"I hereby declare that I have read through this report entitle "**HIGH ACCURACY ANOMALY DETECTION FOR CYBER PHYSICAL SYSTEM USING SELF-ORGANIZING MAP BASED ALGORITHM**" and found that it has complied with requirement for the partial fulfilment for awarding the degree of Bachelor of Mechatronics Engineering."

Signature : ....

Supervisor's Name : ASSOC. PROF. DR. MUHAMMAD FAHMI BIN MISKON

Date : ....

HIGH ACCURACY ANOMALY DETECTION FOR CYBER PHYSICAL SYSTEM USING SELF-ORGANIZING MAP BASED ALGORITHM

KWAN CHAK YIN

A report submitted in partial fulfilment of the requirements for the degree of Bachelor of Mechatronics Engineering

Faculty of Electrical Engineering
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2018

I declare that this report entitles "**HIGH ACCURACY ANOMALY DETECTION FOR CYBER PHYSICAL SYSTEM USING SELF-ORGANIZING MAP BASED ALGORITHM**" is the result of my own research except as cited in the references. The report has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature : ....

Name : KWAN CHAK YIN

Date : ....

To my beloved family

ACKNOWLEDGEMENT

First, I would like to express my sincere thanks and greatest appreciation to my supervisor, Associate Professor Dr. Muhammad Fahmi Bin Miskon for the expertise sharing, truthful and treasured guidance and encouragement extended to me during the undergoing of FYP.

I would like to express my greatest gratitude to my FYP panels, Pn. Nurdiana Bt. Nordin @ Musa and Dr Mohd. Shahrieel B. Mohd. Aras for the valuable yet attentive advice given during the undergoing of FYP. Apart from that, I would like to express my deep sense of gratitude to all the coursemates and friends for being kind and helpful in completing the FYP.

I would like to take this opportunity to express greatest appreciation to my family for their continuous shower of love, unceasing encouragement and support throughout my life. I derived inspiration from their sacrifice, encouragement from their faith, found happiness in their pride and the strength from their unconditional love.

Besides that, my grateful thanks are also extended to Universiti Teknikal Malaysia Melaka (UTeM) for the resources offered in running FYP. Without the help of the particulars that I mentioned above, I might face difficulties during the FYP. Last but not least, my sense of gratitude is recorded to the one and all who, directly or indirectly, have offered their helping hand during the entire period of FYP.

ABSTRACT

In order to embrace challenges of Industry 4.0 (I4.0), failure prediction on the machinery in the Cyber-Physical System (CPS) is important which gives rise to the research in anomaly detection. Recent studies on anomaly detection generally applied to the network security, fraud system and image processing. However, anomaly detection in I4.0 have difficulty in ensuring the designed algorithm is self-adaptive without compromising the accuracy of the prediction. Hence, this project addresses this problem by proposing a habituating SOM-based algorithm to predict the possible failure faced in the system. In the proposed method, the SOM and k-means act as the clustering network for the mechanism, while the habituation function take role as set of habituating synapses that form connection among the network neurons to the output. Weight vector for the neurons is initialized via k-means clustering to ensure reasonable number of cluster and proper distribution of weight vector. Receiver Operating Characteristic (ROC) curve is used to optimize threshold value of Euclidean distance. Accuracy test is carried out by execute the algorithm to different dataset that contain various number of anomalies. The performance of the algorithm is evaluated via the application of confusion matrix in term of accuracy. The proposed algorithm can detect the anomalies occur in the data accurately with minimum accuracy of 98.5% and maximum accuracy of 99.2%. This indicates that there is possibility to use the proposed anomaly detection technique to predict the possible failure in CPS' machinery.

ABSTRAK

Demi untuk menyahut cabaran dalam menerajui I4.0, ramalan ketidakfungsian mesin dalam dalam CPS sangat penting yang mana membangkitkan penyelidikan terhadap pengesanan keganjilan. Kajian yang sedia ada terhadap pengesanan keganjilan kebanyakannya digunakan untuk sistem keselamatan rangkaian, pengesanan penipuan dan pemprosesan imej. Walau bagaimanapun, pengesanan keganjilan dalam I4.0 mengalami kesulitan dalam menjamin algoritma yang direka mempunyai kebolehan pembelajaran kendiri tanpa mengabaikan ketepatannya dalam mengesan keganjilan. Oleh itu, kajian ini menangani masalah ini dengan mencadangkan algoritma berasaskan pembiasaan SOM untuk meramalkan kemungkinan kegagalan yang dialami oleh mesin. Dalam kaedah yang dicadangkan, SOM dan k-means bertindak sebagai rangkaian cluster untuk mekanisme, manakala fungsi pembiasaan memainkan peranan sebagai set sinapsis pembiasaan yang membentuk sambungan antara neuron rangkaian ke hasilannya. Vektor berat untuk neuron diasaskan melalui k-means clustering untuk menjaminkan bilangan neuron dan pengagihan vektor berat yang munasabah. Lengkungan ROC digunakan untuk mengoptimumkan nilai ambang jarak Euclidean. Ujian ketepatan dijalankan dengan melaksanakan algoritma dengan set data yang mempunyai bilangan keganjilan yang berbeza. Prestasi algoritma dinilai melalui aplikasi matriks confusion dari segi ketepatan. Algoritma yang dicadangkan boleh mengesan keganjilan yang terdapat dalam set data dengan tepat dengan ketepatan yang sekurangnya 98.5% dan sebanyaknya 99.2%. Ini menunjukkan bahawa ada kemungkinan untuk menggunakan teknik pengesanan keganjilan yang dicadangkan untuk meramal kegagalan yang mungkin wujud dalam mesin CPS.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	–	Accuracy
AI	–	Artificial Intelligence
ANN	–	Artificial Neural Network
BMU	–	Best Matching Unit
BPNN	–	Backpropagation ANN
CNN	–	Convolutional ANN
CPI	–	Continuous Process Improvement
CPS	–	Cyber-Physical System
DBM	–	Deep Boltzmann Machine
DBN	–	Deep Belief Network
DNN	–	Deep ANN
DWT	–	Discrete Wavelet Transform
FN	–	False negative
FP	–	False positive
FPR	–	False Positive Rate
GGSOM	–	Growing Grid SOM
GHSOM	–	Growing Hierarchical SOM
GNG	–	Growing Neural Gas
I4.0	–	Industry 4.0
IDS	–	Intrusion Detection System
IoT	–	Internet of Things
IR4.0	–	Industrial Revolution 4.0
MCC	–	Matthews Correlation Coefficient
MLP	–	Multilayer Perceptron
PHM	–	Prognostic and Health Management
PPV	–	Precision/ Positive Predictive Value
PSO	–	Particle Swarm Optimization
PSOM	–	Periodic SOM
RBFNN	–	Radial Basis Function ANN

RBM	–	Restricted Boltzmann Machine
RNN	–	Recurrent ANN
ROC	–	Receiver Operating Characteristic
SOM	–	Self-organizing Map
SSOM	–	Scalable SOM
TN	–	True negative
TNR	–	True Negative Rate
TP	–	True positive
TPR	–	True Positive Rate
WEF	–	World Economic Forum

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

The Industry 4.0 (I4.0) is introduced in German Hannover Fair. It ignites the start of the Industrial Revolution 4.0 (IR4.0) with the improvement in the Cyber-Physical System (CPS) and the Internet of Things (IoT) [1]. I4.0 become the main concern in World Economic Forum (WEF) 2016 Conference. Executive chairperson of WEF 2016, Prof. Klaus Schwab state that I4.0 is mainly described the CPS and mainly build on the third industrial revolution that uses electronic and information technology system as basic of the automated production line [2]. Hence, the basic principle of I4.0 is to connect machines, processes, systems as well as business via an intelligent network to control each other autonomously [3]. The main design principles of I4.0 are interoperability that concern about integration that acts as the root of IoT and CPS and cognizance that originated from artificial intelligence (AI) that leads to the decision-making and prognostic maintenance of the system [4].

CPS is the transformative technologies to manage the interconnected system between its physical resources and computational capabilities. Jay Lee et. al. [5, 6] proposed CPS structure consist of five levels, basically known as 5C architecture. The structure of CPS is mainly build up by the advanced connectivity that ensures good interfacing between physical space and cyberspace and the intelligent data analytic that built up the cyberspace. The levels of 5C architecture including smart connection level that provides seamless and tether-free methods for data acquisition system management, data transfer to the central server; data-to-information conversion level that analysed and transformed data collected into valuable information; cyber level that acts as central information hub; cognition level that good in decision-making as well as configuration level that give feedback from cyberspace to physical space as well as take the role of advisory control . [5, 6].

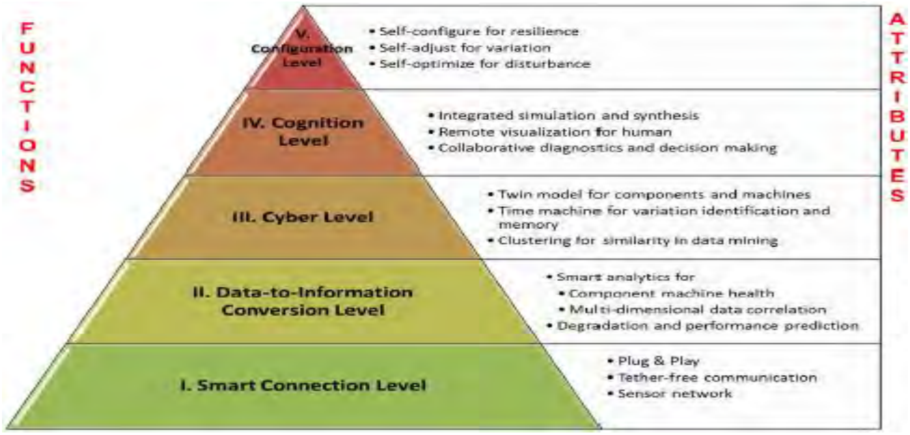
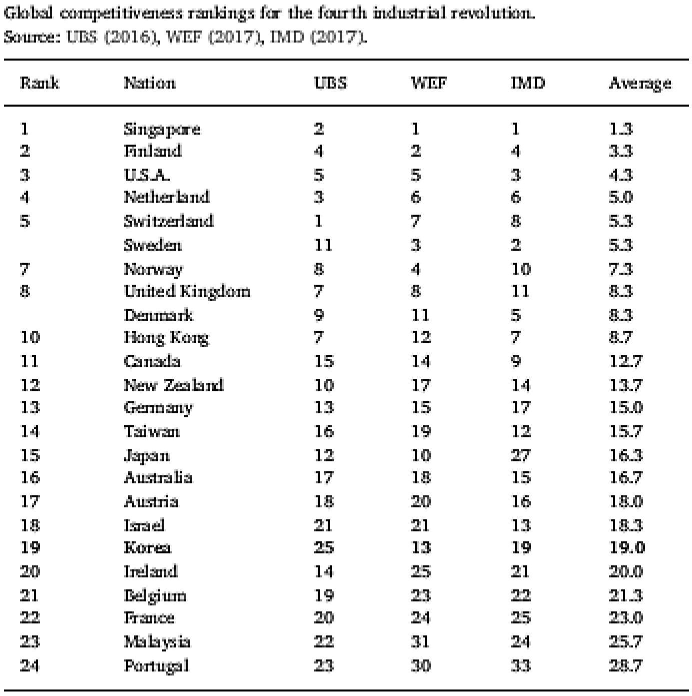

FUNCTIONS
V. Configuration Level
• Self-configure for resilience
• Self-adjust for variation
• Self-optimize for disturbance
IV. Cognition Level
• Integrated simulation and synthesis
• Remote visualization for human
• Collaborative diagnostics and decision making
III. Cyber Level
• Twin model for components and machines
• Time machine for variation identification and memory
• Clustering for similarity in data mining
II. Data-to-Information Conversion Level
• Smart analytics for
• Component machine health
• Multi-dimensional data correlation
• Degradation and performance prediction
I. Smart Connection Level
• Plug & Play
• Tether-free communication
• Sensor network
ATTRIBUTES

Figure 1.1: The implementation of 5C architecture in CPS. [5, 6]

The IR4.0 is concerned to the self-adaptive of the CPS which involve the concept of machine learning that related to the system changes that complete task given via artificial intelligent through techniques such as recognition, prediction and others [7]. It also allows the complex mathematical calculation can be automatically apply to the big data in the cloud for the analysis [8]. Hence, the CPS in I4.0 should not be tied to any fixed program and the changes on the program is allowed in order to enhance the CPS function. Anomaly detection describing a technique of finding patterns in data that do not follow a priori expected behaviour [9]. It is a useful tool to increase productivity with monitoring and failure prevention. Besides, it is also providing an essential way for the machine manufacturers to complete the technology improvement in the continuous improvement process.

1.1 MOTIVATION

Industry 4.0 is unavoidable and bound to happen by TN50 that aimed to rise up the usage of broadband and mobile connectivity, IoT, robotics and Artificial Intelligence (AI) incorporation of formation of “smart city” [10]. The concept of the smart factory had been implement in some country such as United States of America, Germany, Japan, Singapore as well as South Korea, but in terms of automating, Malaysia is still in the infancy stage and back end of the line [11]. According to research on the readiness to embrace challenges of Industry 4.0 by WEF 2016, Malaysia got a relatively low rating on the aspect of infrastructure readiness and legal protection. The relatively low rating on the infrastructure readiness due to the lack of new technology implementation in the industry and the ready facilities and machine in the industry cannot meet the cyber-physical system requirement. Hence, the WEF 2016 evaluated Malaysia as an emerging market. The low maturity level of required technology also became the main concerned challenges in the nine challenges need to face in order to succeed in this industrial revolution [12].



Global competitiveness rankings for the fourth industrial revolution.
Source: UBS (2016), WEF (2017), IMD (2017).

Rank	Nation	UBS	WEF	IMD	Average
1	Singapore	2	1	1	1.3
2	Finland	4	2	4	3.3
3	U.S.A.	5	5	3	4.3
4	Netherland	3	6	6	5.0
5	Switzerland	1	7	8	5.3
	Sweden	11	3	2	5.3
7	Norway	8	4	10	7.3
8	United Kingdom	7	8	11	8.3
	Denmark	9	11	5	8.3
10	Hong Kong	7	12	7	8.7
11	Canada	15	14	9	12.7
12	New Zealand	10	17	14	13.7
13	Germany	13	15	17	15.0
14	Taiwan	16	19	12	15.7
15	Japan	12	10	27	16.3
16	Australia	17	18	15	16.7
17	Austria	18	20	16	18.0
18	Israel	21	21	13	18.3
19	Korea	25	13	19	19.0
20	Ireland	14	25	21	20.0
21	Belgium	19	23	22	21.3
22	France	20	24	25	23.0
23	Malaysia	22	31	24	25.7
24	Portugal	23	30	33	28.7

Figure 1.2: Global competitiveness rankings for IR4.0 [12]

Based on a report by EdgeMarket, 97.3% of business establishment in Malaysia is Small and Medium Enterprise. Only 10% of ICT adoption by SME compared to the ICT adoption in developed countries, which have 50% according to Malaysia Productivity Corporation [13]. Most of the SME are still in the industrial revolution 2.0 that based on mass production empowered by electrically driven [11].

The continuous process improvement (CPI) is important to the industry as a continuing action to improve the processes, services as well as the products of the industry through justifiable changes over a period. The traditional CPI method cannot deal with the current challenges faced that mainly due to the increased of sensors and actuators that lead to multi-parameter production space, big data environment as well as complex dependencies in the production space [14]. Hence, maintenance of the physical assets became the concerned problem in the industry nowadays. Anomaly detection is concerning the cognition level in the CPS to provide continuous improvement process as well as the predictive maintenance. Hence, in order to ensure the completeness in the Industry 4.0, with the use of the anomaly detection algorithms, the anomaly happened can be automatically detect by the CPS itself and making decision once, the anomalies are detect as well as predict in advance the potential failures of the physical system.

1.2 PROBLEM STATEMENT

Anomaly detection consists of two steps, which are learning the normal behaviour from the previous data and identifying the abnormalities in the presence data [14]. Hence, the detection in the CPS based on the big data analytic, which the analysis of data will be done in the cyber-space. In addition, the self-adaptive of the algorithm used as the detection also need to be done in the data-driven orientation. In order to define the metric for maintenance, the algorithm also should be able to predict the performance and predict the possible failure that would occur. Besides, the amount of data in the cloud is very large and lead to the difficulties in clustering the data as well as looking for the anomalies in the data set.

However, anomaly detection in CPS was relatively new. An investigation on the approach to anomaly detection and its accuracy in the existing applications is needed to provide a useful ground for this project.

As the IR4.0 is mainly concerned about the autonomous system, hence the challenge faced in the anomaly detection for IR 4.0 is the self-adaptive of the machine in learning the normal behaviour, comparing the previous with presence, detect the anomalies as well as giving the best decision in the rapid changing manufacturing line in the smart factory. Hence, a neural network algorithm needed to be designed to solve the problem.

Besides, the performance of anomaly detection algorithm is been evaluated in different way in the existing application. The investigation on the method of evaluation for anomaly detection for the existing method is needed in order to choose the most suitable method to evaluate algorithm.

Therefore, an anomaly detection algorithm that is created using the neural network is proposed to solve the problem faced.

1.3 OBJECTIVES

The objectives of this research project are:

i. To investigate the self-adaptive and accuracy of anomaly detection for the cyber-physical system.

ii. To design and develop an anomaly detection algorithm using self-organizing map (SOM).

iii. To validate the anomaly detection algorithm based on the confusion matrix and receiver operating characteristic (ROC) curve.

1.4 SCOPE

The research is focused on creating and design an anomaly detection algorithm that use to detect the unusual performance of the manipulator. Hence, the designed anomaly detection algorithm is used to evaluate the performance of manipulator and define the metric of maintenance of the manipulator via detecting the anomalies present in the manipulator’s behaviour.

Hence, the scopes and limitations of this project are stated as below:

- The CPS investigated in this study is restricted to a manufacturing system with known resources and processes.
- The sample and dataset used in the simulation of algorithm is generated from the V-Rep Robot Simulator.
- The algorithm is mainly designed based on the SOM, while k-means clustering only act as the neuron and neuron weight finder.
- The performance of the algorithm is evaluated based on the accuracy of the algorithm in detecting the true anomalies present in the presence data.
- The algorithm aims to detect the point anomalies.

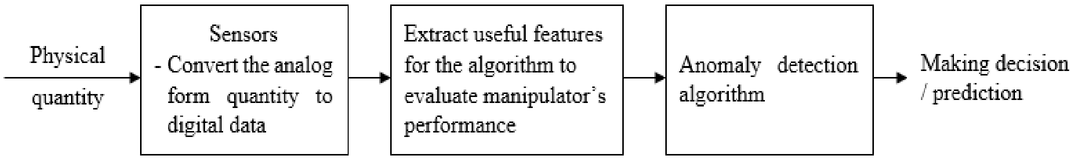

Physical
quantity
Sensors
- Convert the analog form quantity to digital data
Extract useful features for the algorithm to evaluate manipulator's performance
Anomaly detection algorithm
Making decision / prediction

Figure 1.3: The basic block diagram of the system

1.5 OUTLINE OF THE REPORT

This part discusses the role of the rest of the report with a brief explanation on the part of the report. The report is organized as follows:

Chapter 2: Literature Review
This chapter discusses the theoretical background and basic principles that involved for the completion of the report. This part describes the type of anomalies as well as the type of anomaly detection method. This part also reviews and summaries the previous works about the anomaly detection in different applications.

Chapter 3: Research Methodology
This chapter brief the working principle of the method that had been chose to solve the problem of the project. This chapter also discusses the theory related to project solution as well as the method used for the development of the anomaly detection algorithm. The chapter also shows the approach to evaluate the performance of the designed anomaly detection algorithm.

Chapter 4: Result and Discussion
This chapter describes the result that had been achieve in the project. The chapter also gives the analysis and discussion on the result that had been obtained throughout the project.

Chapter 5: Conclusion and Recommendation for Future Work
This chapter summarizes the progress of the project by review all the information obtained for the project. The chapter also provides the recommendation for the future work.

CHAPTER 2

LITERATURE REVIEW

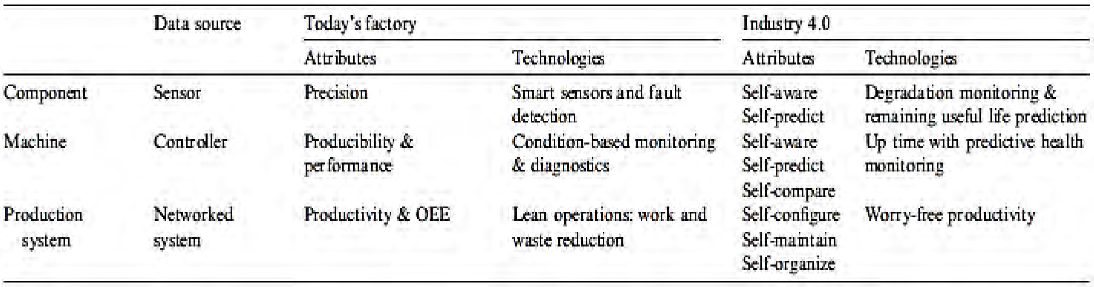
2.1 THEORETICAL BACKGROUND

2.1.1 Cyber-Physical System (CPS)

CPS is a heterogeneous strategy that includes computation and communication devices that link together with sensors and actuators. Reddy [15, 16] state that CPSs is the control system that is distributed, smart, real-time response, self-learning and have the possibility to be connected in a loop. While Ribeiro [17] define CPS as an embedded system with good communication capabilities and the intelligent and permanent connection between the cyber-space which is the logical component and the physical system. With the existence of the CPS, the automated continuous improvement cycle is created.

The existence of preventive maintenance concept and Total Productive Maintenance in 1951 made the Prognostic and Health Management (PHM) start to develop. PHM has the capability to convert data into desired information about the pattern of inefficient performance. Fusion of advanced analytic with communication and the physical machinery leads to CPS and make PHM step closer to completion. Jay Lee et. al. [6] [5] state that increases of usage of sensors and interacted machinery require the system to handle big data. In order to achieve the intelligent and self-adaptive, big data management and leveraging machines' interoperability is necessary.

At the stage of the component, the sensor is not just for the use of precision, but also responsible in capture the sensory data from component and convert it into useful information to offer self-prediction and self-awareness to the system. For the machine stage, the advanced machine data such as controller parameter is not only to supervise the performance of the machine, but also give the self-comparison to the machine by aggregated the advanced machine data to the component information from sensors in order to monitor the status of the machine. For the production stage of the CPS, the aggregated information from the sensors and machines provides the self-maintainability and self-configurability to the factory in order to offer worry-free production with near-zero downtime and improved manufacture scheduling [5].



	Data source	Today's factory		Industry 4.0	
		Attributes	Technologies	Attributes	Technologies
Component	Sensor	Precision	Smart sensors and fault detection	Self-aware Self-predict	Degradation monitoring & remaining useful life prediction
Machine	Controller	Producibility & performance	Condition-based monitoring & diagnostics	Self-aware Self-predict Self-compare	Up time with predictive health monitoring
Production system	Networked system	Productivity & OEE	Lean operations: work and waste reduction	Self-configure Self-maintain Self-organize	Worry-free productivity

Figure 2.1: Comparison between the elements of today's factory and I4.0 [5]

The CPS should can communicate with the integration between the control system, software as well as the communication linkage; uniquely identified, as CPS is part of Internet of Everything (IoE) that can be uniquely addressed with the identifier; have controller, sensors and actuators as the basis of the system to operate; act as the basic building block of I4.0 as well as the enabler of the extra capabilities; have the capabilities to enable smart factory [18].

2.1.2 Anomaly Detection

Anomaly detection is the technique that identifies the data point or pattern that do not follow the expected or normal behaviour, called anomalies or outliers [19]. Anomaly detection is similar to, but have some differences compared to noise removal and novelty detection. Novelty detection is focus on identify the unnoticed pattern in new observation that system is not alert during the training session [20] while noise removal is defined as the process that removing noise from the wanted signal [21]. Major goals of anomaly detection in IR4.0 are automatic monitoring and detect the abnormal events and points on the collected data [22]. Anomaly detection is important to increase the productivity in the manufacturing line, provide continuous improvement to achieve mastery of the technology as well as define the metric for maintenance to detect the malfunction or abnormality in the running machine [23].

2.2 THEORY AND BASIC PRINCIPLE

2.2.1 Type of Anomaly Detection Setup

The mode of the anomaly detection setup is differentiate based on the labels available in the dataset. Hence, the mode is categorized into 3 types whereas:

2.2.1.1 Supervised anomaly detection

Supervised anomaly detection defines the mode that uses the training data contains fully categorized data which is the normal behaviours and anomalies as well as test data. The algorithm is trained first with a set of input with equivalent accurate output and then learned by comparing the actual output with the desired to find an error in order to identify the anomaly [8]. Goldstein and Ushida [24] state that the setup is irrelevant due to the statement that all the outliers are identified and labelled properly as the anomalies for most of the application are unknown in advance. Chandola et. al. [25] state that the supervised method has issues that the number of anomalies is far fewer than the normal data in the training data which the occurrence of imbalanced distribution of data and follow with the high difficulty in getting an accurate label on the unlabelled data.

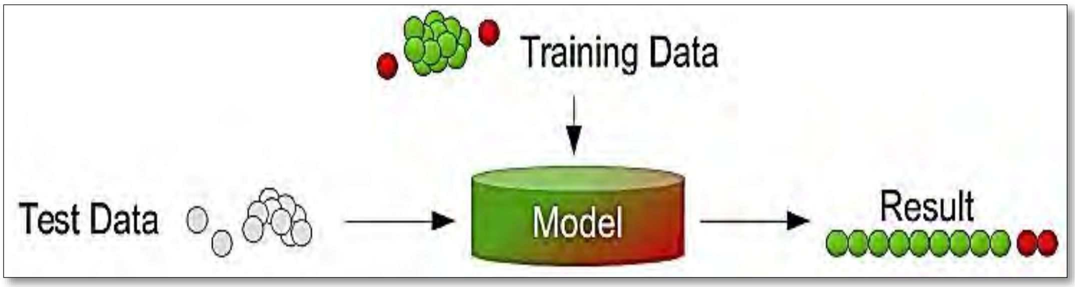


Figure 2.2: The model for supervised anomaly detection mode [24]

2.2.1.2 Semi-supervised anomaly detection

Semi-supervised anomaly detection describes the mode that uses the training data that have normal instances and anomalies to train the algorithm and the test datasets [24]. Chandola et. al. [25] describe the possibility to train the algorithm with the data that only consist of outliers, but this method is not commonly being used due to the difficulty in obtaining the training data that have every possible outlier that might appear in the data.

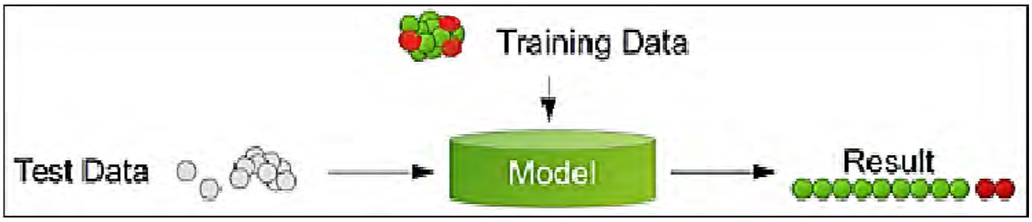


Figure 2.3: The model for semi-supervised anomaly detection mode [24]

2.2.1.3 Unsupervised anomaly detection

Unsupervised anomaly detection is the most widely used method as no training data required. The method is applicable to the data set that has no historical labels to explore the structure of the data [8]. The algorithm estimate the data score based on the fundamental properties of the dataset. Most of the data are labelled based on the distances and densities [24]. If the assumption that the normal data is far more common than an anomaly in the data is false, then resulted in the high false alarm rate. Semi-supervised mode can adapt to an unsupervised mode with the use of a sample of unlabelled data as the training data. The adaptation assumes that very little of outliers contain in the test data and the learning of the algorithm during the training is robust to the outliers appeared in the test data.

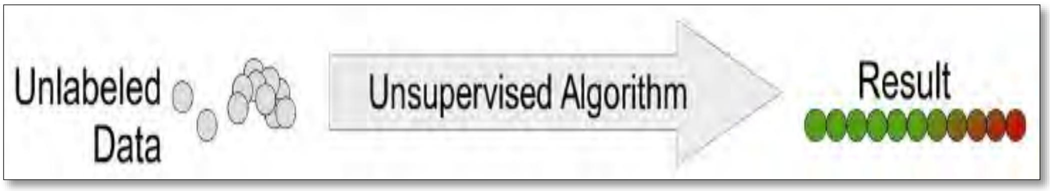


Figure 2.4: The model for unsupervised anomaly detection mode [24]