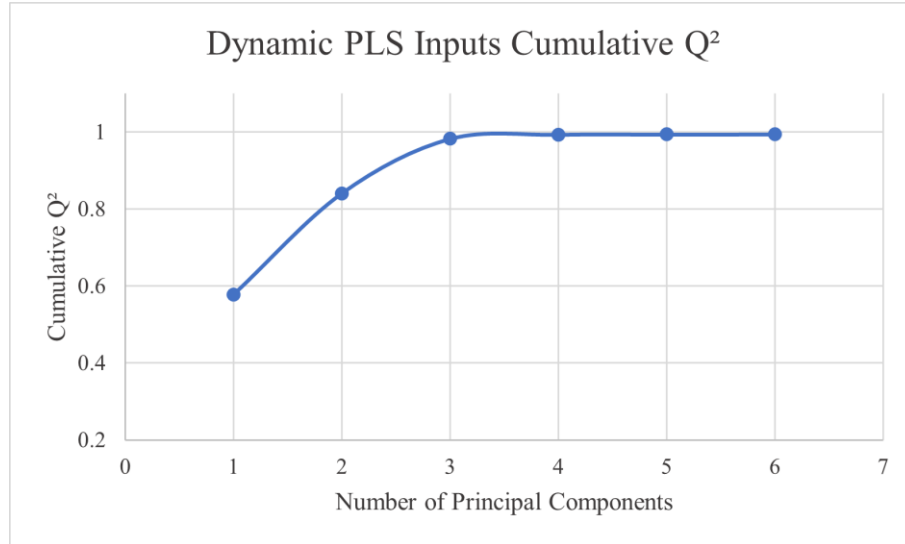


Fig. 24: Dynamic PLS Inputs Cumulative Q^2 - Offline 1

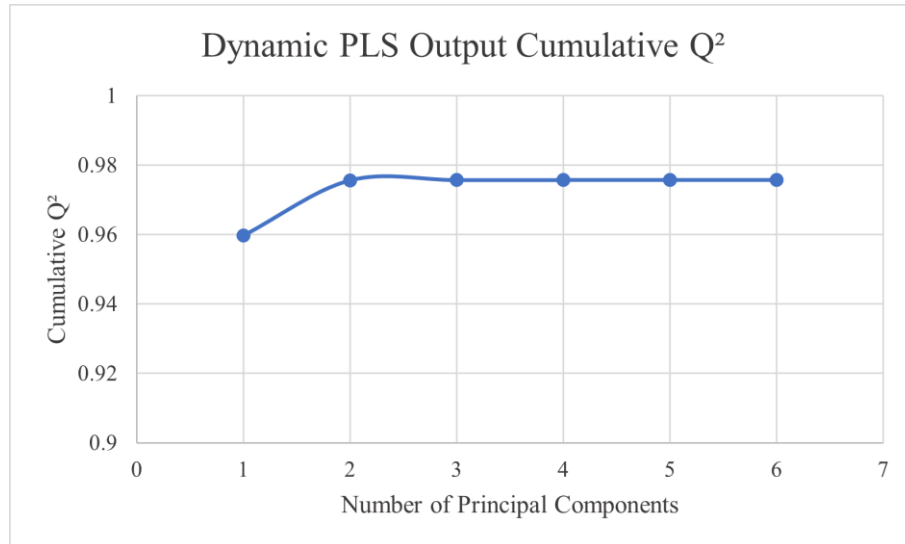


Fig. 25: Dynamic PLS Outputs Cumulative Q^2 - Offline 1

Hotelling's T^2 chart in Fig. 26 shows that unlike static PLS model, dynamic PLS model can detect the leak at an early stage. The T^2 statistic remains below the 95% confidence limit for the period from January to October 2015 except on 25th May 2015 where the value exceeds the 95% confidence limit due to process disturbances. Following that however, the value of Hotelling's T^2 increases to first violate the 95% confidence limit on 20th October 2015. After that, this value continues rising to pass the

99% confidence limit on 20th November 2015. Afterwards, T^2 statistic remains above the 95% confidence limit most of the time until the leak was found, and the equipment was isolated. Like PCA and dynamic PCA, SPE chart shown in Fig. 27 is mostly affected by process upsets and does not provide clear fault detection indications. Moreover, dynamic PLS with two principal components with an output cumulative Q^2 of 0.98 and a test RMSE of 0.16 has also performed well in predicting the HSR product color as shown in Fig. 28.

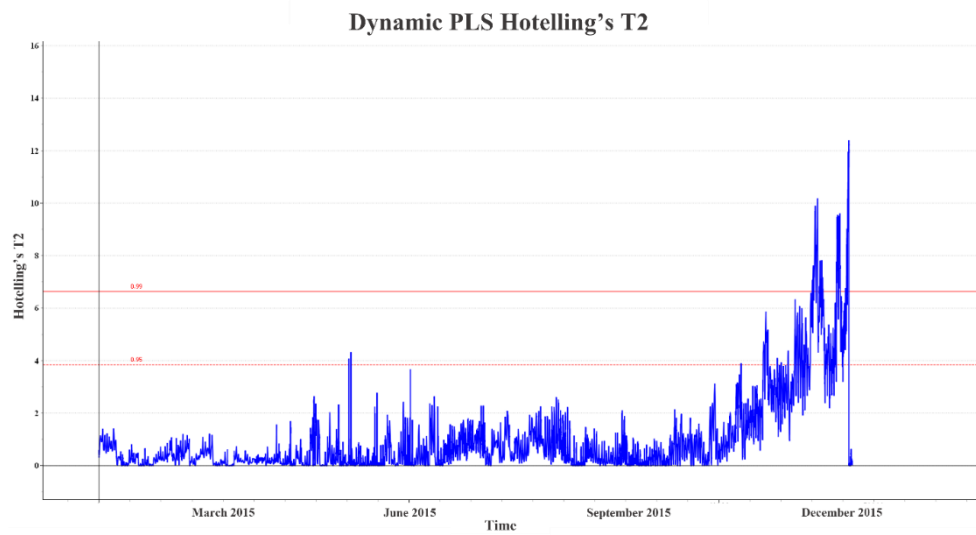


Fig. 26: Dynamic PLS Hotelling's T^2 - Offline 1

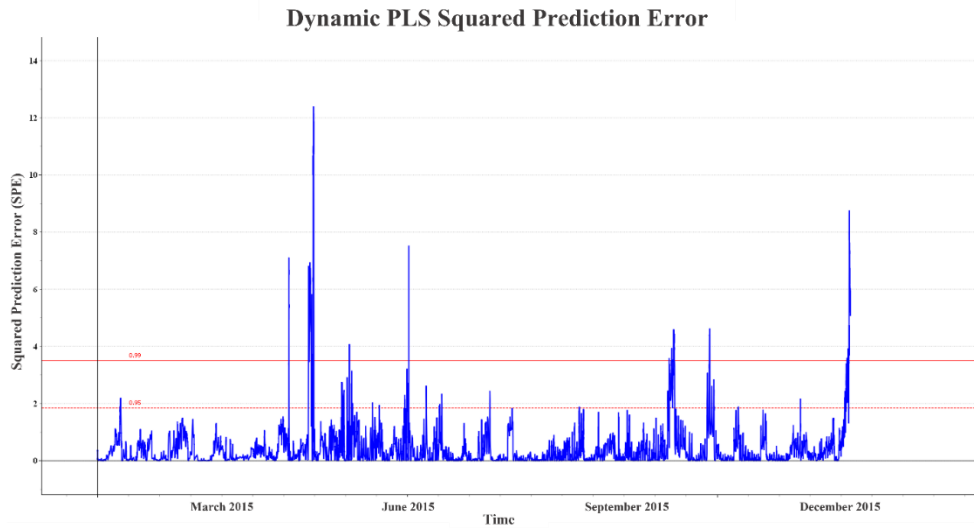


Fig. 27: Dynamic PLS Squared Prediction Error – Offline 1

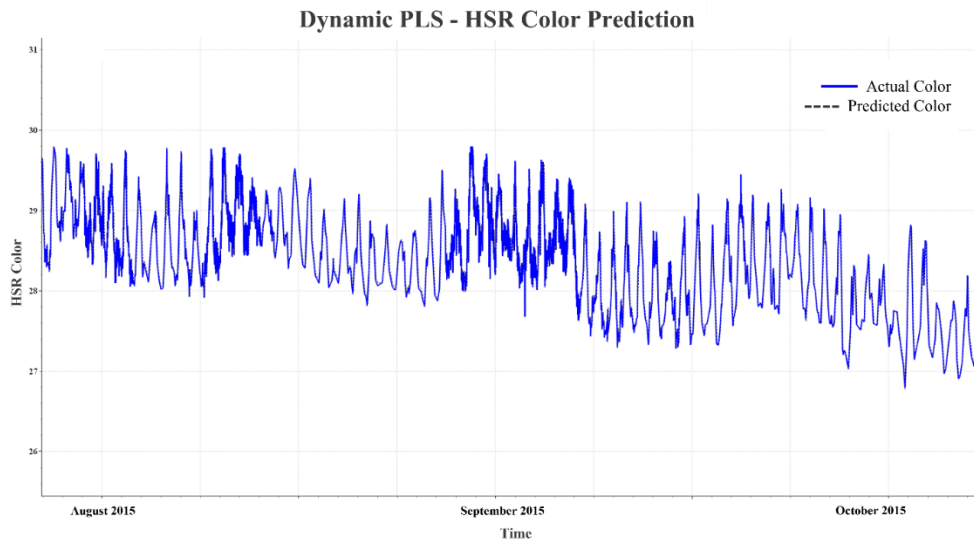


Fig. 28: Dynamic PLS Color Prediction – Offline 1

4.2.2. Online Models

After evaluating the performance of the offline models developed using historical data (1st January 2015 to 9th December 2015), another set of offline models were developed using the past four months data (1st October 2021 to 28th January 2022). The offline models then were deployed online for real time heat exchanger leak detection. This section presents the results of both offline and online models.

4.2.2.1. Offline Models

The performance of offline static and dynamic PCA models were evaluated in terms of accuracy and false alarms generation. The data used to develop the models included a few process upsets without any heat exchanger leaks.

- **Static PCA**

Fig. 29 shows the cumulative Q^2 for static PCA models with different number of principal components. The value of cumulative Q^2 increases from 0.50 for static PCA model with one principal components to 0.82 and 0.94 for static PCA models with two and three components respectively. This value further increases to 0.98 for static PCA model with four principal components.

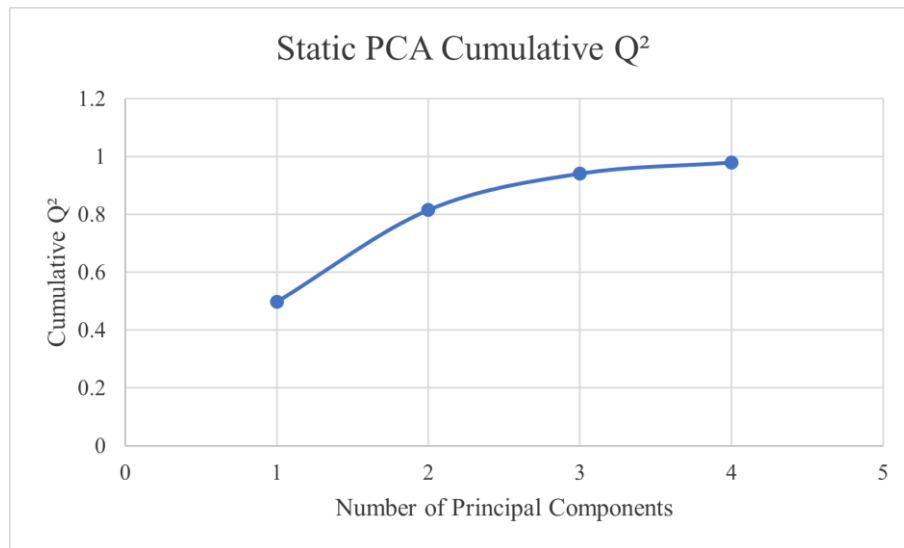


Fig. 29: Static PCA Cumulative Q^2 - Offline 2

Hotelling's T^2 and SPE charts for static PCA model with two components in Fig. 30, and Fig. 31 respectively, demonstrate that both signals remain below the 95% confidence limit most of the time except on a few occasions due to process disturbances. The results also illustrate that the violation of the confidence limits is in the form of spikes

returning below the 95% confidence limit when the process condition is normal again.

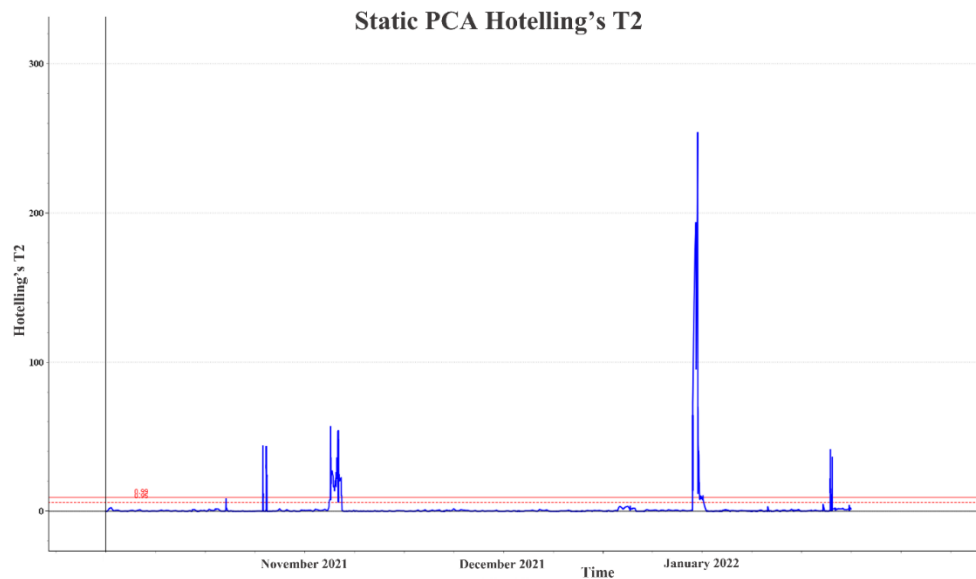


Fig. 30: Static PCA Hotelling's T^2 - Offline 2

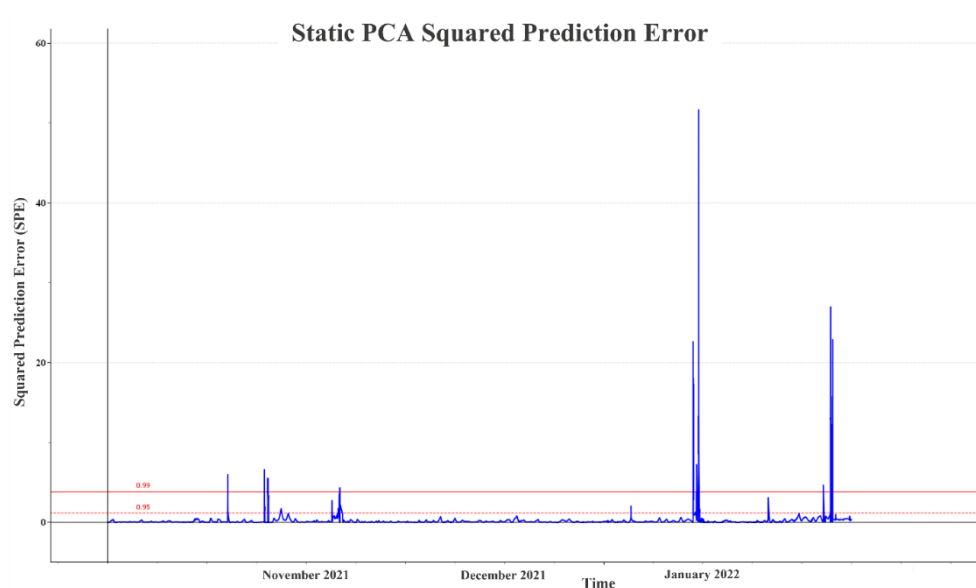


Fig. 31: Static PCA Squared Prediction Error – Offline 2

- **Dynamic PCA**

The results of dynamic PCA models with various number of principal components illustrated in Fig. 32 shows that the value of cumulative Q^2 is 0.45 for the model with one component. It then increases to 0.83 for the model with two principal components. After that this value

raises to 0.88, 0.92, 0.95, and 0.97 for dynamic PCA models with three, four, five, and six principal components respectively.

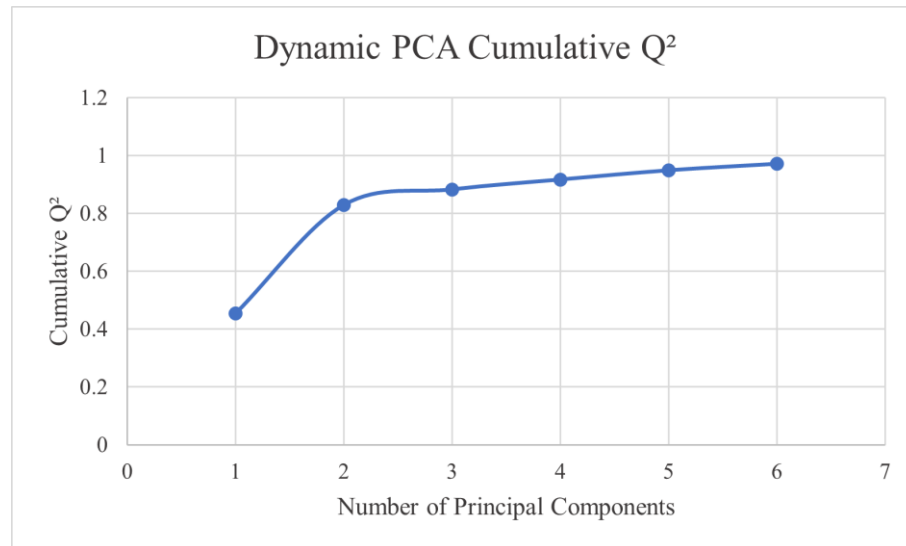


Fig. 32: Dynamic PCA Cumulative Q^2 - Offline 2

Like static PCA, T^2 statistic and SPE for dynamic PCA model with two principal components exceed the confidence limits as spikes only when there is a process upset. Otherwise, these values fall below the 95% confidence limit as shown in Fig. 33 and Fig. 34.

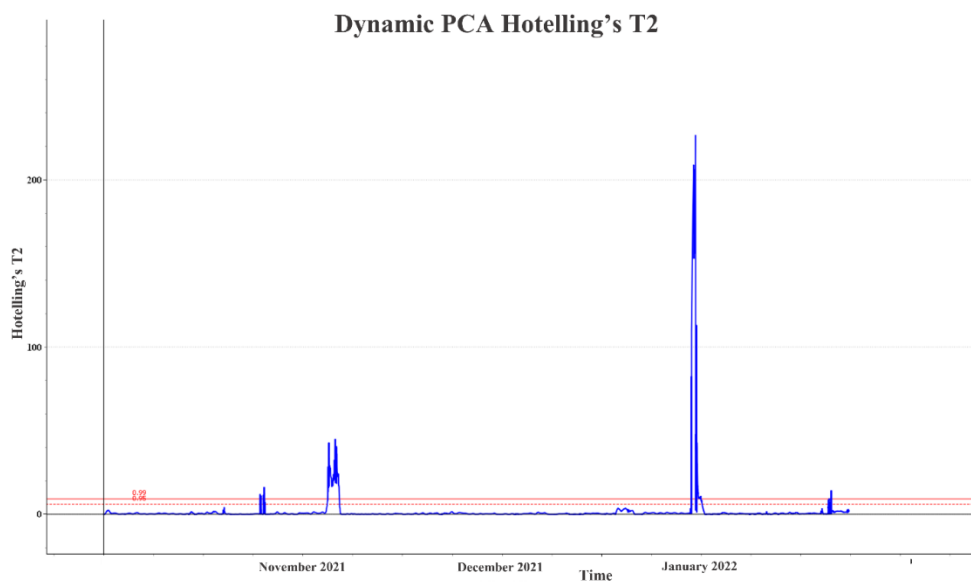


Fig. 33: Dynamic PCA Hotelling's T^2 - Offline 2

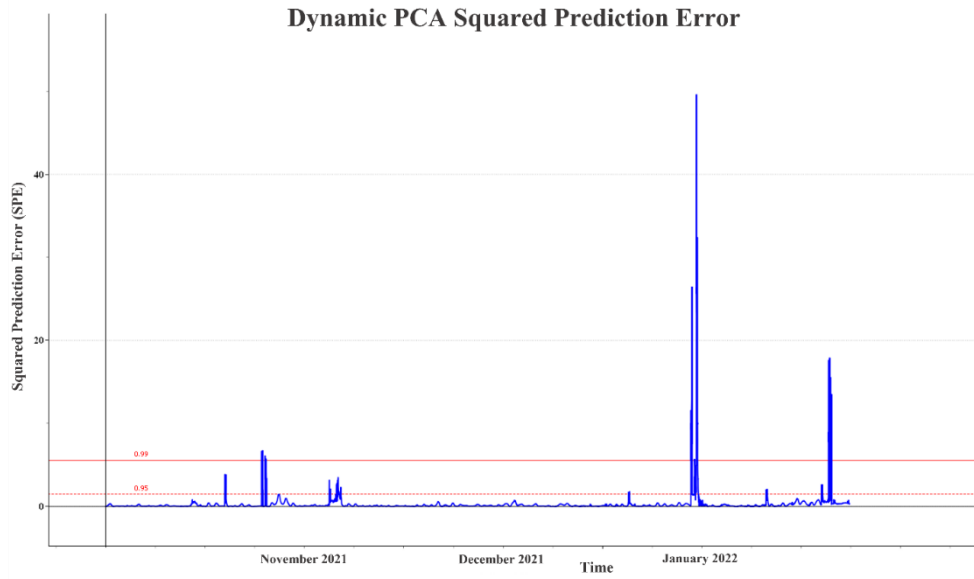


Fig. 34: Dynamic PCA Squared Prediction Error – Offline 2

4.2.2.2. Online Models

After developing offline static and dynamic PCA models, the models with two principal components were implemented online to detect heat exchanger leak on a *real-time* basis. The online models have been monitoring the heat exchanger's performance for a period of one month from 3rd February 2022 to 3rd March 2022. T^2 statistic and SPE charts have been used for online leak detection and to monitor the performance of the models.

- **Static PCA**

Fig. 35 shows the T^2 statistic chart with High alarm level (99% confidence limit) at 9.21 and High-High alarm level (99.9% confidence limit) at 13.82. Fig. 36 however, shows the SPE chart with High alarm level (97% confidence limit) at 2.51 and High-High alarm level (99.5% confidence limit) at 9.69. As stated in the previous chapter, the alarms were configured to be generated only if the value of Hotelling's T^2 or SPE remains above the alarm levels for a period of 2 hours or more. Fig. 35 and Fig. 36 demonstrate that the value of both T^2 statistic and SPE remain below the alarm levels for the entire one-month period except on

two occasions where they exceed the alarm levels in the form of spikes for a short period due to some process disturbances. However, there was no alarm generated since the period is less than 2 hours. This indicates that the heat exchanger is free of any leak during this period.

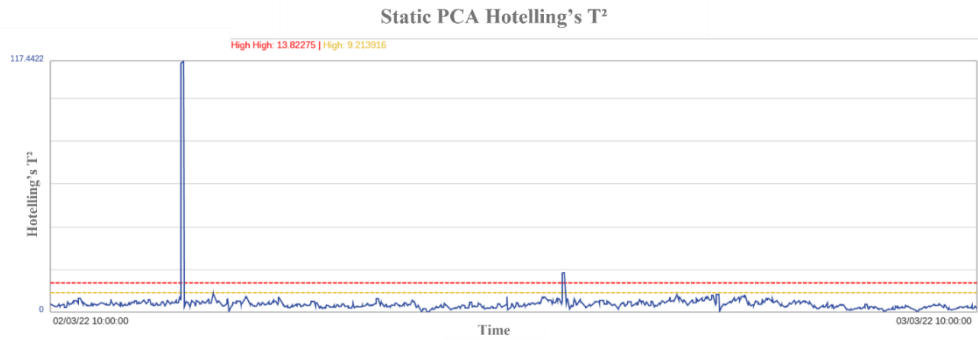


Fig. 35: Static PCA Hotelling's T^2 - Online (real plant data)

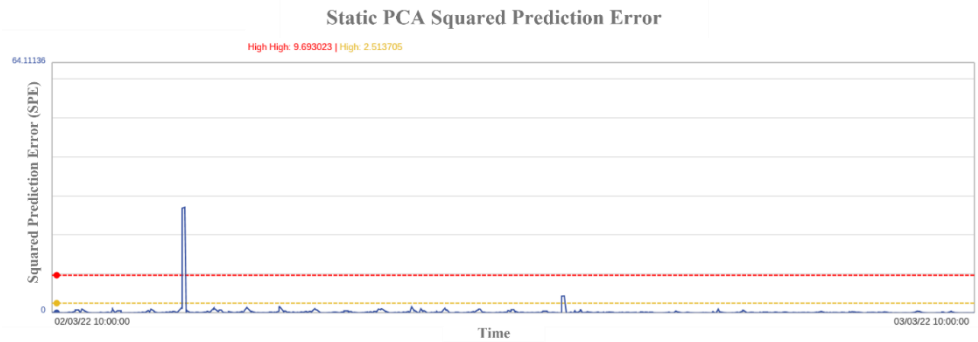


Fig. 36: Static PCA Squared Prediction Error – Online (real plant data)

- **Dynamic PCA**

Online dynamic PCA model also has been deployed for real-time heat exchanger leak detection. Hotelling's T^2 chart with High alarm at 9.21 and High-High alarm at 13.82 as well as SPE chart with High alarm at 2.83 and High-High alarm at 9.29 have been used to monitor the heat exchanger's health. As illustrated in Fig. 37 and Fig. 38, both T^2 statistic and SPE values are below the alarm limits for the entire one-month monitoring period except twice that due to process upsets violate the alarm limits. The model, however, did not generate any alarms

throughout the monitoring window because neither of the two signals exceeded the alarm levels for 2 hours or more.

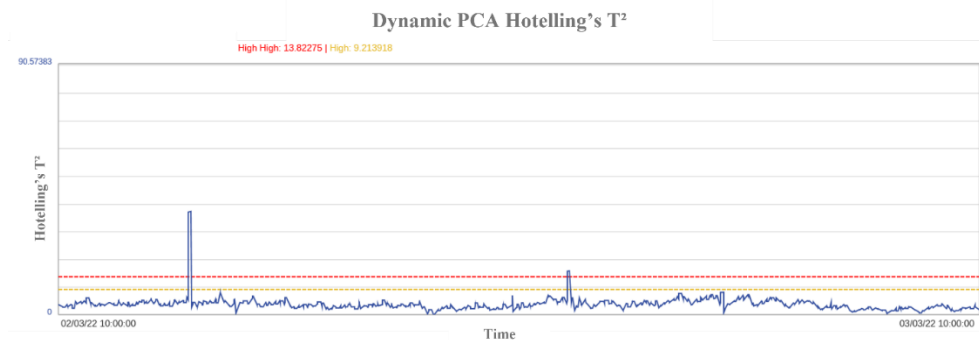


Fig. 37: Dynamic PCA Hotelling's T^2 - Online

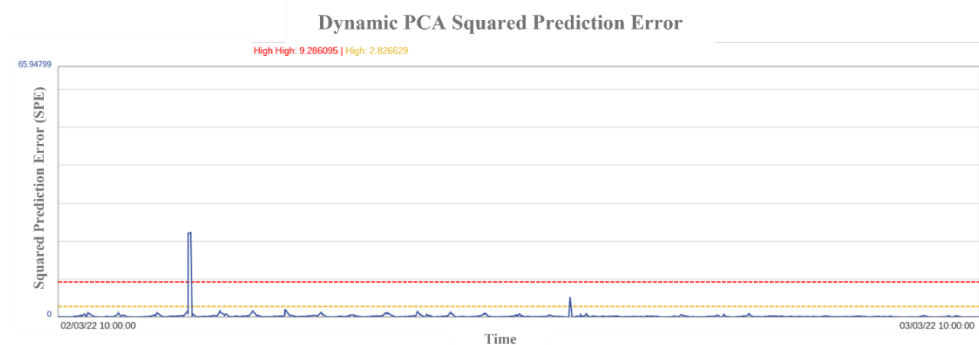


Fig. 38: Dynamic PCA Squared Prediction Error – Online

4.2.3. Economic Benefits

The proposed real-time heat exchanger leak detection system can result in several economic advantages for the local refinery. Due to heat exchanger's tube leak, crude oil mixes with the product resulting in degradation in its quality and specifications. If the leak is not detected at an early stage, the substandard product can join and contaminate the overall products' specifications in a large downstream tank. At worst, the entire product in the tank needs to be reprocessed which results in huge losses. The reprocessing cost of the product in the tank can vary depending on the refinery's operation cost which changes depending on

several factors. This cost for an average of 50,000 barrels of product in the downstream tank can vary from 250,000 United States Dollars (USD) for operations cost of 5 USD/barrel to 350,000 USD for operations cost of 7 USD/barrel as demonstrated in Fig. 39. The proposed system helps mitigate potential losses by offering early leak detection.

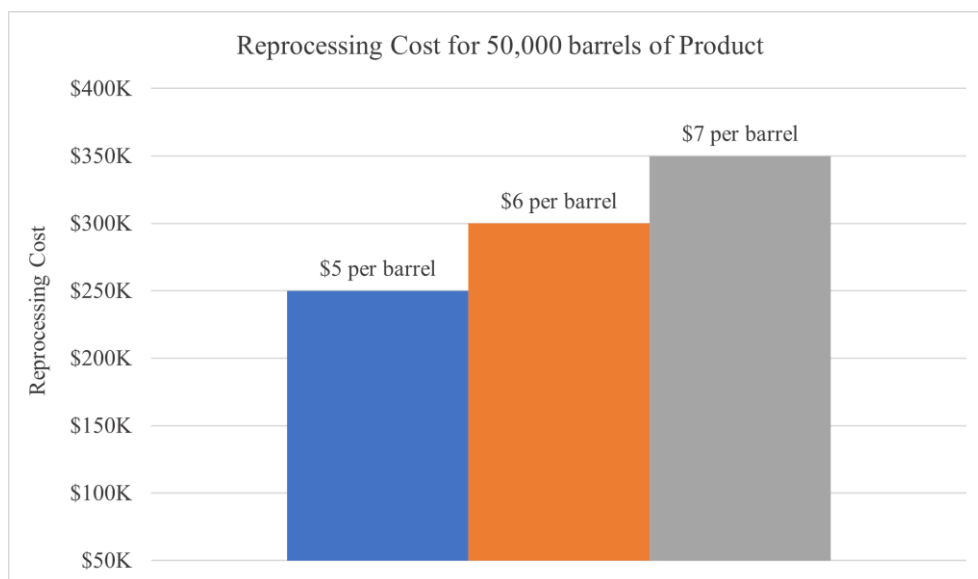


Fig. 39: Reprocessing Cost for 50,000 barrels of Product

Furthermore, when a leak is identified, heat exchangers are isolated for repair and as a result the unit feed is reduced by 10,000 to 15,000 barrels per day during the downtime. This can result in approximately 20,000 USD to 90,000 USD Loss Profit Opportunity (LPO) per day, depending on the reduced feed amount and refinery's net profit margin as illustrated in Fig. 40. Detecting the heat exchanger leak at an early stage allows the maintenance team to formulate a plan and ensure availability and functionality of necessary spare parts. Consequently, the heat exchanger can be restored in a timely manner which minimizes the potential downtime period.

Another benefit of the proposed online leak detection system is that an increased awareness of asset health through rigorous models will

require reduced expenditure on traditional preventive maintenance activities.

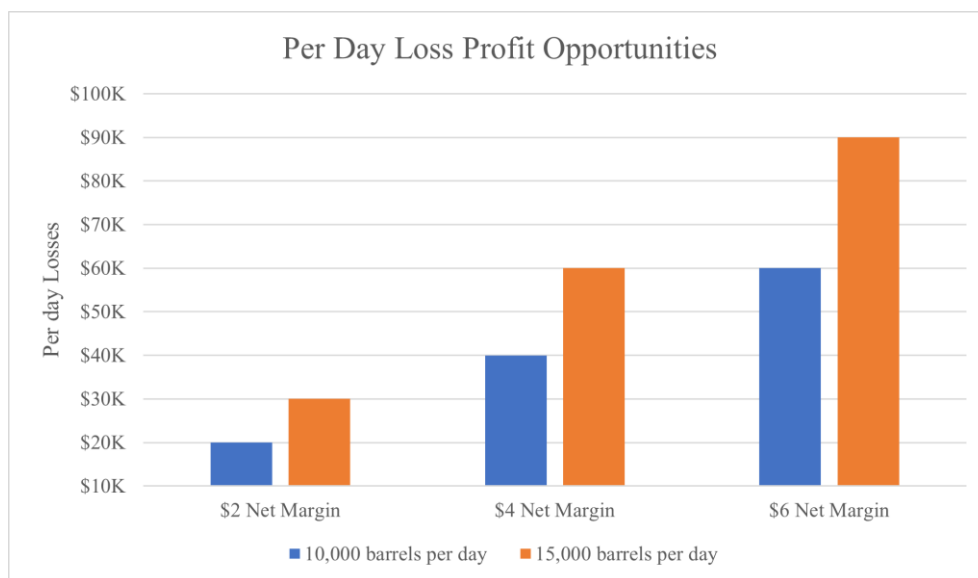


Fig. 40: Per Day Loss Profit Opportunities

4.3. Discussion

In this study four different multivariate statistical approaches specifically static PCA, dynamic PCA, static PLS, and dynamic PLS were adopted to detect a heat exchanger leak in a local refinery.

To assess the performance of the models in detecting leaks, a set of offline models were first developed to detect a heat exchanger leak that occurred in the past using historical data. The results demonstrated that static PCA, dynamic PCA, and dynamic PLS models were able to detect the leak successfully at early stages. The three models performed identically in giving early warnings, generating alerts with 95% confidence limit 50 days and alerts with 99% confidence limit 19 days before the leak was discovered by operators at the facility. The models performed better comparing to least square algorithm implemented by Sun et al. (2002) that detected the leak in a steam-water boiler systems 16, 27, and 13 days before it was found by operators for three different cases.

Furthermore, the models by producing zero or one false alarm performed better compared to the leak detection model proposed by Sun et al. (2002) which have suffered from generating many false alerts.

Moreover, dynamic PCA without generating any false alarms throughout the entire one-year period performed better than static PCA that generated a false alarm on 25 May 2015. This is in line with the study conducted by Choi and Lee (2004) that demonstrated dynamic PCA produces a smaller number of false alarms comparing to static PCA models. In terms of accuracy also dynamic PCA by achieving a cumulative Q^2 of 0.89 for a model with two principal components and 0.95 for a model with three principal components performed better than static PCA that resulted in a cumulative Q^2 of 0.87 and 0.94 for models with two and three principal components respectively. This is also consistent with previous studies (Ku et al., 1995; Chen & Liao, 2002) which illustrated dynamic PCA models achieve a higher accuracy comparing to static PCA models.

Static PLS with an input cumulative Q^2 of 0.59 and an output cumulative Q^2 of 0.41 for a model with two components failed to detect the heat exchanger leak and predict the HSR product color accurately. Dynamic PLS on the other hand, with an input cumulative Q^2 of 0.84 and an output cumulative Q^2 of 0.97 for a model with two components detected the leak and achieved a high product color prediction accuracy. The poor performance of the static PLS model could be due to the absence of HSR product color that is a main indication of the leak in the input variables. In other words, the time-lagged variables included in the dynamic PLS model improved its performance and resulted in achieving better results. However, dynamic PCA model achieved a higher accuracy and a smaller number of false alarms in comparison with dynamic PLS model which is contrary to the study adopted by Heo and Lee (2018).

As a result of changes in refinery's process operations over the past few years the developed offline models could not be implemented online. Moreover, due to the poor performance of static PLS model and software's limitation in deploying the online dynamic PLS model, both static and dynamic PLS models were excluded from the online implementation. Therefore, another set of offline static and dynamic PCA models were built using the recent few months historical data. The developed models then were deployed online to monitor and detect heat exchanger leak in real-time for a period of one month.

Like the previous offline analysis, dynamic PCA outperformed static PCA in terms of both accuracy and number of generated false alarms. However, both static and dynamic PCA models generated more false alerts. This is due to higher number of process upsets during this period comparing to the selected window for the previous offline analysis.

Furthermore, the results of online leak detection models that monitored the heat exchanger performance for a period of one month demonstrated that both static and dynamic PCA models without generating any false alarms have efficiently reflected the actual healthy status of the heat exchanger. Unexpected process disturbances resulted in violation of High and High-High alarm limits on two occasions for a few minutes. However, the models were configured to generate alarms only if the values of T^2 statistic or SPE remained above the thresholds for 2 hours or more. Moreover, as offline analysis illustrated, process disturbances have had a larger impact on online static PCA compared to dynamic PCA model.

The economic benefit analysis presented in the previous section shows that accurate online heat exchanger leak detection models can result in huge savings for the local refinery under study. However, the amount of these savings depends on several factors. First, how soon the

model detects the heat exchanger leak? The earlier the model identifies the leak and notifies plant operators, the less the negative impacts of the leak. This also allows the maintenance team to arrange necessary spare parts and develop an effective repair plan resulting in shortening the downtime duration and more savings. Other factors include the refinery's operations cost and net profit margin. Heat exchanger leak can result in huge losses when the operations cost and net profit margins are high. Therefore, in these circumstances online heat exchanger leak detection models can achieve huge economic benefits for oil and gas industries.

Monitoring and maintaining final products' color are critical for oil and gas industries and any deviation in these parameters result in massive losses. Therefore, the availability of reliable online color readings is crucial for refineries. Industries usually rely on online color analyzers or manual sampling to estimate the color of final products. However, color analyzers are expensive to purchase and maintain. These devices are also prone to malfunction and improper analysis. Manual sampling on the other hand, is a time-intensive activity, requires increased manpower, and does not provide continuous online monitoring.

The study findings illustrated that dynamic PLS can accurately predict and identify the pattern of products' colors. Unlike other studies where online prediction of color requires online properties of the product like density and viscosity (Leam et al., 2020) or an online spectroscopy (Barra et al., 2022), the proposed color prediction scheme in this study relies on historical color and other process parameters data only. Therefore, the proposed method of color prediction in this study using dynamic PLS helps to avoid maintaining such costly instruments and does not rely on other product's properties that might affect the prediction accuracy. Moreover, dynamic PLS model with a Q^2 of 0.98 predicted the

color with higher accuracy comparing to the other two studies (Leam et al., 2020; Barra et al., 2022).

4.4. Conclusions

This chapter first presented the result of PCA, PLS, and their dynamic variants in detecting a heat exchanger leak that occurred in the past. The study findings illustrated that static PCA, dynamic PCA, and dynamic PLS have succeeded in detecting the leak 50 days (with 95% confidence limit) and 19 days (with 99% confidence limit) before it was discovered at the facility. Online static and dynamic PCA models monitoring the heat exchanger's performance in real-time for one month also without generating any false alarms performed well. Finally, the study findings demonstrated that the applied leak detection system can reduce refinery's LPOs and enhance assets' reliability.

Chapter Five

Conclusions, Limitations, and Future Works

5.1. Introduction

5.2. Limitations

5.3. Future Works

5.4. Conclusions

Chapter Five

Conclusions, Limitations, and Future Works

5.1. Introduction

This chapter is divided into three sections. First, the study limitations are highlighted. After that, the recommendations for future works are provided. Finally, conclusion from the overall study results is presented.

5.2. Limitations

This research has faced several challenging limitations. First, there have been limited studies conducted in this field. A few studies in the reviewed literature (AlGhazzawi & Lennox, 2008; Al Tamimi et al., 2019) have investigated the use of data-driven fault detection models in oil and gas industries processes. Furthermore, none of these studies used fault detection approaches to identify leaks in the heat exchangers. There has also been limited studies predicting oil and gas products color (Leam et al., 2020; Barra et al., 2022). Moreover, the application of real-time process monitoring and fault detection in oil and gas industries has not been explored in the reviewed literature. This has limited the researcher's resources to develop the research methodology.

Second, the study was limited to use the existing measurements at the refinery. However, the existing instrumentation infrastructure around the asset lacks several critical measurements such as heat exchanger inlet

and outlet pressure indicators. Such vital variables could've aided to build more accurate fault detection and color prediction models.

Third, the commercial software used in this study (Aspen ProMV) is limited to PCA and PLS models. Therefore, other multivariate statistical and machine learning approaches could not be implemented for real-time fault detection. Due to software limitation dynamic PLS model also could not be deployed online for real-time product color prediction. Finally, the limited research timeframe did not allow to monitor the performance of online heat exchanger leak detection models for an extended period.

5.3. Future Works

The following recommendations are provided for future heat exchanger leak detection studies. As this study considered the existing measurements around the heat exchanger only, adding other critical process variables such as heat exchanger inlet and outlet pressure can improve the accuracy and performance of the models. Furthermore, the focus in this research is to demonstrate the efficacy of using multivariate statistics to detect leak on a single heat exchanger. According to the results obtained it is recommended to scale up the methodology to include a larger number of heat exchangers.

This research adopted multivariate statistical approaches to detect leak in a heat exchanger and predict product color, it is recommended for future studies to detect leak in heat exchangers and predict product color using other machine learning methods such as ANN and LSTM. It is also recommended to leverage heat exchanger's thermodynamics and develop hybrid models using both first principle and data-driven models. Finally, HSR product color was predicted in this study, future researchers are

advised to predict and monitor other quality properties such as flash point, density, and viscosity in real-time.

5.4. Conclusions

This study leveraged different multivariate statistical methods to detect heat exchanger leak at a local refinery in real-time. Static PCA, dynamic PCA, static PLS, and dynamic PLS approaches were adopted first offline, to detect a heat exchanger leak that occurred in the past. The models were then deployed online to monitor and detect the heat exchanger leak on real-time basis for a period of one month. Moreover, HSR product color was also predicted using static and dynamic PLS models. The results of the offline analysis demonstrated that static PCA, dynamic PCA, and dynamic PLS models with detecting the leak 50 days (with 95% significant level) and 19 days (with 99% significant level) before it was discovered by the plant operators performed well in identifying the leak. Dynamic PCA model, however, without generating any false alerts outperformed the other two approaches that generated a single false alarm. Furthermore, dynamic PLS by achieving a Q^2 of 0.98 and RMSE of 0.16 predicted the product color very accurately. Static PLS on the other hand, performed very poorly in both detecting the leak and predicting the color. Moreover, online static and dynamic PCA leak detection models did not generate any false alarms during the entire 1-month monitoring period and efficiently reflected the healthy condition of the equipment.

The proposed leak detection models with detecting the heat exchanger leak at such early stages allows plant operators to prevent contamination of the product in the downstream tank that can save a value up to 350,000 USD depending on the refinery's operations cost.

Early leak detection also allows the maintenance team to formulate a plan and arrange spare parts in advance, minimizing the heat exchanger downtime. This results in saving a daily value between 20,000 USD and 90,000 USD depending on the reduced feed amount and refinery's net profit margin. The analysis in this study has been done for a single heat exchanger. However, the model can be applied on any heat exchanger. Therefore, considering the large number of heat exchangers at refineries, the proposed real-time heat exchanger leak detection system can achieve huge savings for oil and gas industries. Finally, accurate product color prediction scheme demonstrated in this study can be used instead of expensive color analyzers that are susceptible to malfunction and difficult to maintain.

References

Senseye. (2021, June 30). *The Staggering Cost of Unplanned Downtime*.

Www.senseye.io. <https://www.senseye.io/blog/the-true-cost-of-downtime#:~:text=Almost%20%241%20trillion%3A%20the%20staggering%20cost%20of%20unplanned%20downtime%20to%20major%20manufacturers>

Serth, R. W. (2007). *Process heat transfer : principles and applications*.

Elsevier Acad. Press, C.

Panahi, H., Eslami, A., Golozar, M. A., & Ashrafi Laleh, A. (2020). An

investigation on corrosion failure of a shell-and-tube heat exchanger in a natural gas treating plant. *Engineering Failure Analysis, 118*, 104918.

<https://doi.org/10.1016/j.engfailanal.2020.104918>

Ranjbar, K., & Abasi, A. (2013). Failure assessment of crude oil

preheating tubes in mono ethylene glycol–water mixture solution. *Engineering Failure Analysis, 31*, 161–167.

<https://doi.org/10.1016/j.engfailanal.2013.01.048>

Mazlan, S. N. F., Ismail, A., Noh, L. M., & Ahmad, S. (2018). Failure

Analysis on Heat Exchanger Tube Bundle Exposed to Naphthenic Acid Corrosion. *Key Engineering Materials, 791*, 95–101.

<https://doi.org/10.4028/www.scientific.net/kem.791.95>

- Usman, A., & Khan, A. N. (2008). Failure analysis of heat exchanger tubes. *Engineering Failure Analysis*, 15(1), 118–128.
<https://doi.org/10.1016/j.engfailanal.2006.11.051>
- Chiang, L. H., Braatz, R. D., & Russel, E. L. (2001). *Fault detection and diagnosis in industrial systems*. Springer, Cop.
- Isermann, R., & Ballé, P. (1997). Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice*, 5(5), 709–719.
[https://doi.org/10.1016/s0967-0661\(97\)00053-1](https://doi.org/10.1016/s0967-0661(97)00053-1)
- Kesavan, P., & Lee, J. H. (1997). Diagnostic Tools for Multivariable Model-Based Control Systems. *Industrial & Engineering Chemistry Research*, 36(7), 2725–2738.
<https://doi.org/10.1021/ie9606653>
- Jeong, H., Park, B., Park, S., Min, H., & Lee, S. (2019). Fault detection and identification method using observer-based residuals. *Reliability Engineering & System Safety*, 184, 27–40.
<https://doi.org/10.1016/j.ress.2018.02.007>
- Isermann, R. (2005). Model-based fault-detection and diagnosis – status and applications. *Annual Reviews in Control*, 29(1), 71–85.
<https://doi.org/10.1016/j.arcontrol.2004.12.002>
- Weyer, E., Szederkényi, G., & Hangos, K. (2000). Grey box fault detection of heat exchangers. *Control Engineering Practice*, 8(2),

121–131. [https://doi.org/10.1016/s0967-0661\(99\)00132-x](https://doi.org/10.1016/s0967-0661(99)00132-x)

Svard, C. (2009). *Residual Generation Methods for Fault Diagnosis with Automotive Applications* [Thesis].

Kinnaert, M., Vrančić, D., Denolin, E., Juričić, Đ., & Petrovčić, J. (2000). Model-based fault detection and isolation for a gas–liquid separation unit. *Control Engineering Practice*, 8(11), 1273–1283. [https://doi.org/10.1016/s0967-0661\(00\)00064-2](https://doi.org/10.1016/s0967-0661(00)00064-2)

Sun, X., Chen, T., & Marquez, H. J. (2002). Efficient model-based leak detection in boiler steam-water systems. *Computers & Chemical Engineering*, 26(11), 1643–1647. [https://doi.org/10.1016/s0098-1354\(02\)00147-3](https://doi.org/10.1016/s0098-1354(02)00147-3)

Făgărăsan, I., Andone, D., Iliescu, S. St., & Barbulea, S. (2006). Model-Based Fault Detection and Isolation of Thermal Processes. *IFAC Proceedings Volumes*, 39(19), 99–104. <https://doi.org/10.3182/20061002-4-bg-4905.00017>

Khan, A. Q. (2013). *Observer-based Fault Detection in Nonlinear Systems* [Thesis].

Kinnaert, M., Vrančić, D., Denolin, E., Juričić, Đ., & Petrovčić, J. (2000). Model-based fault detection and isolation for a gas–liquid separation unit. *Control Engineering Practice*, 8(11), 1273–1283. [https://doi.org/10.1016/s0967-0661\(00\)00064-2](https://doi.org/10.1016/s0967-0661(00)00064-2)

Hammouri, H., Kinnaert, M., & El Yaagoubi, E. H. (1999). Observer-

- based approach to fault detection and isolation for nonlinear systems. *IEEE Transactions on Automatic Control*, 44(10), 1879–1884. <https://doi.org/10.1109/9.793728>
- Dong, Y., & Qin, S. J. (2015). Dynamic-Inner Partial Least Squares for Dynamic Data Modeling. *IFAC-PapersOnLine*, 48(8), 117–122. <https://doi.org/10.1016/j.ifacol.2015.08.167>
- Wang, J. (2022). *Data-Driven Fault Detection and Reasoning For Industrial Monitoring*. Springer.
- Yan, X.-G., & Edwards, C. (2008). Robust sliding mode observer-based actuator fault detection and isolation for a class of nonlinear systems. *International Journal of Systems Science*, 39(4), 349–359. <https://doi.org/10.1080/00207720701778395>
- Li, L., Ding, S. X., Yang, Y., & Zhang, Y. (2016). Robust fuzzy observer-based fault detection for nonlinear systems with disturbances. *Neurocomputing*, 174, 767–772. <https://doi.org/10.1016/j.neucom.2015.09.102>
- Chiang, L. H., & Colegrove, L. F. (2007). Industrial implementation of on-line multivariate quality control. *Chemometrics and Intelligent Laboratory Systems*, 88(2), 143–153. <https://doi.org/10.1016/j.chemolab.2007.02.005>
- Dunn, K. (2010). *Process Improvement Using Data*.

- Krishnan, R. A., & Pappa, N. (n.d.). Real time fault diagnosis for a heat exchanger — a model based approach. *2005 Annual IEEE India Conference - Indicon*. <https://doi.org/10.1109/indcon.2005.1590128>
- AlGhazzawi, A., & Lennox, B. (2008). Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice*, *16*(3), 294–307. <https://doi.org/10.1016/j.conengprac.2007.04.014>
- Zhang, Y., & Dudzic, M. S. (2006). Online monitoring of steel casting processes using multivariate statistical technologies: From continuous to transitional operations. *Journal of Process Control*, *16*(8), 819–829. <https://doi.org/10.1016/j.jprocont.2006.03.005>
- Yu, Y., Peng, M., Wang, H., Ma, Z., & Li, W. (2020). Improved PCA model for multiple fault detection, isolation and reconstruction of sensors in nuclear power plant. *Annals of Nuclear Energy*, *148*, 107662. <https://doi.org/10.1016/j.anucene.2020.107662>
- Hu, Z., Chen, Z., Gui, W., & Jiang, B. (2014). Adaptive PCA based fault diagnosis scheme in imperial smelting process. *ISA Transactions*, *53*(5), 1446–1455. <https://doi.org/10.1016/j.isatra.2013.12.018>
- Yu, H., & MacGregor, J. F. (2003). Multivariate image analysis and regression for prediction of coating content and distribution in the production of snack foods. *Chemometrics and Intelligent*

Laboratory Systems, 67(2), 125–144.

[https://doi.org/10.1016/s0169-7439\(03\)00065-0](https://doi.org/10.1016/s0169-7439(03)00065-0)

Bounoua, W., & Bakdi, A. (2021). Fault detection and diagnosis of nonlinear dynamical processes through correlation dimension and fractal analysis based dynamic kernel PCA. *Chemical Engineering Science*, 229, 116099. <https://doi.org/10.1016/j.ces.2020.116099>

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K.

(2003). A review of process fault detection and diagnosis.

Computers & Chemical Engineering, 27(3), 327–346.

[https://doi.org/10.1016/s0098-1354\(02\)00162-x](https://doi.org/10.1016/s0098-1354(02)00162-x)

Bounoua, W., Benkara, A. B., Kouadri, A., & Bakdi, A. (2019). Online monitoring scheme using principal component analysis through Kullback-Leibler divergence analysis technique for fault detection. *Transactions of the Institute of Measurement and Control*, 42(6), 1225–1238. <https://doi.org/10.1177/0142331219888370>

Godoy, J. L., Vega, J. R., & Marchetti, J. L. (2013). A fault detection and diagnosis technique for multivariate processes using a PLS-decomposition of the measurement space. *Chemometrics and Intelligent Laboratory Systems*, 128, 25–36.

<https://doi.org/10.1016/j.chemolab.2013.07.006>

L. de Souza, D., H. Granzotto, M., M. de Almeida, G., & C. Oliveira-Lopes, L. (2014). Fault Detection and Diagnosis Using Support

Vector Machines - A SVC and SVR Comparison. *Journal of Safety Engineering*, 3(1), 18–29.

<https://doi.org/10.5923/j.safety.20140301.03>

Chiang, L. H., Kotanchek, M. E., & Kordon, A. K. (2004). Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Computers & Chemical Engineering*, 28(8), 1389–1401.

<https://doi.org/10.1016/j.compchemeng.2003.10.002>

Wang Kai, Bai Ming, Cong Lin, Yang Ting, & Sun huayu. (2011). The Method of Fault Detection and Diagnosis for Induction Motors Based on Wavelets and Independent Component Analysis. *International Journal of Digital Content Technology and Its Applications*, 5(9), 405–410.

<https://doi.org/10.4156/jdcta.vol5.issue9.45>

Zhang, N., Tian, X., Cai, L., & Deng, X. (2015). Process fault detection based on dynamic kernel slow feature analysis. *Computers & Electrical Engineering*, 41, 9–17.

<https://doi.org/10.1016/j.compeleceng.2014.11.003>

Nawaz, M., Maulud, A. S., Zabiri, H., Taqvi, S. A. A., & Idris, A. (2021). Improved process monitoring using the CUSUM and EWMA-based multiscale PCA fault detection framework. *Chinese Journal of Chemical Engineering*, 29, 253–265.

<https://doi.org/10.1016/j.cjche.2020.08.035>

- Al Tamimi, S., Khalaf, O., Abdulla, A., & Bin Shams, M. (2019). Process monitoring and troubleshooting of a local refinery's hydrogen plant using multivariate methods. *International Journal of Oil, Gas and Coal Technology*, 22(3), 346.
<https://doi.org/10.1504/ijogct.2019.10024573>
- Geng, Z., & Zhu, Q. (2005). Multiscale Nonlinear Principal Component Analysis (NLPCA) and Its Application for Chemical Process Monitoring. *Industrial & Engineering Chemistry Research*, 44(10), 3585–3593. <https://doi.org/10.1021/ie0493107>
- Ku, W., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1), 179–196.
[https://doi.org/10.1016/0169-7439\(95\)00076-3](https://doi.org/10.1016/0169-7439(95)00076-3)
- Deng, X., Tian, X., & Chen, S. (2013). Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis. *Chemometrics and Intelligent Laboratory Systems*, 127, 195–209.
<https://doi.org/10.1016/j.chemolab.2013.07.001>
- Choi, S. W., & Lee, I.-B. (2004). Nonlinear dynamic process monitoring based on dynamic kernel PCA. *Chemical Engineering Science*, 59(24), 5897–5908. <https://doi.org/10.1016/j.ces.2004.07.019>

- Muller, K.-R. ., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001).
An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2), 181–201.
<https://doi.org/10.1109/72.914517>
- Kourti, T., Nomikos, P., & MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control*, 5(4), 277–284.
[https://doi.org/10.1016/0959-1524\(95\)00019-m](https://doi.org/10.1016/0959-1524(95)00019-m)
- Jia, R., & Mao, Z. (2013). Robust PLS model based product quality control strategy for solvent extraction process. *2013 25th Chinese Control and Decision Conference (CCDC)*.
<https://doi.org/10.1109/ccdc.2013.6561110>
- Wang, G., & Yin, S. (2015). Enhanced quality-related fault detection approach based on OSC and M-PLS. *IEEE Transactions on Industrial Informatics*, 1–1.
<https://doi.org/10.1109/tii.2015.2396853>
- Xu, X., Liu, Q., & Ding, J. (2018). A Modified Dynamic PLS for Quality Related Monitoring of Fractionation Processes. *IFAC-PapersOnLine*, 51(18), 315–320.
<https://doi.org/10.1016/j.ifacol.2018.09.319>
- Marjanovic, O., Lennox, B., Sandoz, D., Smith, K., & Crofts, M. (2006). Real-time monitoring of an industrial batch process. *Computers &*

Chemical Engineering, 30(10-12), 1476–1481.

<https://doi.org/10.1016/j.compchemeng.2006.05.040>

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6(6), 329–348. [https://doi.org/10.1016/0959-1524\(96\)00009-1](https://doi.org/10.1016/0959-1524(96)00009-1)

Lennox, B., Montague, G. A., Hiden, H. G., Kornfeld, G., & Goulding, P. R. (2001). Process monitoring of an industrial fed-batch fermentation. *Biotechnology and Bioengineering*, 74(2), 125–135. <https://doi.org/10.1002/bit.1102>

Patton, R. J., Chen, J., & Siew, T. M. (1994). Fault diagnosis in nonlinear dynamic systems via neural networks. *International Conference on Control '94*. <https://doi.org/10.1049/cp:19940332>

Jack, L. B., & Nandi, A. K. (2002). Fault Detection Using Support Vector Machines and Artificial Neural Networks, Augmented by Genetic Algorithms. *Mechanical Systems and Signal Processing*, 16(2-3), 373–390. <https://doi.org/10.1006/mssp.2001.1454>

Heo, S., & Lee, J. H. (2018). Fault detection and classification using artificial neural networks. *IFAC-PapersOnLine*, 51(18), 470–475. <https://doi.org/10.1016/j.ifacol.2018.09.380>

Iqbal, R., Maniak, T., Doctor, F., & Karyotis, C. (2019). Fault Detection and Isolation in Industrial Processes Using Deep Learning

- Approaches. *IEEE Transactions on Industrial Informatics*, 15(5), 3077–3084. <https://doi.org/10.1109/tii.2019.2902274>
- Zhao, H., Sun, S., & Jin, B. (2018). Sequential Fault Diagnosis Based on LSTM Neural Network. *IEEE Access*, 6, 12929–12939. <https://doi.org/10.1109/access.2018.2794765>
- Gravanis, G., Dragogias, I., Papakiriakos, K., Ziogou, C., & Diamantaras, K. (2022). Fault detection and diagnosis for non-linear processes empowered by dynamic neural networks. *Computers & Chemical Engineering*, 156, 107531. <https://doi.org/10.1016/j.compchemeng.2021.107531>
- Wang, P., Zhang, J., Wan, J., & Wu, S. (2022). A fault diagnosis method for small pressurized water reactors based on long short-term memory networks. *Energy*, 239, 122298. <https://doi.org/10.1016/j.energy.2021.122298>
- Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, 22(9), 1567–1581. <https://doi.org/10.1016/j.jprocont.2012.06.009>
- Wu, H., & Zhao, J. (2018). Deep convolutional neural network model based chemical process fault diagnosis. *Computers & Chemical*

Engineering, 115, 185–197.

<https://doi.org/10.1016/j.compchemeng.2018.04.009>

Chen, J., & Liao, C.-M. (2002). Dynamic process fault monitoring based on neural network and PCA. *Journal of Process Control*, 12(2), 277–289. [https://doi.org/10.1016/s0959-1524\(01\)00027-0](https://doi.org/10.1016/s0959-1524(01)00027-0)

Zhu, H. (2007). Modeling Method of Monitoring Nonlinear System by Combining Neural Network with Partial Least Squares. 2007 *International Conference on Mechatronics and Automation*.
<https://doi.org/10.1109/icma.2007.4304059>

Prosensus. (2014). *Introduction to Principal Component Analysis*.
<https://prosensus.com/>

Prosensus. (2014). *Introduction to Projection to Latent Structures*.
<https://prosensus.com/>

Malhi, A., & Gao, R. X. (2004). PCA-Based Feature Selection Scheme for Machine Defect Classification. *IEEE Transactions on Instrumentation and Measurement*, 53(6), 1517–1525.
<https://doi.org/10.1109/tim.2004.834070>

Song, F., Guo, Z., & Mei, D. (2010). *Feature Selection Using Principal Component Analysis*. IEEE Xplore.
<https://doi.org/10.1109/ICSEM.2010.14>

Yoo, Y.-J. (2019). Fault Detection of Induction Motor Using Fast Fourier Transform with Feature Selection via Principal Component

Analysis. *International Journal of Precision Engineering and Manufacturing*, 20(9), 1543–1552. <https://doi.org/10.1007/s12541-019-00176-z>

Onel, M., Kieslich, C. A., & Pistikopoulos, E. N. (2019). A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process. *AIChE Journal*, 65(3), 992–1005. <https://doi.org/10.1002/aic.16497>

Rauber, T. W., Boldt, F. A., & Munaro, C. J. (2020). Feature selection for multivariate contribution analysis in fault detection and isolation. *Journal of the Franklin Institute*, 357(10), 6294–6320. <https://doi.org/10.1016/j.jfranklin.2020.03.005>

Leam, J. J., Khor, C. S., & Dass, S. C. (2020). Saybolt color prediction for condensates and light crude oils. *Journal of Petroleum Exploration and Production Technology*, 11(1), 253–268. <https://doi.org/10.1007/s13202-020-01031-y>

Barra, I., Alaoui Mansouri, M., El Moatassem, T., Kharbach, M., Abou Oualid, H., Cherrah, Y., Hanafi, M., Mostafa Qannari, E., & Bouklouze, A. (2022). Prediction of diesel fuel quality indicators using FT-MIR spectroscopy and chemometrics. *Infrared Physics & Technology*, 122, 104096. <https://doi.org/10.1016/j.infrared.2022.104096>

Appendices

الملخص

تعتبر المعدات الميكانيكية ذات أهمية حيوية لعمليات الصناعات التحويلية، لذلك يؤدي فشل هذه الأجهزة إلى انقطاع الأعمال بالإضافة لخسائر كبيرة. يساعد الاكتشاف المبكر لهذه الأخطاء المصانع في صياغة خطة صيانة مناسبة لتجنب فترات التوقف غير المخطط لها. كما يؤدي هذا أيضًا إلى تحسين اعتمادية المعدات وصيانتها وتوافرها. لذا تهدف هذه الدراسة للاستفادة من طرق التحليل إحصائية متعددة المتغيرات للبيانات لاكتشاف تسرب مبادل حراري في مصفاة محلية في الوقت الفعلي. ولتحقيق ذلك تم جمع البيانات من سجلات البيانات المتاحة للمبادل الحراري الذي يعمل كجزء أساسي في شبكة التسخين المسبق لوحدة تقطير النفط. ثم تم تطوير مجموعة من النماذج باستخدام تحليل العنصر الرئيسي (PCA)، و PCA الديناميكي، والمربعات الصغرى الجزئية (PLS)، و PLS الديناميكي لاكتشاف ما حدث من تسربات سابقة في المبادل الحراري. بعد ذلك تم تنفيذ النماذج لمراقبة أداء المبادل الحراري واكتشاف التسربات في الوقت الفعلي. أظهرت نتائج الدراسة فعالية PCA الثابت، و PCA الديناميكي، و PLS الديناميكي في اكتشاف التسربات 50 يومًا (بحد ثقة 95%) و 19 يومًا (بحد ثقة 99%) قبل أن يتم ملاحظتها من قبل المشغلين في المصفاة. بالإضافة لذلك، عملت النماذج المتصلة بالبيانات اللحظية بشكل جيد في مراقبة أداء المبادل الحراري في الوقت الفعلي دون إصدار أي إنذارات خاطئة.

جامعة البحرين

قسم برامج الدراسات العليا

كلية العلوم



نظام قائم على البيانات لكشف وتشخيص الأعطال لمصفاة محلية في الوقت الفعلي

أطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

عبدالرحيم محمد صالح

20134188

إشراف

د. محمد بن شمس
أستاذ مشارك

جامعة البحرين

مملكة البحرين
مارس 2022