ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF MECHANICAL AND INDUSTRIAL ENGINEERING

INDUSTRIAL ENGINEERING STREAM

ADDIS ABABA, ETHIOPA

Designing a Multivariate Process Control Procedures for Production System

Case of Ethio Cement PLC

Daniel Ashagrie

A Dissertation Submitted to Addis Ababa Institute of Technology in Partial Fulfillment of Doctor of Philosophy in Industrial Engineering

Addis Ababa University

Addis Ababa Institute of Technology

School of Mechanical and Industrial Engineering

May 2023

AUTHOR'S DECLARATION

I hereby declare that this dissertation entitled "**Designing a Multivariate Process control procedures for Production System: Case of the Ethio Cement PLC** " is my research work. It has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

……………………………………… ……………

Daniel Ashagrie Date

We hereby certify that this dissertation entitled "**Designing a Multivariate Process control procedures for Production System: Case of the Ethio Cement PLC** "is conducted under our supervision. It has not been, and will not be, submitted in whole or in part to another University for the award of any other degree for the Candidate.

……………………………………… ………………

Prof. Daniel Kitaw Date

……………………………………… ………………

Dr. Eshetie Berhan Date

Authors

Daniel Ashagrie, Ph.D. Student in Addis Ababa University (AAU), Addis Ababa Institute of Technology (AAiT), School of Mechanical and Industrial Engineering, Addis Ababa, Ethiopia

E-mail: daniel.ashagrie@yahoo.com, daniel.tashagrie@gmail.com

Dr. Eshetie Berhan, Associate Professor at Industrial Engineering Chair, Addis Ababa University (AAU), Addis Ababa Institute of Technology (AAiT), School of Mechanical and Industrial Engineering, Ethiopia

E-mail: berhan.eshetie@yahoo.com

Professor Daniel Kitaw, Professor at Industrial Engineering Chair, Addis Ababa University (AAU), Addis Ababa Institute of Technology (AAiT), School of Mechanical and Industrial Engineering, Ethiopia

E-mail: danielkitaw@yahoo.com

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY (AAIT)

SCHOOL OF MECHANICAL AND INDUSTRIAL ENGINEERING

INDUSTRIAL ENGINEERING STREAM

ADDIS ABABA ETHIOPIA

Designing a Multivariate Process Control Procedures for Production System

Case of the Ethio Cement PLC

Name of the student: Daniel Ashagrie ID No.GSR/7174/09

Signature_______________ Date__________________

Approved by Board of Examiners: -

Prof. Daniel Kitaw ________	________________	________________
Advisor	Signature	Date
Dr. Eshetie Berhan ________	________________	________________
Co-advisor	Signature	Date
Dr. Kassu Jilcha ________	________________	________________
Internal Examiner	Signature	Date
Dr. Yeneneh Tamrat ________	________________	________________
External Examiner	Signature	Date
Dr. Gezahegn Tesfaye ________	________________	________________
Chair, Industrial Engineering	Signature	Date
Dr. Araya Abera ________	________________	________________

Dean, SMiE Signature Date

Dr. Sosina Mengistu ______________ ________________

____

Associate director, Post graduate Signature Date

ACKNOWLEDGMENTS

First of all, I would like to express my sincere thanks to the almighty God for providing me with peace and strength to embark on this academic path. This work wouldn't have been possible without His blessings.

Professor Daniel Kitaw has given me the privilege of studying my Ph.D. at Addis Ababa University (AAU). I am deeply grateful for his patience and support that gives me enough time and freedom to pass through my interest and complete my dissertation.

I would also like to acknowledge the invaluable contribution of Dr. Eshete Birhan, whose guidance, advice, and unwavering support have been instrumental in my study area. His knowledge and insightful feedback have significantly influenced the quality and direction of my research. Without his guidance, I would not have been able to reach the conclusion of this dissertation.

Then, I extend my thanks to Dr. Birhanu Beshah, my advisor during my MSc studies, for introducing me to the captivating field of Multivariate Statistics and inspiring me to pursue further education. His mentorship and encouragement have played a pivotal role in shaping my research interests and academic growth.

During my stay in Germany, I had the privilege of interacting with supervisors and Ph.D. students from the University Technische Hochschule Nürnberg NCT research center. Their valuable input and discussions provided me with a fresh perspective and inspired me to explore new avenues in my research. I am deeply grateful for their support and for helping me strengthen my research direction.

I am thankful to Hawassa University for sponsoring my Ph.D. program, and I extend my special appreciation to Dr. Fasika Bete for encouraging me to embark on this academic endeavor. Their financial support and belief in my potential have been instrumental in the successful completion of this dissertation.

I would like to express my profound gratitude to the operators and engineers at Ethio-cement Company for their cooperation and assistance during the data collection and analysis phases of my research. Their willingness to share their expertise and provide valuable insights has greatly enriched the empirical foundation of my study.

ACKNOWLEDGMENTS(Continued)

The support, follow-up, and valuable comments from my colleagues and senior Industrial Engineering Ph.D. students at AAU have been truly invaluable, and never forget how you worried about my research direction. I would like to mention some of them; Yichalewal, Wogene, Temesgen, Ashenafi, Asrat, and Ale. Their constructive criticism and thought-provoking discussions have greatly contributed to the refinement of my research work.

Last but not least, I am profoundly grateful to my family. The continuous support, understanding, and encouragement at every step of my life are delightful. Their love and encouragement were my source of strength, and I am forever indebted for their presence in my life.

In conclusion, I would like to express my deepest gratitude to all the individuals and institutions who have contributed to the completion of this dissertation. Their support, guidance, and encouragement have been invaluable, and I am honored to have had the opportunity to work with such remarkable mentors, colleagues, and family members.

Designing a Multivariate Process Control Procedures for Production System

Case of the Ethio Cement PLC

Daniel Ashagrie

ABSTRACT

This dissertation explores the application of Statistical Process Control (SPC) techniques in the manufacturing sector. Industries demand multivariate process monitoring technique capable of identifying cause of variation, and conducting fast and accurate fault detection analysis. However, the existing techniques fall short of satisfying this demand. Hence, the research question is devised as follows: how to design a multivariate process control procedure that can effectively monitor and control the production system, identify the root causes of variation, and provide solutions for improvement. The literature review conducted in this study revealed that while SPC techniques have been extensively studied and applied in various industries, the multivariate analysis of identifying cause of variation is relatively limited. Their practical implementation and adaptation to the industry have not been thoroughly explored. The main objective of this research is to design a procedure that can bridge the theoretical gap that exist in the manufacturing sector. By addressing this gap, it is anticipated that the productivity, quality, and overall performance of the production system can be improved. To address these limitations, a novel approach called the GANNT chart is introduced in this research. The GANNT chart incorporates three key theories: Graph theory (G), Artificial Neural Networks (ANN), and Hotelling T^2 (T). By combining these theories, the proposed approach aims to enhance the process control technique used in the production system. The GANNT chart mimics human decision-making processes and serves as a decision support system for both process engineers and operators.

ABSTRACT(Continued)

The GANNT chart methodology offers several advantages. Firstly, it analyzes the correlation effects between variables using Hotelling T^2, allowing for a more comprehensive understanding of process variation. Secondly, it leverages graph theories to retain and utilize knowledge from previous successful operations, facilitating continuous improvement. Lastly, the system is trained using Artificial Neural Networks, enabling it to provide solutions to future challenges based on learned patterns from past operations. The proposed model is validated in the cement industry to assess its effectiveness and practicality. The results demonstrate that the GANNT chart effectively addresses the identified gaps in the application of SPC techniques to the cement production process. The model's ability to accurately detect process deviations and provide insights into the causes of variation contributes to improved productivity, quality, and overall performance. As a future research direction, this study highlights two suggestions. The first one is examining and extending the assumptions to design this model in such a way that it considers different scenarios not covered by this research. The second direction is extend the implementation of GANNT chart to various industries, including service giving industries, and study and explore its applicability. In conclusion, bridging the gap between theory and practice, this research aims to contribute to the advancement of multivariate process control to the industry, ultimately leading to enhanced operational efficiency and product quality.

Keywords: Neural Network, Graph theory, Hotelling T2, GANNT chart, VARIMA model, cement process control

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AAiT……….	Addis Ababa Institute of Technology	LCL………..	Lower Control Limit
AAU………..	Addis Ababa University	MSPC………	Multivariate statistical Process Control
ANN………..	Artificial Neural Network	MCUSUM…	Multivariate Cumulative Sum
ARL ……...	Average Run Length	MEWMA….	Multivariate Exponentially Moving Average
CCP………	Control Chart Pattern	NN………….	Neural Network
CL…………	Centerline	PCA………...	Principal Component Analysis
CUSUM…..	Cumulative sum	PLS………..	Partial Least Square
DT………..	Decision Tree	SPC …………	Statistical Process Control
DNN………	Deep Neural Network	SQC………….	Statistical Quality Control
EPC………..	Engineering Process Control	T^2……………	Hotelling T-Square
EWMA…….	Exponentially Weighted Moving Average	UCL………….	Upper Control Limit
GT…………	Graph Theory	VSM…………	Vertical Support Machine
IID………..	independent and identically distributed		
K-NN……..	K-Nearest Neighbor		

Chapter One

BACKGROUND AND JUSTIFICATION OF THE STUDY

This chapter introduces the main concern of the dissertation, which is Statistical Process Control (SPC). It further discusses the background of the study where the research gap is identified in the industry. Then the statement of the problem is clearly stated followed by the objective of the research, significance, scope, and limitation of the study. Finally, the organization of the document is described.

1.1 Introduction

Industries strive to continually improve their productivity and performance by optimizing their processes and promptly returning to normal operation in the face of disturbances. Failure to do so can result in the wastage of valuable resources and render them unable to compete effectively in the market. Recognizing the importance of overcoming this challenge, extensive research has been conducted to develop a multivariate process control procedure capable of enhancing the monitoring and controlling capabilities of process engineers and operators(Woodall & Montgomery, 2014).

Multivariate Statistical Process Control (MSPC) encompasses a technique or procedure aimed at simultaneously monitoring or controlling two or more quality characteristics within a process. Building upon the foundations of Univariate Statistical Process Control, MSPC provides a comprehensive framework for analyzing multiple variables and their interdependencies. (Bersimis et al., 2005) Alongside widely recognized techniques such as Hotelling's T2, Multivariate Exponentially Weighted Moving Average (MEWMA), and Multivariate Cumulative Sum (MCUSUM), a multitude of researchers has made substantial contributions to the field, expanding the repertoire of control procedures for related variables.

The management of a multiple input and multiple output system presents unique challenges due to its complexity. In such systems, there is a need to consider the interactions and dependencies among various process parameters and quality characteristics. This complexity necessitates the application of Multivariate Statistical Process Control (MSPC). The utilization of multivariate statistical process control

(MSPC) techniques in industries have witnessed further enhancements through the adoption of more advanced process control methods(Qin & Chiang, 2019). These cutting-edge techniques encompass data mining and machine learning, both of which are integral components of artificial intelligence. With the advancements in sensor technology and online data collection systems, an abundance of relevant data can now be collected within a significantly shorter time frame. The process of extracting valuable information from these vast data sets is the primary function of data mining, while the development of algorithms that emulate the human mind and comprehend the flow of data falls within the domain of machine learning. The widespread introduction and implementation of these techniques in various industries have yielded substantial improvements in both productivity and product quality(Dhini, 2016).

The field of data mining plays a pivotal role in the realm of process control (Noskievicova et al., 2020). It involves the exploration and analysis of large data sets to discover patterns, correlations, and hidden insights. By leveraging sophisticated data mining techniques, organizations can uncover valuable information that was previously obscured within the vast volumes of collected data. This information can then be utilized to identify factors affecting process performance, detect anomalies, and optimize process parameters. Moreover, data mining enables organizations to gain a deeper understanding of complex relationships among multiple variables, contributing to more effective decision-making and improved control over the manufacturing process.

In addition to improving process control, data mining and machine learning techniques also contribute to advancements in other areas of industrial operations(Demircioğlu Diren et al., 2020). For instance, these techniques can be applied to predictive maintenance, enabling organizations to identify maintenance needs proactively and avoid unplanned downtime. They can also support demand forecasting and supply chain optimization, facilitating better inventory management and resource allocation. Hence, Multivariate process control techniques are part of the industry system that should be integrated to different industrial operations.

Furthermore, the widespread adoption of data mining and machine learning techniques has led to the development of intelligent systems capable of autonomous decision-making and adaptive control. These systems continuously learn from new data, refine their models, and evolve their control strategies, leading to optimized and self-improving processes. This level of automation and intelligence has the potential to revolutionize industries by enabling more efficient and flexible production systems.

The focus of this research is to address the problem of identifying and resolving issues in complex production systems efficiently. By leveraging multivariate process control techniques, industries aim to gain a comprehensive understanding of the interdependencies and interactions among various process variables. This allows for the early detection and diagnosis of deviations, which is crucial in swiftly determining the root cause of the problem. By promptly identifying the underlying issue, industries can take appropriate corrective measures, minimizing disruptions and ensuring a rapid return to normal operation.

1.2 Background and Study Justification

In today's manufacturing industries, the utilization of complex process control system models with multiple inputs and multiple outputs has become increasingly exercised. This technological advancement has brought about substantial benefits, particularly within the cement sector. By employing these intricate models, the cement industry has been able to revolutionize its operations and processes, leading to improved efficiency, enhanced productivity, and optimized resource allocation.

The utilization of complex process control models in the cement industry also enhances decision-making processes. By providing a comprehensive overview of the production system, these models enable managers to make data-driven decisions based on accurate and real-time information. They facilitate forecasting the process flow, ensuring optimal resource allocation and reducing the risk of process disruptions.

The cement industry is the backbone of the construction sector. Infrastructures like bridges, dams, power stations, hospitals, schools, city development plans, housing programs, industries, and similar mega-construction projects are realized due to the supply of cement. Cement demand is increasing every year as a result of industrialization, urbanization, and population growth. Such noticeable consumption growth has changed the lifestyle and quality of life of the population. Furthermore, the construction industry employs a large number of workers. Hence cement has a significant impact on social, cultural, and economic development.

Currently, the cement industry is challenged by sustainability issues as a result of the high volume of depletion of resources, high energy consumption, and large impact on the environment. Cement is the second most consumed material on the planet next to the water, with about 4.10 billion tonnes/year for the year 2018(Miller et al., 2021). It is estimated that one-third of the total natural resources consumed by industries are used for cement production. To manufacture one ton of Ordinary Portland Cement (OPC), approximately 1.5 tons of raw materials are required. As a result, the depletion rate of natural resources is increasing exponentially as production volume is increasing every year.

The cement industry is the third-largest industrial energy consumer, with 7% of the global industrial energy use(IEA, 2018). Each tone of clinker needs 2800 MJ of thermal energy and 103 to 110KWh of electrical energy. Hence, the energy reduction strategy in the cement industry is the most sensitive issue to sustain in the market.

Cement has emitted about 900kg of CO_2 into the atmosphere per ton of production of clinker. It is estimated that 8% of global CO_2 emission is from both the chemical decomposition of raw materials and the thermal combustion of fuel to burn the raw material in the production of cement(Poudyal & Adhikari, 2021)

Long-term and short-term strategies are designed to overcome the above challenges faced by cement industries. Large-scale replacement of cement with other binding materials is the long-term strategy. This strategy is applied in pilot projects and it was successful, however, cannot be possible to replace the current type of cement in the next decade due to the large volume of global demand per year (Habert et al., 2020). Optimization across the value chain in the short-term strategy aims at reducing CO_2 emission, energy consumption, waste reduction, and optimal raw material consumption. One of the optimization techniques is statistical quality control (SQC) which applies statistical tools to improve the productivity and quality of the product (Matthew Carlyle et al., 2000).

The statistical process control (SPC) charting method is the foundation of SQC which is introduced only a decade while mass production of Portland cement was introduced. Two motives drive us to study SPC as an optimization technique for the cement industry. These are the production system and the technological advancement of the cement industry. Cement production is a complex system. In general, there are five stages to completing the cement production process. These are the raw material preparation stage, raw mix, and milling stage, clinker forming stage, cement milling stage, and packing and delivering stage. Each stage comprises many parameters that influence the production system. SPC plays the key role in material selection in quarries; raw material blending according to the raw mix design; optimizing both thermal and electrical energy consumption in the clinker forming stage; maximizing additives to the cement milling stages; and monitoring the compressive strength, fineness, and other product quality parameters.

The cement industry has recorded remarkable technological advancement over the last century. This technological change in types of machinery and control devices has influenced the way SPC is applied to the industry. Online data collection systems, smart sensor and equipment technology, condition monitoring equipment, transformation to digitalization, application of intelligent optimal control and industry 4.0, and decision support systems further drive the application of process control techniques to make the cement production system more productive(John et al., 2019). The technological progress is summarized and presented by Xu et. al(Trout, 2019) as shown in **Error! Reference source not found.**.

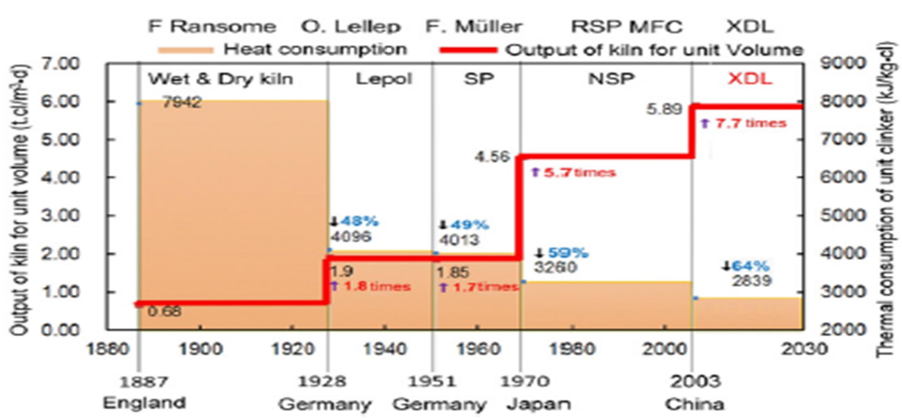

Figure 0-1 The technical progress of the cement industry since 1887(Blezard,1998)

Both statistical process control (SPC) techniques and cement production systems are profoundly improved over the last 100 years. However, the authors observed that there are limited publications on the application of SPC techniques to cement industries as indicated in Figure 0-2. The current world demands the cement production system to meet environmental, economic, societal, and technological requirements to sustain itself in the market. Hence, cement industries design strategies to fulfill these requirements and search for opportunities and maximize outcomes from optimization techniques. On the other hand, wide research ideas are introduced on statistical process control like variable sample size and sampling interval methods, economic designs, attribute data methods, charts based on auto-correlated observations, multivariate and non-parametric methods as well discussed in the literature review on this paper. Hence, this research follows the following research directions to examine this topic. What are the basic researches done on statistical process control chart design techniques? Which direction the basic research is proceeding? Do these basic research findings satisfy the demand of cement industries? The main objective of the research is then to survey the advances of statistical process control chart techniques over the last century and investigate how far they can fill the demand of the cement industries.

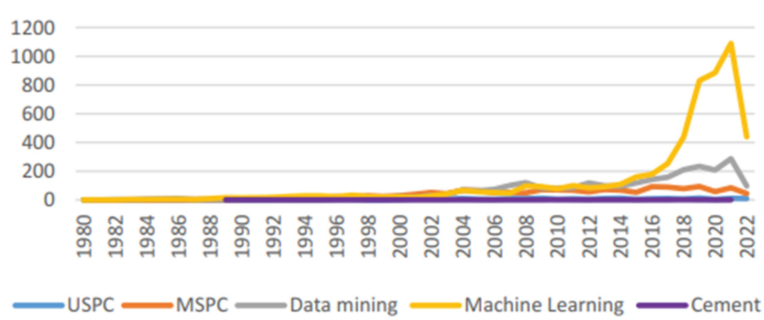

Figure 0-2 Number of articles published per category

1.3 Statements of the problem

In the global context, there is a significantly increasing number of multivariate processes monitoring applications in industries. At the same time, the rapid growth of automatic data capture systems has provided massive data collection to analyze valuable information. A selection of a method that is capable of conducting fast and accurate fault detection analysis is extremely important. The method should not only detect the out-of-control data accuracy, but it should also be able to immediately identify the assignable cause. Only a few studies have been conducted in integrating the out-of-control detection and identifying the variable which causes a control signal (Bersimis et al., 2021)(Bersimis et al., 2005),(Mason et al., 1997).

Literature discloses that various multivariate control charts like Hotelling's T2, multivariate cumulative sum, and multivariate exponentially weighted moving average charts have been designed for detecting mean shifts. The main problem of such charts is that they can detect an out-of-control event but do not directly determine which variable or group of variables is causing the out-of-control signal (Bersimis et al., 2021);(Guh & Shiue, 2008); (Mason et al., 1997). This theoretical gap is identified by researchers and it is an active research area for statisticians and mathematicians, and it can be filled by theoretical formulation of methods.

The major drawback of most multivariate control chart procedures is that they don't directly provide the information an operator needs when the control charts signal. To be more specific, the operator needs to know which variable(s) caused the out-of-control signal(Bersimis et al., 2017), (Bersimis et al., 2021).

Some attempts have been done to integrate data mining and machine learning techniques into MSPC problems. However, there is no consensus yet that stated a specific technique is the most suitable for any circumstances(Shao & Lin, 2019). Only a few efforts have been performed to integrate these algorithms and control charts. Hence, more studies are required to fill this knowledge gap.

1.4 Research question

Research discloses that there are many unsolved problems in the area of statistical process control. Some of them are multivariate methods, the effect of estimation error, short-run method, auto-correlated data, variable sampling methods, economic design methods, change point estimation, engineering process control, and SPC, non-parametric approaches are some of the most active research areas in SPC(Firat & Çilan, 2000), (William H. Woodall, 1999). This research focuses on one of the research directions, which is multivariate methods.

This dissertation focuses on addressing three key research questions related to Multivariate Statistical Process Control (MSPC) techniques:

- What are the major limitations of the existing MSPC techniques in monitoring and controlling manufacturing process?
- What strategies and approaches can be employed to develop a procedure for a multivariate statistical process control (MSPC) method that effectively addresses the limitations of existing techniques and enhances its accuracy, adaptability, and performance in monitoring and controlling multivariate process?
- What are the effective approaches and techniques for validating a developed model in order to assess its performance, reliability, and generalizability?

The first research question aims to identify and understand the limitations and constraints associated with the current MSPC techniques. Through an extensive review of the literature, analysis of case studies, and examination of real-world applications, the dissertation seeks to provide a comprehensive understanding of the challenges and shortcomings faced by the existing MSPC techniques. By identifying these limitations, the research aims to lay the groundwork for developing a more effective and robust procedure.

The second research question focuses on the development of a new procedure for multivariate statistical process control that overcomes the limitations identified in the first research question. This objective entails designing an innovative framework that addresses the specific challenges and shortcomings of the current MSPC techniques. By incorporating advanced statistical methods, data analysis techniques, and relevant domain knowledge, the dissertation aims to propose a comprehensive procedure that enhances the monitoring, control, and optimization of multivariate processes. The research seeks to provide practical guidelines and recommendations for developing an improved MSPC method.

The third research question centers on the validation of the developed model or procedure. Once the new procedure for multivariate statistical process control is designed, the dissertation aims to validate its effectiveness and applicability. This involves conducting rigorous testing and evaluation of the model in real-world scenarios or using simulated data. The research seeks to assess the performance, accuracy, and robustness of the developed model, ensuring its reliability and suitability for practical implementation. By validating the model, the dissertation aims to provide empirical evidence of its efficacy and establish its credibility in addressing the limitations of existing MSPC techniques.

By addressing these research questions, the dissertation aims to contribute to the advancement of knowledge in the field of MSPC. The findings and recommendations derived from this research endeavor will offer valuable insights for researchers, practitioners, and industries seeking to enhance their process control practices and improve the quality and efficiency of their operations.

1.5 Objectives of the Study

1.5.1 General Objective

The general objective of this research is to develop a comprehensive process control procedure that takes into account the multivariate impact of the variables within the specific production system.

By recognizing the interconnected nature of these variables, the research aims to design a robust and efficient control mechanism that effectively monitors and regulates the entire system. Through the integration of advanced statistical methods, data analysis techniques, and domain-specific knowledge, the objective is to create a procedure that optimizes process performance, enhances product quality, and ensures the overall efficiency and sustainability of the production system. By achieving this objective, the research aims to provide valuable insights and practical guidelines for process engineers and operators to implement a multivariate-focused process control approach in their respective industries.

1.5.2 Specific Objective

The research at hand encompasses specific objectives that aim to address critical aspects related to Multivariate Statistical Process Control (MSPC) procedures. These objectives serve as guiding principles throughout the research process, enabling a comprehensive exploration of the limitations, the development of an algorithm to overcome these limitations, and the validation of the model in a practical setting. The specific objectives are outlined as follows:

- Identify the limitations of existing MSPC procedures
- Develop an algorithm that fills the gap of the existing limitation
- Validate the model in the case company and demonstrate the applicability of the model

The first objective revolves around conducting an in-depth analysis and evaluation of the current MSPC procedures in place. This entails identifying and understanding the inherent limitations and shortcomings associated with the existing methods employed in process control. By thoroughly examining the strengths and weaknesses of these procedures, the research aims to gain valuable insights into the areas that require improvement and enhancement.

The second objective builds upon the findings from the first objective, the second objective involves the development of an algorithm that specifically targets and addresses the identified limitations. This objective entails the creation of a novel and innovative approach that fills the gaps left by the existing MSPC procedures. The algorithm will be designed to provide a more robust and comprehensive framework for monitoring, controlling, and optimizing multivariate processes. By developing this algorithm, the research strives to offer a practical solution that surpasses the constraints of the current procedures.

The third objective centered on the practical application and validation of the developed algorithm. This involves implementing the model in a real-world case company, specifically selected for its relevance to the research context, such as a manufacturing or industrial setting. The research aims to demonstrate the efficacy and applicability of the model by conducting rigorous testing and evaluation within this specific case company. By validating the model, the research seeks to showcase its effectiveness in addressing the limitations of existing MSPC procedures and its potential for wider adoption in similar industries or contexts.

In pursuing these specific objectives, the research strives to make significant contributions to the field of MSPC. By identifying and addressing the limitations of existing procedures, developing an innovative algorithm, and validating its applicability in a practical setting, the research aims to enhance the overall effectiveness and efficiency of multivariate process control. Through this work, the research seeks to advance the knowledge and understanding of MSPC, provide practical solutions for process engineers and operators, and ultimately contribute to the improvement of quality and productivity in various industries.

1.6 Significance of the Study

In today's fiercely competitive market, companies are constantly striving to develop effective strategies that will enable them to survive and thrive. One such strategy involves delivering high-quality products to customers and promptly responding to their requirements. Achieving this goal necessitates the implementation of controlled and well-designed production systems that are capable of meeting customer demands and ensuring product quality.

Introducing Multivariate Statistical Process Control (MSPC) methods offers numerous benefits and advancements to the field of process control. Firstly, it provides a fresh and improved technique for effectively controlling production processes. By utilizing MSPC methods, process engineers and operators can enhance their decision-making processes, leading to more informed and optimized production decisions. This, in turn, reduces the overall decision-making time and minimizes waste, thus improving operational efficiency.

One notable contribution of MSPC methods is the introduction of a novel approach to creating process control charts. By combining traditional and advanced process control techniques, MSPC methods offer a new way of monitoring and controlling production processes. This integration allows for a more comprehensive analysis of process data and facilitates the detection of potential deviations or abnormalities, leading to timely corrective actions.

Furthermore, the introduction and implementation of MSPC techniques contribute to the overall awareness and understanding of these methods within the industry. As more organizations adopt and utilize MSPC methods, the knowledge and proficiency in these techniques are improved across the board. This increased awareness fosters a culture of quality and emphasizes the significance of maintaining high standards in production and manufacturing processes.

Undoubtedly, conducting research on MSPC and seeking out better methods for improving quality in the production and manufacturing industries is of paramount importance. By continually exploring and enhancing MSPC techniques, researchers and practitioners can contribute to the continuous improvement of product quality, customer satisfaction, and overall competitiveness in the market. This research not only empowers organizations to meet and exceed customer expectations but also solidifies the importance of the quality movement within the industry.

In summary, given the intense competition in today's market, companies must embrace strategies that prioritize the delivery of high-quality products and responsiveness to customer needs. Implementing controlled and well-designed production systems is essential for achieving these objectives. The introduction of Multivariate Statistical Process Control methods brings a range of benefits, including improved process control techniques, enhanced decision-making processes, reduced waste, innovative process control chart creation, increased awareness of MSPC techniques, and the reinforcement of quality movement. Therefore, conducting research and continually seeking better methods for quality improvement in production and manufacturing industries is a vital endeavor that drives progress and success in today's business landscape.

1.7 Scope of the Study

This research endeavors to make significant contributions to the development of an algorithm capable of analyzing complex multiple input multiple output (MIMO) systems. The primary objective is to design an algorithm that can effectively identify and interpret patterns within the cement production system, which serves as the case company for this study. As a result, the scope of this research is specifically focused on the cement industry and the unique challenges it poses.

In the realm of multivariate statistical process control (MSPC) techniques, a well-designed algorithm plays a pivotal role in defining and understanding the underlying patterns within a plant. It is essential to recognize that all the data collected, including information about raw materials, process data, and product data, follow a discernible pattern. Consequently, the ultimate aim of the algorithm is to identify and extract this pattern from the data. The performance and effectiveness of the algorithm are assessed by comparing the actual collected data with the expected output derived from the identified pattern. It is within this scope that this research seeks to make valuable contributions.

By developing an algorithm specifically tailored to the cement production system and validating it using real-world data from the case company, this research strives to enhance the understanding of multivariate process control in this industry. The algorithm's ability to accurately capture and interpret patterns within the complex MIMO system will be evaluated and measured against the expected outcomes. Through this research, the goal is to bridge the gap between theoretical advancements in MSPC techniques and their practical application within the cement production context.

Given the limited scope of this study, focusing solely on the cement production system, it is important to acknowledge the inherent challenges and complexities associated with the industry. The algorithm's success in effectively analyzing the multivariate data collected from the cement factory provides valuable insights into optimizing processes, improving product quality, and enhancing overall operational efficiency. Moreover, this research aims to contribute to the broader field of multivariate statistical process control by showcasing the potential of tailored algorithms in tackling the unique complexities of specific industries.

1.8 Limitations of the Study

The field of multivariate statistical process control has received significant attention in academic publications, demonstrating its status as an intensively researched area. However, despite the wealth of research, only a limited number of approaches have gained popularity and practical implementation in industries. In this research, the algorithm that has been developed is assumed to apply to any chemical industry, although the data collection and analysis will be constrained to cement factories. This limitation arises from resource constraints, such as financial limitations for data collection and time constraints to observe the algorithm's results in other industries, such as plastic, pulp and paper, pharmaceuticals, and others.

To achieve the primary objective of this research, a comprehensive approach has been adopted, integrating concepts and applications from various disciplines. For instance, principles of quality management from industrial engineering, multivariate analysis from statistics, graph theory from mathematics, and programming and database expertise from computer science have been planned to be employed. However, it is important to note that integrating such diverse disciplines for a specific purpose is a time-consuming task. The limited time available for the study, coupled with the depth of the research, may potentially impact the overall quality and breadth of the study's findings and conclusions.

Despite these challenges, the research aims to make valuable contributions to the field of multivariate statistical process control and its applications in the chemical industry, specifically within the context of cement factories. By focusing on this industry, the study can delve deeply into the intricacies and nuances specific to cement manufacturing, offering insights that can be potentially extrapolated to other chemical industries. While the limitations of resources and time may pose constraints, the research is designed to provide practical and applicable solutions to enhance process control and quality management in cement factories.

It is important to acknowledge that the constraints imposed by the limitations of resources and time do not invalidate the significance and potential impact of the research. Instead, they serve as valuable insights for future studies and highlight the importance of conducting further research to explore the application of the developed algorithm in a broader range of industries. Despite the potential limitations, this research represents a step forward in advancing the understanding and implementation of multivariate statistical process control, providing a foundation for future investigations in related disciplines and industries.

1.9 Organization of the Study

This dissertation is a comprehensive exploration of multivariate process control in the context of cement technology. It is organized into seven chapters, each contributing to a deeper understanding of the subject matter.

Chapter one serves as the introduction, providing an overview of the general background and justifications for the study. This chapter is divided into nine sections, beginning with an introduction and background of the study. It then delves into the statement of the problem, research question, and objectives of the research in sections three, four, and five, respectively. The significance, scope, and limitations of the study are discussed in sections six, seven, and eight. The final section, section nine, outlines the organization of the entire document, giving readers a roadmap of what to expect in subsequent chapters.

Chapter two focuses on the literature review, where a comprehensive examination of existing research and knowledge on the topic is conducted. This chapter consists of eight sections, with the first two sections discussing the framework and methods employed in conducting the literature review, as well as the resources and materials utilized. The subsequent sections delve into various aspects of process control, including univariate, multivariate, and projection-based techniques, as well as data mining and machine learning in statistical process control (SPC). The final section provides a summary of the key findings derived from the literature review.

Chapter three is dedicated to the research design and methodology, offering insights into the approach taken to conduct the study. This chapter details the research design, data collection methods, data analysis techniques, and any ethical considerations that were taken into account during the research process.

In chapter four, algorithm development for multivariate process control analysis takes center stage. This section provides a detailed explanation of the algorithms developed and utilized in the analysis of multivariate process control data. The chapter highlights the steps taken to ensure the accuracy and effectiveness of the algorithms.

Chapter five serves as a pivotal point in this dissertation as it shifts the focus towards the practical implementation of the model that was developed in chapter four, specifically within the context of the cement industry. This chapter delves into the intricacies and challenges associated with integrating the developed model into the existing processes and systems within the cement industry. By exploring the practical aspects of implementing the model, such as data integration, software integration, and process adaptation, this chapter aims to provide a comprehensive understanding of the feasibility and effectiveness of applying the developed model in real-world scenarios. The chapter also examines the potential benefits, limitations, and considerations that need to be taken into account when deploying the model within the cement industry. Overall, chapter five serves as a bridge between theory and practice, shedding light on the practical implications and outcomes of implementing the model in a tangible industrial setting.

Chapter Six focuses on the validation of the developed model for a specific case company within the cement industry. This section discusses the process undertaken to validate the model's applicability and effectiveness in a real-world scenario. It examines the results obtained from the case study and draws conclusions based on the findings.

The final chapter, chapter seven, concludes the dissertation. It is divided into three sections: the conclusion, recommendation, and future research direction. The conclusion offers a concise summary of the study's main findings, highlighting the contributions and implications of the research. The recommendations section provides actionable suggestions based on the study's outcomes, offering insights into potential improvements or interventions that can be implemented in the cement industry. Lastly, the future research direction section identifies areas for further exploration and suggests potential avenues for future studies.

In conclusion, this dissertation provides a comprehensive exploration of the theory of statistical process control techniques to overcome the limitation of process control in the context of cement technology. It encompasses a range of chapters, covering the introduction, literature review, research design, algorithm development, cement technology overview, case study validation, and concluding remarks. The dissertation aims to contribute to the field by providing valuable insights, practical recommendations, and avenues for future research within the cement industry and multivariate process control domain.

Chapter Two

LITERATURE REVIEW

This chapter presents an extensive literature review on statistical process control techniques. This includes both traditional and current practices in process control within production systems. Furthermore, this chapter seeks to identify gaps in knowledge regarding process control techniques and proposes a conceptual framework as a means for filling them.

2.1 Introduction

Initial surveys on statistical process control literatures have exposed researchers' rapid adoption of various process control techniques throughout time. Literature in this field is abundant and approaches have varied over the years; one approach utilized as a benchmark was innovating statistical process control (SPC) techniques, so to track its development a research study used a semi-systematic or narrative literature review approach; this flexible framework permits detailed investigation while offering cohesive synthesis through narrative means.

To gather relevant information, prominent academic sources such as Science Direct, Google Scholar, JSTOR, Scopus, Mendeley, Emerald, Elsevier, Taylor and Francis, and IEEE Explore were utilized. A thematic search was conducted within these sources, resulting in the extraction of over 190 pertinent articles from 34 different journals. Additionally, books, theses, and dissertations were consulted as supplementary references to the main research topic.

The first section of this chapter is an introduction. Statistical process control has a century-long history since its inception, during which numerous methods have been introduced. These methods are classified based on their evolution and approach, providing a framework for summarizing the extensive literature.

Subsequent sections of the chapter delve into each category of statistical process control techniques. Section two focuses on the evolution and development of univariate statistical process control techniques. Section three discusses the extension of these techniques into the multivariate domain, while section four explores advanced non-statistical techniques within Multivariate Statistical Process Control methods. Sections five and six are dedicated to data mining and machine learning techniques, respectively. Finally, section seven presents a summary of the findings from the literature review and identifies any remaining gaps.

The advancement of the SPC methods can be categorized into four periods. The first period started at the time when the II Industrial Revolution has immerged. The II Industrial Revolution is characterized by the introduction of mass production and electrical energy. The shift from batch production to mass production increases the production volume and the method for quality control during that time which was inspection is challenged so sampling and application of control charts were introduced. The second period is introduced during the mid of 20th century when the Second World War exploded. During this time, production volume is increasing alarmingly and precision of product quality was highly demanded. This period is characterized by the introduction of multivariate statistical process control methods.

The third period is introduced during the middle of the 1970s. This is the time when the III Industrial Revolution was introduced. This time is marked by the introduction of computers and sensors to automate production systems. Due to abundant information accumulated in computer within a short period, data-mining techniques to monitor the production system is introduced. Engineering process control and statistical process control methods are married during this time. The fourth period is the time when machine learning is introduced into the production system. An intelligent system that learns from the environment is introduced during this time.

Within all these periods, the contribution of statisticians to the development of industrial revolutions to develop quality products and services was remarkable. It is reasonable to conclude that statistics is still the foundation of the concept of quality improvement.

2.2 Univariate Statistical Process Control

The following search queries are used for the first identified keyword.

TITLE-ABS-KEY (*univariate* AND *statistical* AND *process* AND *control*)**.**

The search is conducted on the Scopus database and it generates 654 documents. To ensure a comprehensive search of the topic, other databases are also reviewed and the result database is collected in the same folder. Use inclusive and exclusive criteria like the language of the document, subject area, and keywords to the specific area and reduce the number of documents. To further explore the research landscape, employ the co-citation analysis method and consider the cited-authors as a unit of analysis in the VOSviewer software, all the influential authors in the database are identified as shown in Figure 0-1.

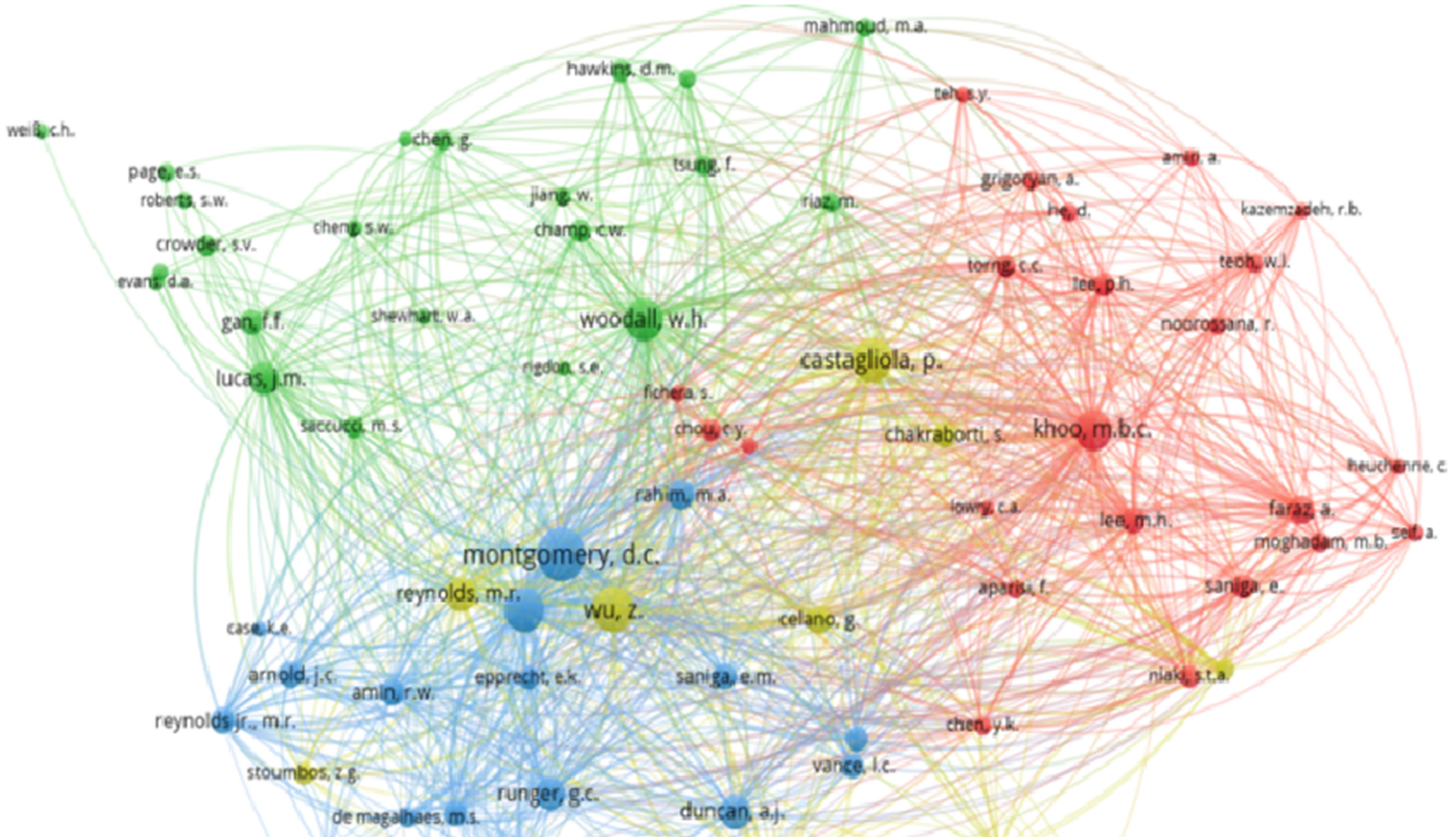

Figure 0-1 Authors in Univariate Statistical Process Control

The review of the major authors, identified by the VOSviewer software reveals that the research direction to the types of univariate statistical process control was categorized into the Shewhart control chart, cumulative sum control chart, and exponentially weighted moving average control chart, As well as the design aspect is categorized into the heuristic, statistical, economical, and the economic statistical control chart design.

During the initial phase of field development, rigorous mathematical and statistical analyses were applied to create procedures to detect changes over time in processes. By this time, Shewhart first proposed his groundbreaking idea of statistical process control (SPC). Submitting his first article for approval within his organization in 1924, then published it as an article (Shewart, 1926). Shewhart demonstrated the significance of control charts to quality concepts by publishing two books: in 1931 “Economic Control of Quality in Manufactured Products" by himself (Shewhart) and then in 1939 with Deming (Shewhart & Deming) respectively. "The Western Electric Company" later built on these efforts by publishing their statistical quality control handbook.

Shewhart's control chart was devised as an answer to mass production challenges during the Second Industrial Revolution. Based on the theory of distribution and sampling, this design allows the monitoring of a single quality variable. Assuming normality in data distribution, the control chart design involves three parameters. They include "selection of sample size", "sampling interval", and "width of control chart". Statistical criteria determine appropriate control limits by specifying a maximum probability for type I errors (False Alarm) and/or type II errors (failure to sound an Alarm). The control chart, a visual display of quality characteristics, can serve as empirical evidence to test whether or not your process is under statistical control (Murthy & Rambabu, 1997). A point plotting within the control limits is equivalent to failing to reject the hypothesis of statistical control while outsider plots indicate rejection(Murthy & Rambabu, 1997).

Heuristic Shewhart Control Chart Design is a straightforward method that takes into account statistical criteria and practical experience and suggests using samples of size 5 with three Sigma Control Limits and a sampling frequency of one hour for its creation. Such charts were effective tools for monitoring variability in parameters as well as displaying process patterns such as normal, increasing trend decreasing trend, cyclic systematic mixture upward shift or downward shift patterns; only normal being consistent with an unassignable cause process and all others not unnatural or unexpected.

Shewhart's pioneering contributions, bolstered by influential publications and resources like his books and the statistical quality control handbook, provided an important basis for the advancement and widespread adoption of statistical process control techniques within quality management. These seminal works laid a strong foundation for subsequent research as well as the continued evolution of SPC methodologies over subsequent decades.

Duncan introduced an improvement to the Shewhart control chart known as an economic model for designing Shewhart-type charts(Duncan, 1971). When selecting sample sizes, intervals, and control widths optimally, various risk and cost considerations must be considered, along with several process variables like frequency of occupancy per shift in the process, cost of sampling/inspection fees/penalty costs for defectives/false alarms as well as investigation fees incurred while trying to assignable causes as well as process correction costs are taken into account. Duncan also considers cases with multiple assignable causes in his paper and developed a general model to estimate total costs concerning various parameters.(Gibra, 1975), (D. C. Montgomery, 1980), (Lorenzen & Vance, 1986), and (Ho & Case, 1994) conducted literature surveys on the economic design of control charts. Lorenzen & Vance proposed an overall model for such designs (Lorenzen & Vance 1986).

Saniga(Saniga, 1989) proposed an economic statistical control chart design with statistical constraints placed upon economic models to meet the industry's demand for low process variability and long-term product quality. He found this method superior to statistical control chart designs as it ensures long-term quality while decreasing the variance of quality characteristics distributions. This type of economic design ensures long-term product quality with lower variance than economic designs while increasing long-term product quality and decreasing variance over time.

Research also proposed an adaptive control chart design. With this model, one or more design parameters may change as a function of process data, with control charts created based on statistical criteria or economic statistics criteria (Table 0-1). A selection literature review summary can be seen here:

Table 0-1 Adaptive control chart literature

Type	statistical Criteria	Economical Criteria	economic statistical Criteria
Variable sampling size (VSS)	Flaig(Flaig, 1991), Daudin(Daudin, 1992), Prabhu at al.(Prabhu et al., 1994), Costa(Costa, 1994), Zimmer et al(Zimmer et al., 1998), Castagliola et al.(Castagliola et al., 2012)	Flaig(Flaig, 1991), Park and Reynolds(Park & Reynolds, 1994)	
Variable sampling interval (VSI)	Reynolds et al. (Reynolds et al., 1988), Reynolds(Reynolds, 1989), Runger and Pignotiello(Runger & Pignatiello, 1991), Zhang et al.(Zhang et al., 2012)	Das & Jain(Das & Jain, 1997), Bai and Lee(Bai & Lee, 1998), Reynolds(Reynolds, 1989), Yu et al.(Yu et al., 2007)	Lee et al.(T. H. Lee et al., 2016)
Variable sampling size and sampling interval (VSSI)	Prabhu et al.(Prabhu et al., 1994), Costa(Costa, 1997), Carot et al.(Carot et al., 2002)	Park and Reynolds(Park & Reynolds, 1999),	Prabhu et.al (Prabhu et al., 1997)
Fully adaptive (VP)	Costa(Costa, 1999)	De Magalhaes et al.(De Magalhães et al., 2001), Costa and Rahim(Costa & Rahim, 2013)	De Magalhaes et.al (De Magalhães et al., 2002)

Shewhart-Control Charts have limitations that render them less sensitive to small and medium process shifts; as a result, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) have been introduced as alternative measures of process analysis. Shewhart's Mean Control Chart has proven extremely efficient when applied with a shift magnitude of 1.5s or larger(D. Montgomery, 2009).

Page (Page, 1961) introduced one and two-sided CUSUM procedures for detecting small mean shifts. In this method, known as tabular or logarithmic plotting, the cumulative sum of sample means against their reference values is plotted against sample numbers; when this exceeds some threshold called decision interval H then corrective action must be taken immediately; giving CUSUM analysis an element of memory. Page also introduced average run length (ARL), defined as the average number of samples taken before an out-of-control signal; analytical studies exist for selecting parameters K and H that will give optimal ARL performance.

Bernard (Barnard, 1959) first proposed an alternate approach for designing CUSUM charts known as a V-mask, determined by distance 'd' behind the leading point and angle o between arms of V and the horizontal plane. Lucas proposed his modified V-mask, consisting of a parabolic-shaped mask, as an alternative method. Ewan & Kemp ((W. D. Ewan & Kemp, 1960) and Kemp (Kemp, 1962) provide nomograms showing relationships among L, H, and O for normally distributed variables.

CUSUM has been studied extensively by numerous authors, such as Ewan & Kemp((W. D. Ewan & Kemp, 1960), Page (Page, 1961), Kemp ((Kemp, 1962), Beattie(Beattie, 1962), Ewan(A. D. Ewan, 1963), Lucas(Lucas, 1976), Hawkins((Hawkins, 1981),(Hawkins, 1993) & Gan (Gan, 1991), Woodall & Adams((Woodall & Adams, 1993). Research by Bissell (Bissell, 1969) highlighted the effectiveness of Cumulative Sum Control (CUSUM) charts versus Shewhart's for detecting smaller variations in average levels. CUSUM charts are widely utilized across industries as powerful and easy-to-use solutions that quickly identify out-of-control processes ((Koshti, 2011) so corrective action may be taken on them to rectify them.

Taylor first investigated the economic design of CUSUM charts ((Taylor, 1968). Subsequently, Goel & Wu (Goel & Wu, 1973) and Chiu (Chiu, 1974) proposed similar models and algorithms for determining economically optimal designs of CUSUM charts as well as reporting some results of sensitivity analyses.

Research also explored adaptive control chart designs for cumulative sum charts. Reynolds, Amin, Arnold(Reynolds et al., 1990), Reynolds(Reynolds, 1995) presented VSS CUSUM charts while Anandi et al((Keats et al., 1995) studied VSSI CUSUM charts; VSSI CUSUM charts are investigated by Arnold & Reynolds(Arnold & Reynolds, 2001).

Robert (Roberts et al., 1959) first introduced a geometric moving average chart as one method for univariate statistical process control, later to become known as exponentially weighted moving average (EWMA). Reynolds and Arnolds(Reynolds & Arnold, 2001) and Jones (Jones, 2002) proposed statistical and economic designs of EWMA control charts with estimated parameters respectively while Montgomery et al. (D. C. Montgomery et al., 1995) proposed economic designs while Park et al (Park et al., 2004) offered economic-statistical designs; more recently Gracia Diaz Aparisi (García-Díaz & Aparisi, 2005) suggested economic-statistical designs with estimated parameters and estimated parameters respectively.

Researchers have developed an adaptive control chart design for EWMA. Sacuucci et al.(Saccucci et al., 1992) presented VSSI EWMA charts while Reynolds(Reynolds, 1996) , also studied VSS EWMA charts; these findings are detailed further by Reynolds and Arnold((Reynolds & Arnold, 2001) who conducted further investigations of VSS EWMA charts as well.

Nantawong (Nantawong et al., 1989) conducted experiments to assess the influence of three factors (sample size, sampling interval, and magnitude of shift) on three control charts - Shewhart, CUSUM and geometric moving average charts - using profit as the evaluation criterion, without optimizing any. Ho & Case(Ho & Case, 1994) conducted brief economic comparisons among them and concluded that CUSUM and EWMA charts perform significantly better economically.

2.3 Statistical-based MSPC Methods

Follow the same procedure as the second identified keyword.

TITLE-ABS-KEY (*multivariate* AND *statistical* AND *process* AND *control*)**.**

The Scopus database generates 4409 documents. Search this keyword for other databases and collect the documents in the same folder. Apply inclusive and exclusive criteria and using the type of analysis co-citation and unit of analysis cited-authors in the VOSviewer software, we get all the influential authors in the database as shown in Figure 0-2 Authors in MSPC.

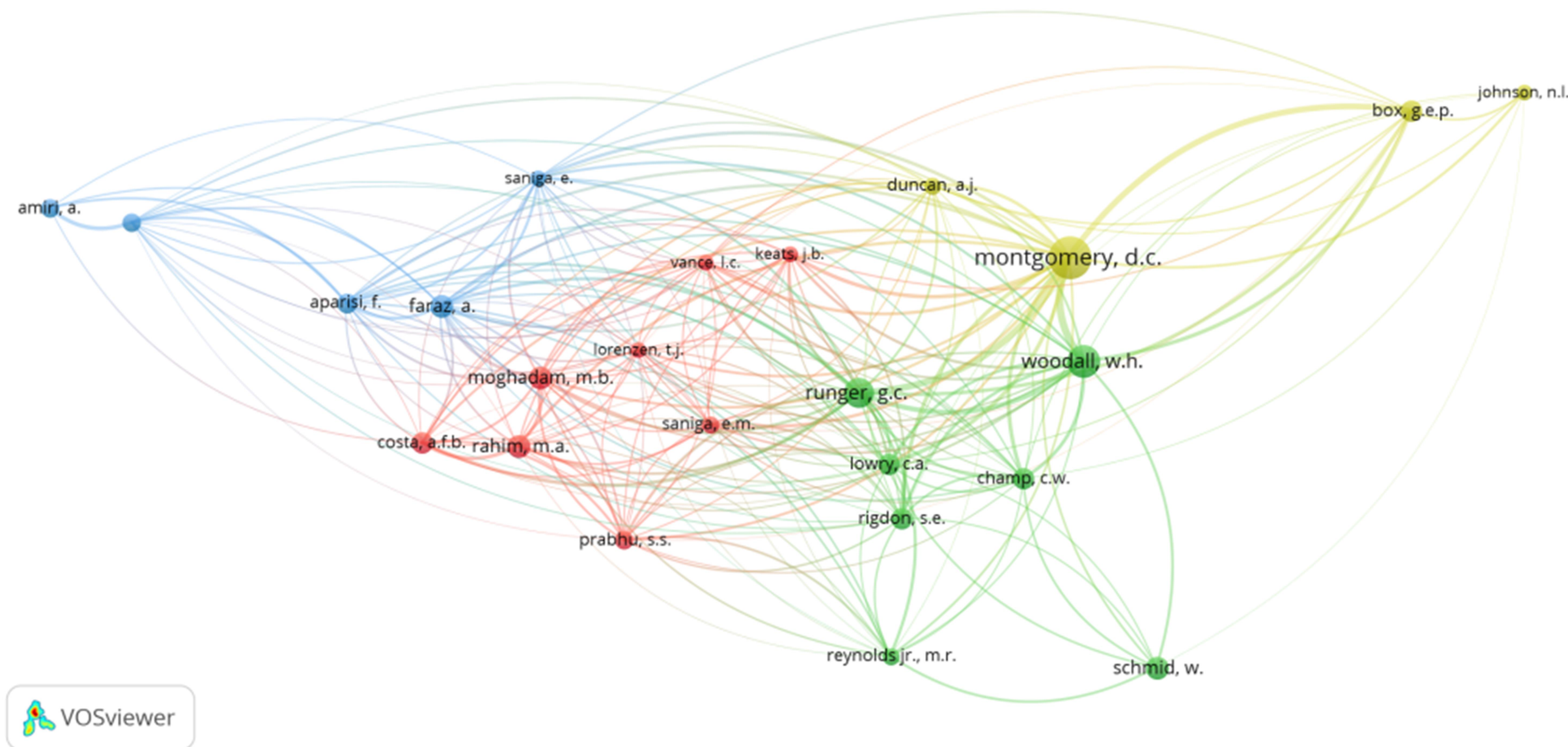

Figure 0-2 Authors in MSPC

Realistically, product quality evaluation requires more than simply one factor alone; multiple aspects contribute to its quality. Therefore, relying solely on univariate control charts alone for multivariate problems can produce inaccurate monitoring results due to its assumption of no correlation among variables and that variables are identically and independently distributed (iid).

Researchers and statisticians have long recognized the limitations of univariate control charts, leading to multivariate methods to address challenges associated with monitoring multiple variables simultaneously. One pioneering development in this area was Hotelling's introduction of T2 charts in 1947 as an extension of Shewhart controls, providing a framework for monitoring multiple variables simultaneously while considering their correlations; furthermore, T2 charts enabled a more comprehensive assessment of product quality by considering all interdependencies among multiple variables simultaneously.

Hotelling's T2 chart provided an early foundation, while in the 1950s additional advancements were made with the development of two additional multivariate control charts: MCUSUM (Multivariate Cumulative Sum) and MEWMA (Multivariate Exponentially Weighted Moving Average) charts. These were extensions of CUSUM and EWMA techniques designed specifically to analyze multivariate data; specifically tailored for multivariate data analysis MCUSUM chart allowed the detection of shifts or changes to the mean vector of multivariate data while the MEWMA chart provided an effective means of monitoring small but gradual shifts within the mean vector.

Multivariate control charts represented an exciting breakthrough in statistical process control, providing a more precise and thorough evaluation of product quality. By considering correlations and interdependencies among multiple variables, these methods provided a stronger framework for monitoring and controlling processes within industries. The next sub-sections provide further discussion for each class of Multivariate techniques.

2.3.1 Multivariate Schewart chart

A multivariate Shewhart chart, using its χ- square value, is the extension of a univariate Shewhart chart based on value to account for two or more variables

(Hotelling, 1951). Also referred to as Hotelling's T2, its name commemorates his introduction of multivariate statistical process control for the first time.

The concept of statistical distance dates back to 1908 when Gosset, then known as a student, examined the ratio of mean to standard deviation of samples taken. His distribution was more rigorously determined in 1925 by R.A. Fisher; Harold Hotelling then generalized this test for significance testing by using fisher-student t-tests that measured means, differences of means, and linear functions of observations within one variate (Hotelling, 1931).

Hotelling T2 examines covariance and correlation among variables using statistical distance, making industrial settings much more challenging when trying to control or interpret individual variables independently of others. Hotelling T2 provides a method of monitoring multiple variables simultaneously while extracting all relevant information from its system.

<u>Hotelling T2 procedure</u>

Hotelling T2 is to be presented as follows. Suppose there are two uncorrelated variables X1 and X2. Then the statistical distance (SD) of these two variables is computed as follows

$$(SD)^2 = \frac{(x_{1-}\mu_1)^2}{\sigma_2{}^2} + \frac{(x_2-\mu_2)^2}{\sigma_2{}^2} \quad \ldots\ldots (1)$$

Where μ is the mean of the population and σ2 is the variance of the population. In analytical geometry, this equation is the equation of an ellipse.

Now in the case of correlated variables, consider the general equation of the ellipse

$$a_{11}x_1{}^2 + a_{12}x_1x_2 + a_{22}x_2{}^2 = c \quad \ldots\ldots (2)$$

Where the a's are specified constants satisfying the relationship $(a_{12}{}^2\text{-}4a_{11}a_{22}) < 0$ and C is a fixed value. By properly choosing the a_{ij} in eq. (2) we can rotate the ellipse while keeping the scatter of the two variables fixed until a proper alignment is obtained. Centered at the mean of the two variables yet rotated to reflect the correlation between them.

For the case of two variables, and assuming that the (x_1,x_2) can be described jointly by a bivariate normal distribution, squared statistical distance (SD) is computed as

$$(SD)^2 = \frac{1}{(1-\rho^2)}[\left(\frac{x1-\mu_1}{\sigma 1}\right)^2 - 2\rho(\frac{x1-\mu_1}{\sigma 1})(\frac{x2-\mu_2}{\sigma 2}) + \left(\frac{x2-\mu_2}{\sigma 2}\right)^2] \ldots\ldots\ldots\ldots\ldots\ldots \text{(3)}$$

Where ρ is the correlation of the two variables. If ρ is positive the ellipse tilted upward and if the ρ is negative the ellipse tilted downward.

Taking the square root of the squared statistical distance gives the actual statistical distance (SD). Equation (3) is the equation of an ellipse. It is also has been labeled as Mahalanobis distance or Hotelling T2 or simple T2.

Now denote a multivariate observation on any number of P variables in vector form as $X'=\{X_1, X_2, X_3, ...X_n\}$.

$$(X-\mu)^{-1}\,\Sigma^{-1}(X-\mu) = (SD)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(4)}$$

Where $X'=(x_1,x_2)$ $\mu'=(\mu_1, \mu_2)$ and Σ^{-1} is the inverse of the matrix Σ

Multivariate analysis' main purpose is to extract valuable information contained within multiple variables. To evaluate their statistical distance (SD), correlation measurements need to be obtained - typically through covariance analysis which quantifies how closely variables vary with one another.

However, one can recognize the limitation of T^2 as a statistical measure: its lack of direct interpretability. As T^2 encompasses all components in an exhaustive way (Hawkins 1991), deciphering its meaning is difficult and takes practice to interpret directly. Regardless, the multivariate analysis still finds T^2 valuable for examining overall relationships and interdependence among variables of interest.

2.3.2 Multivariate Cumulative Sum (MCUSUM)

Woodall & Ncube (Woodall & Matoteng, 1985) proposed the first multivariate CUSUM control chart; then Crosier (Crosier, 1986),(Crosier, 1988) provided two CUSUM vector schemes and two of T(COT) statistics schemes using this technique for CUSUM control charts. Healy(Healy, 1987) then employed this procedure on multivariate normal distributions to detect changes to either mean vector or covariance matrix shift. Pignatiello and Runger(Pignatiello & Runger, 1990) then proposed two additional multivariate CUSUM charts named MC1 and MC2.

Woodall and Ncube(Woodall & Matoteng, 1985) introduced a multivariate CUSUM procedure composed of multiple one-sided and two-sided univariate CUSUM procedures combined into one multivariate procedure. Hawkins (Hawkins, 1991) termed this procedure, known as MCX schemes, which examines successive observations of independent multivariate normal random variables $X_n=(X_{1n},X_{2n}......X_{mn})^T$. Each random vector is assumed to possess a variance-covariance matrix S and its variance-covariance coefficient u_n is computed for target value T=0. After applying one or two-sided CUSUM procedures on sequences of random variables X_{in} (i=1, 2, etc) an interpretable signal can be immediately identified; unlike procedures using T.

Crosier (Crosier, 1988) presented two multivariate CUSUM quality control procedures. The first process reduces each observation to Hotelling's T statistic and then calculates their cumulative sum, known as COT (Cumulative of T). For multivariate CUSUM, direct CUSUM vector formation occurs directly from observations.

In the analysis of COT, firstly the CUSUM is calculated as

$S_n = \max(0, S_{(n-1)} + T_n - k)$, ……………………………………….………… (5)

where $S_0 \geq 0$, $and\ k > 0$. The COT scheme signals when $S_n > h$

A vector-valued CUSUM scheme can be "derived" by replacing the scalar quantities of a univariate CUSUM scheme with vectors. i.e $Sn = \max(0, S_{(n-1)} + (Xn - a) - K)$

Pignatiello and Runger(Pignatiello & Runger, 1990) control charts Multivariate CUSUM#1(MC1) and Multivariate CUSUM#2(MC2) differ by centering around where accumulation (i.e. sum) takes place; Multivariate CUSUM I (MC1) accumulates X vectors before producing quadratic forms while multivariate CUSUM II (MC2) calculates quadratic forms for every observation then accumulates these quadratic forms; with Multivariate CUSUM #1 collecting observations before producing quadratic forms while Multivariate CUSUM #2 calculates and accumulates these quadratic forms before finally adding all observations together into quadratic forms that represent mean vectors before producing quadratic forms that represent mean vectors.

Hawkins(Hawkins, 1991) introduced the multiple regression concepts that Consider the multiple regression of X_i, the i^{th} component of X, on all other components of X. Write this regression as

$(X_i - \mu_i) = \sum_{j\, i} \beta_{ij}(X_j - \mu_j) + \varepsilon_i$ ………………………………………………………. (6)

and write τ_{ii} for the residual variance-the variance of the ε_i. and introduce Z_i where

$Z_i = [(X_i - \mu_i) - \sum_{j\, i} \beta_{ij}(X_j - \mu_j)] / \tau_{ii}^{1/2}$ …………………………………………. (7)

whose null distribution is N(0, 1). This (scalar) Zi is the residual when Xi is regressed on all other components of X rescaled to unit variance. The Z chart found a better performance than the Shewhart X chart. The group charts of Z which are MCZ, is the MCUSUM proposal applied to Z, and ZNO, which is the Euclidean norm of the five Z CUSUM's is also found to be more effective.

2.3.3 Multivariate Exponentially Weighted moving averages

Lowry, Woodall, Champ, and Rigdon(Lowry et al., 1992) introduced the Multivariate Exponentially Weighted Moving Average (MEWMA). Their paper generalized EWMA's initial application as an instrument for monitoring process stability.

In the Multivariate case, a natural extension is to define vectors of EWMA's,

$Z_i = RX_i + (IR)Z_{i-1}$…………………………………………………………..(8)

Where $R=\mathrm{diag}(r_1, r_2, \ldots\ldots r_p)$, $0<r_j\leq 1$

If r1 =r2 = ……. rp = r then the MEWMA vectors can be written as

$Z_i = rX_i + (1r)Z_{i-1}$………………...…………………………………………(9)

The MEWMA chart gives an out-of-control signal as soon as

$$T_i^2 = Z_i' \Sigma_{zi}^{-1} Z_i > L \quad \text{..}(10)$$

Where Σ is the covariance matrix of Zi and calculated as

$$\Sigma_{zi} = \{\frac{r[1-(1-r)^{2i}]}{2-r}\}\Sigma \quad \text{...}(11)$$

L (L>0) is chosen to achieve a specified in-control ARL

If r=1, the MEWMA chart is equivalent to Hotelling's T2 chart.

As with univariate SPC methods, their multivariate counterparts have also been studied utilizing statistical, economical, and statistical-economical design approaches. Many authors have explored the economic design of multivariate T2 control charts; initial work in this area can be found by Montgomery & Klatt (D. C. Montgomery & Klatt, 1972), Heikes (Heikes et al., 1974), and Alt(Alt, 1976).

Lee and Khoo (M. H. Lee & Khoo, 2006) suggest the optimal statistical design of MCUSUM charts for multivariate individual observations using average run length (ARL) and median run length (MRL) measurements, while Prabhu and Runger(Prabhu & Run2.4ger, 1997) proposed statistically designed MEWMA charts, Linderman & Love(Linderman & Love, 2000) developed methods for Economic & economic statistical design of MEWMA control charts, while Molnau et al.(Molnau et al., 2001) proposed statistically constrained economic designs of MEWMA charts.

2.4 Projection-Based Methods

Hoteling T2, MCUSUM, and MEWMA encounter challenges related to dimensionality and collinearity, alternative techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) have been employed as potential solutions. PCA provides one approach by reducing the dimensionality of datasets containing many correlated variables; PCA does this by transforming these into orthogonal latent variables with lower dimensions(Kazmer et al., 2008).

PLS takes into account both process variables and product quality characteristics for which there exists latent structure, taking into account their relationships to maximize covariance while decreasing collinearity. Feerer(Ferrer, 2007) and Braga(Braga et al., 2018) introduced MSPC using PCA; MacGregor & Kourti (MacGregor & Kourti, 1995) and Kresta (Kresta et al., 1991) also demonstrated PLS' effectiveness for multivariate analysis within process control framework.

2.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an effective MSPC technique capable of handling large numbers of highly correlated variables, measurement errors, and missing data sets efficiently. PCA offers an ideal means of reducing variable dimensions while still retaining vital information.

An individual dataset may include many correlated variables that make multivariate analysis challenging, making the identification of patterns and relationships between variables difficult. PCA provides an easy solution by transforming original variables into uncorrelated principal components and then ranking these in order of importance to focus on those contributing the most while eliminating those with lesser contributions.

PCA excels at handling measurement errors and missing data, with its robust nature providing more reliable analysis despite measurement inaccuracies. Furthermore, its capability of imputing missing values based on available information and patterns seen within data enables more complete and accurate analyses to take place.

PCA simplifies complex datasets by reducing their dimensionality of variables, making them easier to interpret and visualize the underlying structure. Dimensionality reduction also provides a means of dealing with overfitting issues as well as computational complexity issues in machine learning and data analysis tasks.

This method is researched well to apply to process control (Mason et al., 1997)(Doganaksoy et al., 1991)(Murphy, 1987)(Maravelakis et al., 2002), however, the author of this research argues that this method has also an inherent limitation of interpretation of the new variables generated from the principal axis.

2.4.2 Partial Least Square (PLS)

Herman Wold introduced the Partial Least Square (PLS) method in the 1960s as an effective approach for developing predictive linear regression models with many highly collinear explanatory variables. This technique can be utilized in both univariate and multivariate analyses(Hawkins, 1991).

PLS simultaneously reduces the dimensionality of both process data (X) and product quality data (Y), by using both types of variation to construct explanatory factors that serve as explanatory variables. MacGregor & Kourti(MacGregor & Kourti, 1995) and Kresta (Kresta et al., 1991) provided MSPC models based on PLS.

The latent variables of the process and quality variables are defined as a linear combination of the original variables. The number of variables selected in the model is based on some criteria like Akaike Information Criterion (AIC), Final Prediction Error Criterion (FPE), Bayesian Information Criterion (BIC), Law of Iterated Logarithms Criterion (LILC), Normalised Residuals Sum of Squares (NRSS), Multiple Correlation Coefficient (R2), Adjusted Multiple Correlation Coefficient, (Ra2), Overall F-test of the Loss Function (OVF) and Mallow's statistic (Cp)(Li et al., 2002).

2.5 Data Mining-Based MSPC

Follow the same procedure as the third identified keyword.

TITLE-ABS-KEY (*data-mining* AND *statistical* AND *process* AND *control*)**.**

The Scopus database generates 876 documents. And the result will be observed in Figure 0-3

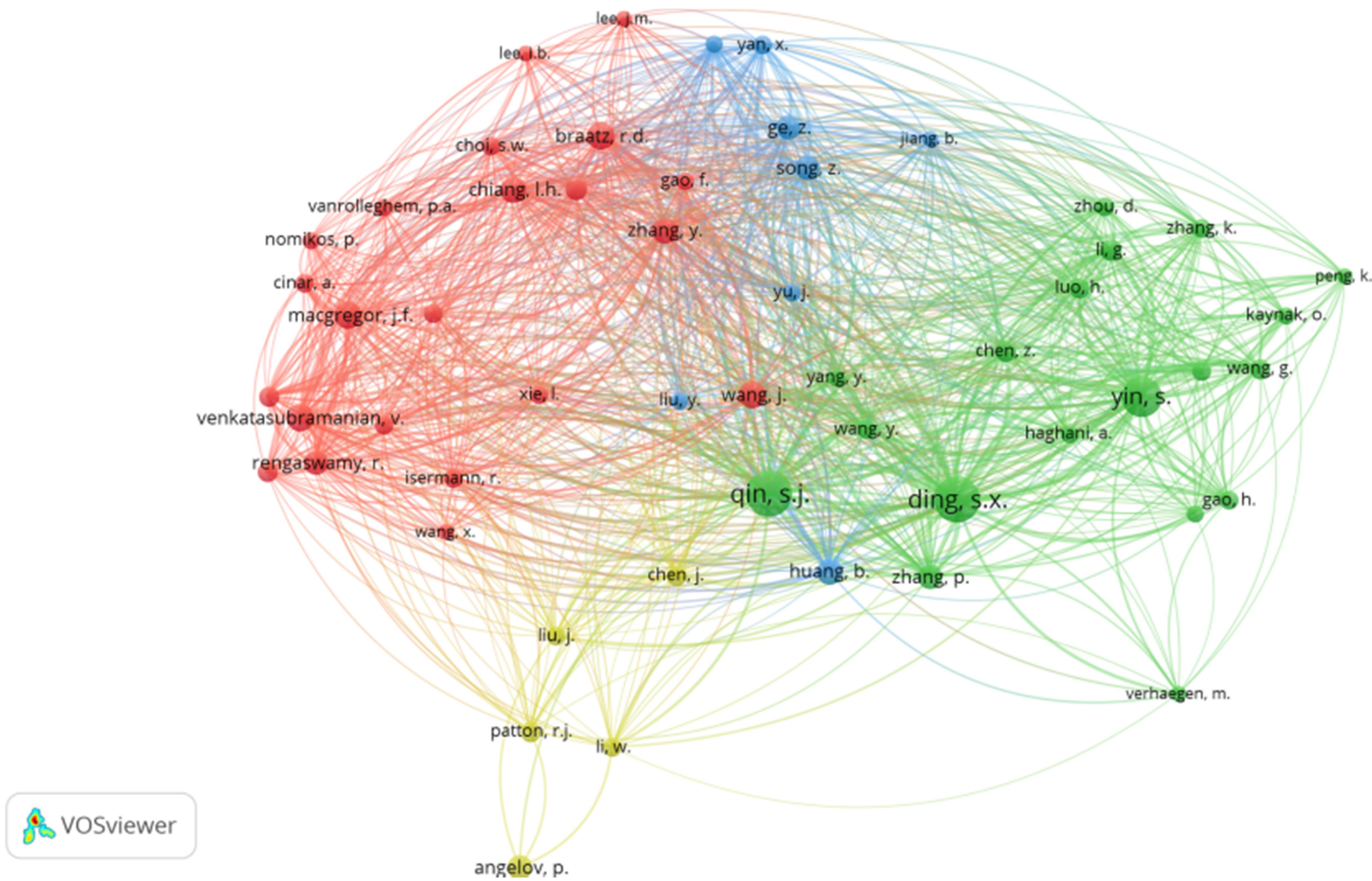

Figure 0-3 Authors in Data Mining

As technology advanced over time, limitations of MSPC techniques were observed. The rigorous mathematical and statistical analysis of the procedure, its time-consuming process during analysis, unsuitable for complex, non-linear, and non-parametric problems found to be not convenient for industrial application. As a result, advanced process control techniques like data mining are introduced. However, application of data mining to manufacturing is limited(Vazan et al., 2017).

Data mining refers to the practice of extracting insights from large volumes of stored computerized information(Olson & Delen, 2008). Data mining is a discipline focused on extracting meaningful and potentially useful information from databases by uncovering hidden patterns or relationships that exist among them (Dogan & Birant, 2021). Technological advancement in data collection and analysis by computers has propelled statistical analysis beyond traditional MSPC methods to data mining stages. Modern databases contain large volumes of data sets which are less susceptible to outlier detection and missing data issues, nonstationary population profiles (making online data analysis suitable with fluctuating populations), quick extraction of patterns quickly and efficiently while dealing with non-numerical information make data mining increasingly preferable(Hand, 1998).

Namikka (Namikka & Gibbon, 2003) divides Data Mining methods into five major categories; classification, deviation detection, clustering, association, and sequential patterns. Some of the traditional data mining techniques studied under these categories include Decision Trees (DT), Neural Networks (NN), KNN (K Nearest Neighbor), Support Vector Methods (SVM), Genetic algorithms fuzzy data mining approach, and rough set methods. Literature shows that almost all data mining techniques outperformed traditional MSPC approaches by some distance (Mason & Young, 2002).

2.5.1 Classification method

Classification involves creating a function that maps data items into one of several predetermined classes by inputting training data sets and developing models of class attributes based on other attributes(Tjortjis & Keane, 2002). Zaki(Zaki & Wong, 2001) define classification as learning a target function that maps each attribute set x to one of several predefined class labels y; it can also be understood as mapping an input attribute set into class labels y.

Classification techniques are effective tools for predicting or describing data sets with binary or nominal categories, yet less suitable when dealing with ordinal ones as they don't account for implicit order among categories(Kesavaraj & Sukumaran, 2013).

Classification techniques (or classifiers) provide a systematic method for creating classification models from an input data set. Classification techniques include decision tree classifiers, rule-based classifiers, neural networks, support vector machines, nearby-neighbor classifiers, naive Bayes classifiers, and ensemble methods.

(Tan et al., 2006) have developed an approach that uses training sets and test sets as separate entities to build classification models that are applied directly to unknown class labels (Test Set) records in the test set, followed by their application by applying this classification model against them for evaluation. Performance measurement involves counting correctly predicted records as opposed to incorrect predictions or calculating the errors for predicting by different by different classification model(Namikka & Gibbon, 2003).

Errors committed by classification models can generally be divided into two groups: training errors and generalization errors. Training errors refer to misclassification errors committed on training records; generalization errors indicate the expected misclassification errors of the model on new records that it has never seen(Tan et al., 2006).

2.5.1.1 Decision Trees

Decision Trees were popularized in statistics by Breiman (Breiman et al., 1984) and machine learning by Quinlan (Quinlan, 1986). Breiman introduced his CART algorithm; Quinlan later introduced ID3 with its extensions C4.5 multivariate tree induction techniques gained popularity later; this review and comparison of many datasets by Yildiz & Alpaydin in 2000 (Alpaydin, 2014) became influential; Murty & Rambabu(Murthy & Rambabu, 1997) provided another overview of the work related to Decision Trees.

Guh and Shiue (Guh & Shiue, 2008) proposed a decision tree (DT) model for quality control purposes that successfully detects mean shifts as well as assignable causes. Their research demonstrated that the learning speed of this DT learning-based model is much faster (25 times) than its neural network-based counterpart, and using recognition accuracy and average run length as performance indices showed the former to outperform its counterpart when applied to control pattern recognition problems.

2.5.1.2 Support Vector Machine

Support Vector Machine (SVM) is one of the more frequently studied data mining techniques first proposed by Vapnik (Vapnik, 1999). Later Tax & Duin (Tax & Duin, 2004) introduced Support Vector Data Description (SVDD), inspired by support vector classification. Scholkopf et al. (Schölkopf et al., 2001) further refined this idea. Scholkopf introduced one-class support vector machines (OCSVM). Sun and Tsung's (Sun & Tsung, 2003) K-Chart is a support vector data description (SVDD)-based control chart used for multivariate SPC. Gani et al.(Gani et al., 2011) demonstrated the industrial application of K-chart as being highly sensitive to small shifts in mean vector, outstripping T2 control chart in terms of ARL. Ning and Tsung(Ning & Tsung, 2013) provided an optimized design method for an SVDD-based K-chart design method.

Cai and colleagues(Cai, 2006) proposed support tensor machine models, while Chen (Chen et al., 2016) proposed one-class support tensor machines as extensions to one-class support vector machines (OCSVM), to take into account tensor data input. Sokchotrat (Sukchotrat et al., 2010) developed one class classification-based control charts using the Mahalanobis kernel for multivariate process monitoring; Maboudou-Tchao suggested several forms including support matrix data description(SMDD), support tensor vector data description, and support tensor vector data description(Maboudou-Tchao, 2019).

He and Zhang(S. He & Zhang, 2011) proposed the Support Vector Data Distribution (SVDD)-based MCUSUM chart known as S-MCUSUM for data distribution with support vector data distribution (SVDD). Their research compared this technique with the Crosier COT chart which can analyze free distribution. In their simulation model, they have demonstrated that S-MCUSUM outperformed COT in terms of banana-shaped distribution typical of free distribution.

Xia et al. (Xia et al., 2018) developed the D-MCUSUM control chart which is constructed using modified Multivariate CUSUM charts based on support vector data description algorithms. The advantage is sensitivity to small shifts due to its Multivariate CUSUM algorithms as well as learning abilities from support vector data description algorithms.

2.5.1.3 Neural Networks NN

An early foundation for Neural Networks can be found in the late 19th and early 20th centuries, when scientists such as Hermann von Helmholtz, Ernest Mach, and Ivan Pavlov conducted inter-dependability work across physics, psychology, neurophysiology, and vision studies. Early NN work focused on general theories of learning and vision conditioning rather than specific mathematical models for artificial neuron operation.

Warren McCulloch and Walter Pitts(Mcculloch & Pitts, 1943) laid the groundwork for modern neural networks during World War II with their work showing that artificial neurons could, in principle, compute any mathematical or logical function imaginable using networks of artificial neurons; their work is generally recognized as the beginning of the modern neural network field. Later Donald Hebb proposed that classical conditioning as discovered by Pavlov is present due to individual neuron properties; their work led to further advancement of our knowledge in this area.

Frank Rosenblatt first proposed practical applications of Artificial Neural Networks (ANNs) in 1957 with his creation of perceptron networks and their respective learning rules(Frank Rosenblatt, 1960). At approximately the same time, Bernard Widrow and Ted Hoff introduced "MADALINE" and used its Least Mean Square algorithm (LMS) to train adaptive linear neural networks which had similar structures and capabilities as Rosenblatt's perceptron. Unfortunately, both Rosenblatt's and Widrow's networks can only solve linearly separable problems; this was widely noted by Marvin Minsky and Seymour Papert; nonetheless, some significant work continued into the 1970s. In 1972, Teuvo Kohonen and James Anderson independently created new neural networks which could act as memories. Stephen Grossberg was also actively researching self-organizing networks at this time; among his research were James Anderson's brain-state-in-a-box (BSB) model and Kunihiko Fukushima's neocognitron model of self-organizing networks.

Two concepts introduced during the mid-1980s contributed significantly to NNs' revival: statistical mechanics was pioneered by John Hopfield(Hopfield, 1982), explaining its operation by specific classes of recurrent networks that could act as an associative memory; another key development from that decade was backpropagation for multilayer perceptron training which was independently discovered by Paul Werbos(Werbos, 1988), David Parker(Parrker,1985) and David Rumelhart(Rumelhart et al., 1986) as backpropagation algorithms; James McClelland found these more helpful in answering criticism made against them by Minsky and Papert during their debates from 1960s.

The single neuron is represented by Figure 2-4.

P is the input. The weight, w, corresponds to the strength of synapses, the cell body is represented by the summation (Σ) and the transfer function (f) and the neuron output (a) represents the signal on the axon.

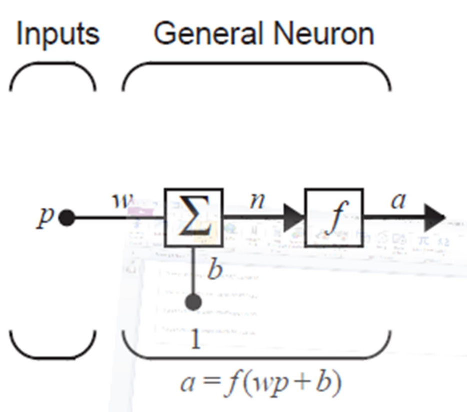

Figure 0-4 Representation of single neuron

From the collection of neurons arises an Artificial Neural Network (ANN). These highly parallel computational systems are comprised of interconnected artificial neurons or processing units(Perry et al., 2001).

Based on their connection patterns, artificial neural networks can generally be divided into Feed forward network architecture and Recurrent (backward) network architecture, as depicted in Figure 0-5. Single layer perceptron, multilayer perceptron, and radial basis function belong to feed-forward neural network while competitive networks, Kohnen's Self Organizing Map (SOM), Hopfild network, and adaptive resonance theory (ART) fall under Recurrent (backward) neural network architecture.

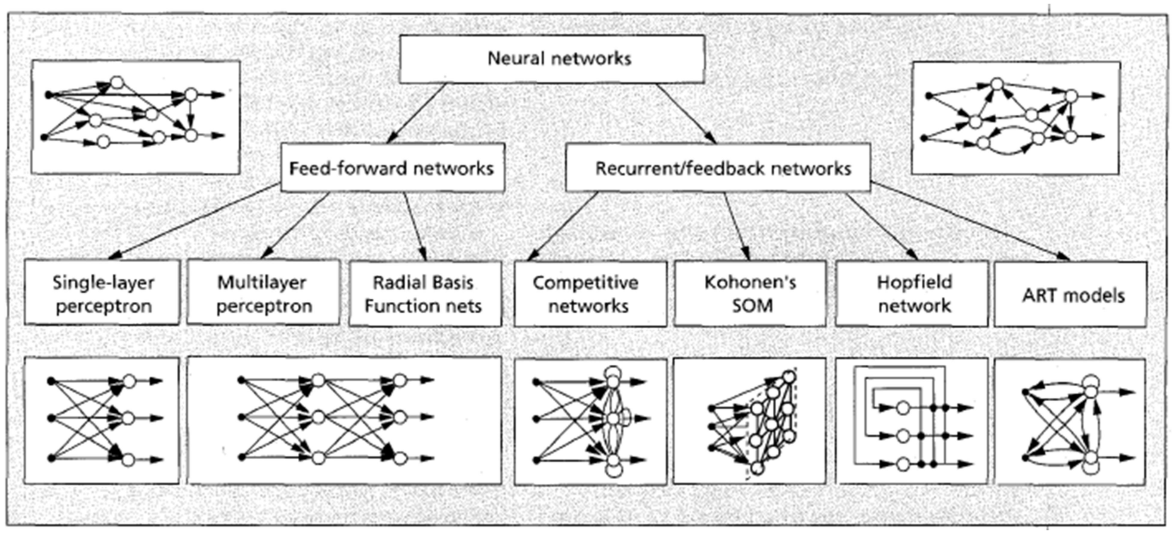

Figure 0-5 Types of ANN Architectures

Perceptrons consist of one input layer of McCulloch-Pitts neurons feeding forward into one layer of McCulloch-Pitts neurons in an arrangement called the Perceptron (Frank Rosenblatt, 1960). Widrow and Hoff developed the gradient descent algorithm (delta rule). Single-layer perceptron can handle linearly separable sets by adjusting connection weights and bias.

Multilayered Neural Networks consisting of one or more hidden layers are known as Multilayered Neural Networks. Multilayer perceptron (MLP), trained via a back-propagation algorithm, can overcome similar input patterns leading to similar output patterns in SLP networks. Both MLP and RBF networks can approximate any continuous function with some degree of accuracy.

Broomhead & Low introduced the radial basis function Neural Network in 1988 and developed it further since(Broomhead & Lowe, 1988). Their proposed input-output multivariate relationship is given by. For optimal operation of such networks, adjustments must be made to both their centroids (represented by Ci) and scaling factors sj.

If the processing units in a group of instars are connected in an on-center off-surrounding feedback network, their network is known as a competitive layer, **Figure 0-6**

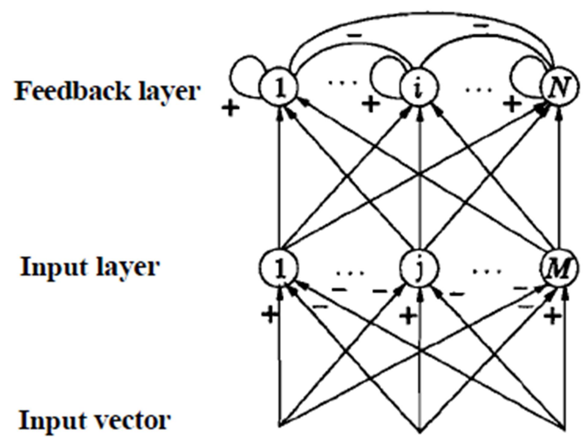

Figure 0-6 Competitive learning

SOM is an unsupervised learning algorithm in which target outputs are not known or given. A Kohonen unit computes the Euclidian distance b/n an input X and its weight vector w. A Kohonen layer(Kohonen, 1982) acts as a "winner-take-all"(WTA) layer where only one output of its layers is reported as "1". Thus for any given input vector only one kohenon layer yields 1 while all others produce 0(Yin, 2008).

Hopfield(Hopfield, 1982) proposed an associative model, one of the cornerstones of which has contributed significantly to NNs' resurgence today. Hopfield Neural Network (HNN) is a single-layered and recurrent network, in which each neuron is fully connected. Given two neurons i and j, there exists a connectivity weight wij b/n them which is symmetrical (wji=wji). Conversely, wii =0 signifies no connectivity whatsoever between them.

Grossberg created the ART model in the 1970s as an autonomic competitive neural network, composed of three components (Comparision field and recognition field composed of neurons, Vigilance parameters, and Reset modules)(Grossberg, 2013).

Neural Networks require some computational method of weight adjustment. There exist various learning rules or methodologies for accomplishing this task(Perry et al., 2001) these include 1. Backpropagation (BPN), 2. Simulated Annealing (SA), 3.

Learning Vector Quantization(LVQ), 4, Probabilistic Neural Network(PNN), 5. Cascade Correlation Learning. Huang and colleagues developed the Extreme Learning Machine (ELM).

2.5.2 Clustering method

Cluster analysis has long been considered one of the oldest data mining techniques and its literature since it first came into use in 1955 is quite extensive. Jain(Jain, 2010) defined Clustering analysis as: "the formal study of methods and algorithms of grouping or clustering objects according to perceived intrinsic characteristics or similarity". Tan (Tan et al., 2006) described Clustering analysis as: "dividing data into groups (clusters) which are either meaningful, useful, or both; meaningful groups should capture the natural structure of data while useful groups serve other purposes such as summarization".

Clustering analysis is an unsupervised classification method that groups data objects based on information in their respective datasets that describe them and their relationships. Clusters can be identified by shapes, sizes, and densities that define the clustering pattern; an ideal number of clusters must also be determined for successful clustering analysis to occur; within any one cluster should be similar in characteristics compared with objects from another group represented by dense patterns; the greater this homogeneity within one group and greater differences between groups, the more precise or distinct will be the clustering results;

Tan(Tan et al., 2006) conducted a systematic analysis of clustering techniques, categorizing them into one of three types. He divided these categories as hierarchical (nested) vs partitional (unnested), exclusive versus overlapping versus fuzzy, and complete versus partial. Jain (Jain, 2010) explored hierarchical (nested) versus partitional (unnested) clustering. If we allow clusters to contain sub-clusters, this results in hierarchical clustering: an organized set of nested clusters organized as a tree structure. Partitional (Nested) Clustering finds all clusters at once, as a partition of data, without creating hierarchies within each cluster. Agglomerative Hierarchical Clustering is a clustering approach that encompasses several related techniques to produce hierarchical clustering by starting each point as its singleton cluster and merging any two closest ones until one comprehensive cluster remains.

Exclusive, Overlapping, and Fuzzy Clustering techniques differ according to how objects are assigned into classes. When each object belongs to only one cluster it is classified as exclusive; when an object appears multiple times within one cluster this is classified as overlapped (i.e. one object can belong simultaneously to more than one class). Fuzzy C-means(FCM) was first introduced by Jim Bezdek in 1981 as one method that can apply these classification techniques effectively.

Complete clustering differs from partial clustering in that all objects must be assigned a cluster, while in partial clustering not necessarily all the objects can be classified; an example would be outliers that do not belong to any particular cluster.

Clustering techniques can be divided into five distinct groups: well-separated, prototyped, graph-based, density-based, and shared property (conceptual cluster). Clusters that are well separated tend to exhibit larger distances between any two points from different groups than between any two points within one group(J. He et al., 2015). Clusters identified using prototyped-based clustering are groups of objects where any object is closer to its prototype than any other prototype of any other cluster. K-means is a prototype-based partitional clustering technique that attempts to find K clusters that represent centroids for many types of data, or prototype-based clusters as we often refer to them.

Toussaint(Toussaint, 1980) first introduced the Relative Neighborhood Graph (RVG), which uses graph theory for cluster analysis, back in 1980. A graph representing data with nodes representing objects and links representing connections among them can then be represented as an RVG; its edges being defined using distance(Jaromczyk & Toussaint, 1992). If data were presented as nodes representing objects and the links representing connections among them; clusters can then be defined as connected components i.e. groups that share some connections but do not connect outside this group - with Gabriel and Delaunay graph extensions taking it further and becoming known as proximity graph

Density-based clustering relies on the number of points per hyper-volume unit which are higher within a cluster than outside it (P. Singh & Meshram, 2018). If an object clusters together within an environment of lower density surrounding it then they form density-based clusters. DBSCAN clustering is one such density-based algorithm that produces partitional clustering in which the number of clusters is automatically determined by its algorithm.

Conceptual Clustering (Conceptual Clumping, or Concept Clusters, for short) is an informal method of learning by observation or concept formation from an existing set of concepts (Michalski, 1980). When objects are clustered based on shared properties (conceptual clusters), this form of clustering provides both an intentional description for each cluster as well as its intentional set (Fisher & Langley, 1980).Perez-Suarez (Pérez-Suárez et al., 2019) reviews literature related to conceptual Clumping).

2.5.3 Association method

Association Rule Mining (ARM) is one of the most well-researched data mining techniques. First introduced by Agrawal(Agrawal et al., 1993), who proposed an algorithm to generate associations between items purchased frequently by customers in large customer transaction databases and the customers themselves. Retainers use association rule mining techniques to investigate customers' purchasing habits; these techniques extract correlations between variables, frequent patterns, associations, or informal structures between multiple items or databases - this technique being widely employed across industries like medical diagnosis, marketing and sales, finance, text mining networks inventory control, and counseling among others.

ARM offers several algorithms such as Apriori, FP-Growth, and ECLAT algorithms. Agrawal(Agrawal et al., 1993) first proposed Apriori using its classic implementation based on the Apriori principle; which states that if an item set is frequent then all its subsets must also be frequent. IF/THEN statements help uncover relationship rules among related data stored in a repository. Three criteria such as support, confidence, and lift are utilized in association rule mining to identify relationships and strengths used when mining association rules. Support (S) of an association rule can be defined as the percentage of records that match both conditions, about total database records containing these elements. Confidence (C) of an association rule is defined as the ratio of transactions that contain both elements X and Y to all records containing X; it serves as an indicator of its strength. Lift(L) can be defined as total support divided by independent support i.e. The traditional Apriori algorithm suffers from high computational cost and time consumption due to repeated scanning of databases;

Singla & Malik(Singla & Malik, 2014) reviewed eight approaches that enhance it further.

The Frequent Pattern Tree-Growth Algorithm was proposed as an alternative to the Apriori algorithm's limitation by compressing the database into a compact data structure using the divide-and-conquer strategy(Gupta, 2011). Once this FP-tree is built it may become too large; so optimization techniques such as COFI-tree Mining, CT-PRO, and FPgrowth Algorithms are introduced as solutions.

ECLAT algorithm is an enhancement method that uses vertical databases instead of horizontal ones for its algorithmic scan of databases.

2.5.4 Deviation detection

Aggrawal(Aggarwal & Yu, 2001) asserts that outliers may be defined as points outside a set of defined clusters that lie outside their set; or alternatively as outliers that fall outside this set but remain separate from noise.

Outliers occur due to human error, mechanical faults, instrument error, natural population deviations, fraudulent behavior, or changes in system behavior(Hodge & Austin, 2004).

2.5.5 Sequential patterns

Aggrawal and Srikant (Agrawal & Srikant, 1995) first introduced the sequential pattern mining problem by exploring potential patterns within sequence databases. Their methodology relies heavily on customer purchase sequences; as follows: "Given a set of sequences where each element consists of several items, and given a user-specified minimum support threshold, sequential pattern mining involves searching for all frequent subsequences within this set whose frequency of appearance in said set exceeds min support(Han et al., 2004)".

Recent studies have identified two general classes of sequential pattern mining methods. 1) Generation-and-test approaches such as GSP: Horizontal format-based sequential pattern mining methods and SPADE: vertical format-based sequential pattern mining methods 2) Sequence pattern growth techniques like prefixspan (horizontal format-based sequential pattern mining methods) with its further expansion such as Clospan for mining closed sequential patterns are among them.

2.6 Machine Learning

The fourth keyword in the literature survey is machine learning. From the preliminary literature review, it is identified that machine learning techniques can be categorized into seven groups. These are supervised learning, semi-supervised learning, unsupervised learning, reinforcement learning, ensemble learning, instance-based learning, and multitask learning. For each technique that belongs to a single group, documents are identified following the same procedure as indicated above.

For example, support-vector-machine is one of the supervised machine learning techniques. The following keywords are used to search the articles.

TITLE-ABS-KEY (*support* AND *vector* AND *machines* AND *pattern* AND *recognition*). The result will be shown in Figure 0-7.

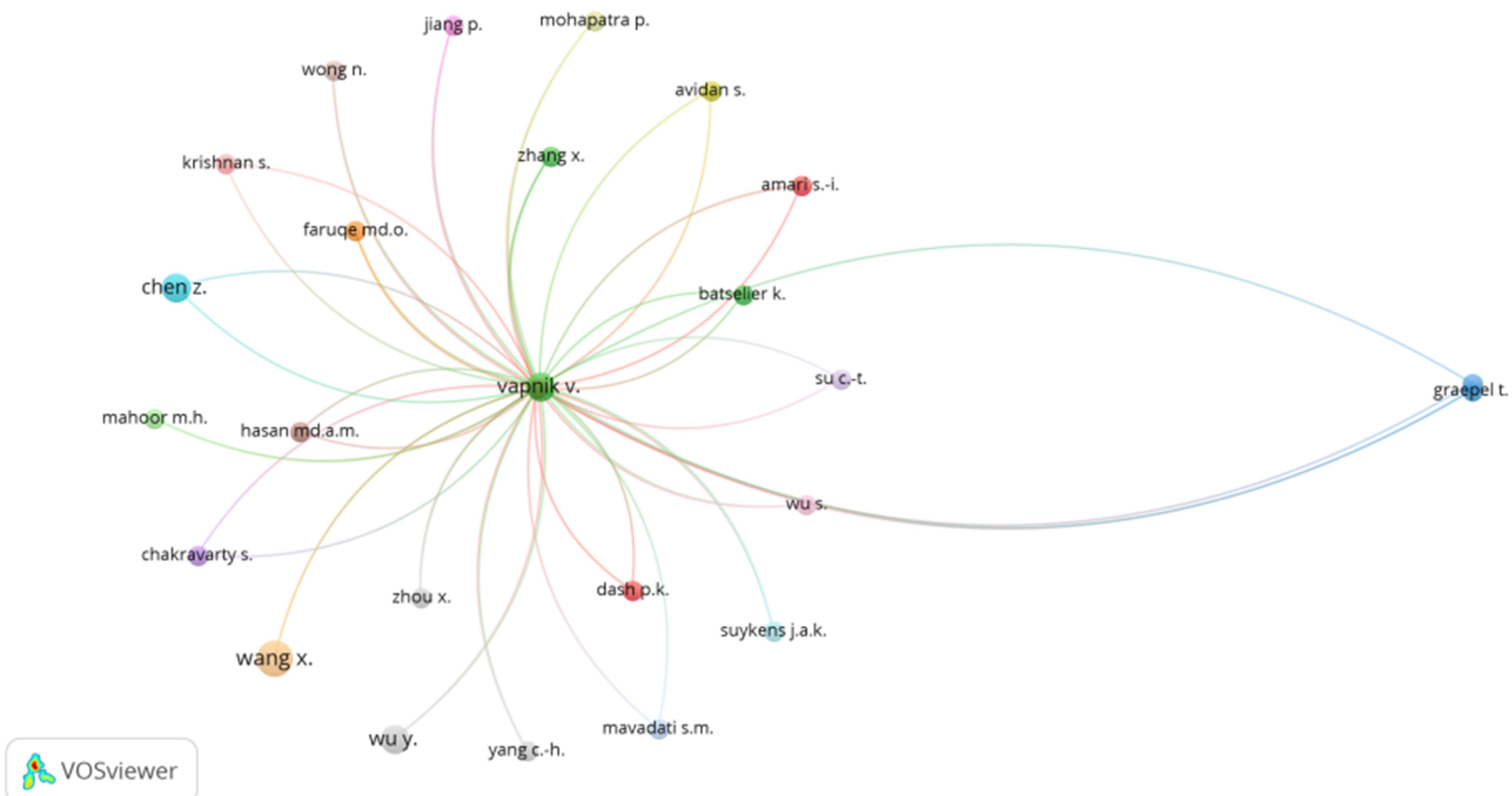

Figure 0-7 Support Vector Machines writers

Arthur Lee Samuels, an IBM researcher, created one of the earliest Machine Learning programs-a set of learning programs for playing checkers and popularized the term. Today, Machine learning serves to teach machines how to handle data more efficiently (Dey, 2016). Machine Learning algorithms can be divided into eight distinct groups; these include supervised learning, unsupervised learning, semi-supervised learning reinforcement learning multitasking learning Ensemble Neural Networks instance-based learning, represented by Figure 0-8.

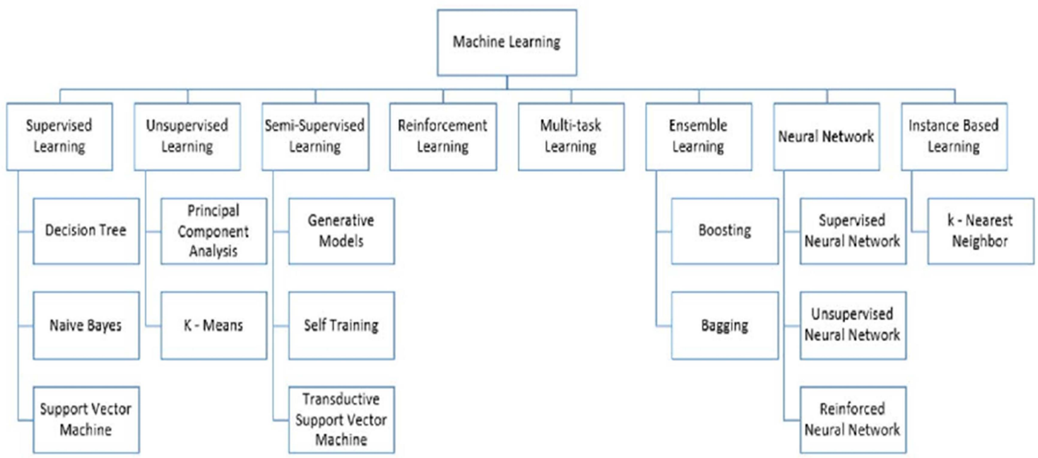

Figure 0-8 Classification of Machine Learning algorithms

2.6.1 Supervised learning

If instances are given with known labels (i.e. the corresponding correct outputs) the learning is called supervised whereas if the instances are unlabeled it is unsupervised. Decision Tree, Naïve Bayes, and Support Vector Machine are the three types of Supervised Learning algorithms.

Decision Trees

Decision trees are used to classify instances based on feature values(Kostiantis, 2007). A decision tree consists of nodes, edges, and leaves: these nodes represent tests for attributes while edges connect them to nodes or leaves that predict an outcome. Finally, leaves serve as terminal nodes that help predict this result.

Decision trees can be translated into rules by creating one rule for every path from root to leaf in a decision tree. Alternatively, rules may also be directly induced from training data using various rule-based algorithms; Furnkranz (Furnkranz, 1999) provided a great overview of existing work in rule-based methods.

naïve Bayes

The Naïve Bayes machine learning algorithm is based on the theorem of probability proposed by Thomas Bayes in the 1740s. Bayes theorem states that

$$p(x/y) = \frac{p(y/x)p(x)}{p(y)}$$

Machine learning's Bayes theorem can be used for classification. Let h represent the attribute set and D represent the class variable; if there is no non-deterministic relationship between D and its attributes, we can treat h and D as random variables and capture their relationship probabilistically using p(D/h). This conditional probability is known as its posterior probability for D; otherwise known as its prior probability p(D).

At this stage of training, it is necessary to learn the posterior probabilities for every combination of D and H using information gleaned from our training data. With these probabilities in hand, a test record h' can be classified by finding which class D maximizes its posterior probability p(D/h').

Support Vector Machine

Support Vector Machine (SVM), first proposed by Vapnik in 1966, has quickly become one of the most sought-after techniques due to two benefits. SVM provide satisfaction from a theoretical viewpoint while producing impressive performances in real-life applications. SVM was recently introduced as an additional technique for solving numerous learning, classification, and prediction problems(Cheng & Cheng, 2007).

Support Vector Machine's objective is to locate a mapping function or the maximum separating hyperplane in geometric terms from training data of the form (x1,y1).... (**xn**, yn) Є Rn X {1,-1}; this corresponds to n-dimensional vector patterns(vectors)xi with their labels yi as illustrated below; their following equations depict them.

$$x_i w + b \geq +1 \qquad for\ yi = +1 \ldots\ldots (20)$$

$$x_i w + b \leq -1 \qquad for\ yi = -1 \ldots\ldots (21)$$

In this case, the maximum separating margin of d which is equal to 2/||w|| is a constraint optimization problem represented by minimizing ½||w||2 subject to yi(wxi+b)≥1.

When classes cannot be separated by a linear hyperplane, the kernel trick may be applied. A kernel function serves to map nonlinearly separable data into linearly separable high-dimensional feature space; various kernel functions have been identified specifically for this task: polynomial kernel functions are identified here such as Gaussian radial basis function, exponential radial basis function, multilayer perceptron, etc.

2.6.2 Unsupervised Learning

Unsupervised learning refers to a type of training data wherein only input vectors (X) exist, with no associated output observations available for training purposes. From this row dataset, useful information is extracted. Principal component analysis and clustering are among the most frequently employed unsupervised learning techniques - they will both be covered here in detail.

Principal Component Analysis

K. Pearson introduced Principal Component Analysis in his paper entitled: On Lines and Planes of Closest Fit to Systems of Points in Space"(Pearson, 1901). Hotelling (Hotelling, 1933) developed further this science in psychometry while Jolliffe(Jolliffe, 2002) acknowledged him as contributing to its advancement and dissemination to various disciplines.

Principal Component Analysis (PCA) is a technique by which a set of variables, X1,X2,......Xp is transformed into new variables Y1,Y2...Yp by applying an equation called E=EX; where E represents columns of Eigenvectors of Variance-Covariance Structure or Correlation Matrix of X. Principal components are new variables Y1, Y2,...Yp represents directions with maximum variability while remaining uncorrelated between themselves. Principal components are linear combinations of variables with maximum variance. As more variance is represented by fewer variables, principal components provide advantages in more efficiently representing systems.

PCA is one of the many multivariate quality control and monitoring methods. It can reduce the dimensionality of monitoring spaces while the process itself can then be monitored using one, two, or three-dimensional control charts that maintain all the simplicity and clarity associated with conventional single-variable SPC charts(Kourti & MacGregor, 1996).

k-means clustering

K-means clustering is an unsupervised learning technique in which an algorithm iteratively assigns each data point to one of the k-groups using the features provided. More broadly, cluster analysis encompasses algorithms and methods for grouping or classifying objects(Jain & C.Dubes, 1988).

Let X={x_i}, i=1,…..,n be the set of n-dimensional points to be clustered into a set of k-clusters, C={ck, k=1,….k}. The K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μk be the mean of cluster Ck. The squared error between μk and the points in cluster CK is defined as

$J(Ck) = \sum_{xi\epsilon Ck} \|xi - \mu k\|^2$……………………..……………….……………(22)

The goal of k-means is to minimize the sum of the squared error over all k clusters

$J(C) = \sum_{k=1}^{k} \sum_{xi\epsilon Ck} \|xi - \mu k\|^2$…………………………………………………..(23)

2.6.3 Semi-supervised learning

Semi-supervised learning algorithms are a technique that combines the power of both supervised and unsupervised learning. It contains both labeled and unlabeled data where the labeled data are small in number and unlabeled data are by far large in number.

Generative models

The generative Model originated to address the problem of optimal use of a training set Q composed of l labeled samples $\{(X_1,\theta_1)(X_2, \theta_2),\ldots\ldots(Xl, \theta l)\}$ and u unlabeled samples $\{X1,\ldots\ldots Xu\}$ in the construction of a classifier that discriminates between two classes of observation. (Castelli & Cover, 1996) It has a probabilistic approach to classification.

The generative Model considers the class conditional densities p(X/Ck) as well as the class prior is p(Ck) and then uses these to compute posterior probabilities p(Ck/X) through Bayes' theorem. Once it is specified a parametric functional form for the class-conditional densities p(x/Ck), the values of the parameters are determined, together with the prior class probabilities p(Ck) using the maximum likelihood function.

self-training

Self-training involves first training a classifier with small amounts of labeled data, before using this classifier to classify unlabeled points with predicted labels that were added later to the training set and added back into it for classification(Zhou & Li, 2005). Yarowsky(Yarowsky, 1995) makes use of self-training for word sense disambiguation purposes.

co-training

Co-training operates under the assumption that features can be divided into two sets, each sub-feature set being sufficient to train an effective classifier. Given any given class, both sets should be conditionally independent of one another. Two separate classifiers are first trained on their respective sub-feature sets separately using labeled data from both sub-feature sets respectively before classifying unlabelled examples (predicted labels by themself) to teach the other classifier more confidently (Blum & Mitchell, 1998) / Mitchell 1999/(Zhou & Li, 2005)). Once complete, this process repeats itself.

Transductive support vector machine

A Transductive Support Vector Machine (TSVM) is an extension of support vector machines (SVM). TSVMs use both labeled and unlabeled data, with their objective being to assign class labels such that the best support vector machine(Bennett & Demiriz, 1999). Furthermore, regularization techniques enable this model to maximize separation between labeled and unlabelled data through regularization (Wang 2000).

Graph-based

The Graph-based method is the recent introduction of the semi-supervised learning method. Graph-based Learning algorithm (GBL) defines two sets in semi-supervised learning. The first set DL={(Xi,Yi)}i=1l contains labeled data, and the second set Du={Xi}i=l+1n with n=l+u contains the unlabeled data. Each data point lies in a d-dimensional vector space Rd. The goal of the GBL algorithm is to infer the label of Du Via an undirected weighted graph G= (V, E, W) where V is data points in Dl and Du and E is the undirected edges on the graph, weighted by wij Є W. (Zhu & Lafferty, 2005) (Zha et al., 2009) Nodes in the graph corresponds to labeled and unlabeled data points and edges reflect the similarities between data points.

2.6.4 Reinforcement Learning

The term "reinforcement" and "reinforcement learning" were used in the engineering literature since 1960(Minsky, 1961). Widrow devised a reinforcement learning algorithm called "punish/reward" or "Bootstrapping " in the mid-1960s. A related reinforcement learning approach was later explored in a classic paper by Barto, Sutton, and Anderson on the "Credit Assignment "Problem.

Reinforcement learning is characterized by a long-term interaction between the learning agent and a dynamic, uncertain environment. Reinforcement learning has the following elements: Agent, Environment, Policy, Reward signal, Value function, and optionally a Model of the environment(Sutton & Barto, 2015). The agent senses the state of the environment and chooses an action to perform in the environment to meet the maximum reward of the environment as a result the state of the environment will be changed. This action is determined by the policy of the agent. The agent sense again the state of the environment and the next action to the best reward of the environment will proceed. This iteration will be done continuously and the value function is defined from the cumulative long-run rewards. Diagrammatically, it is represented by Figure 0-9.

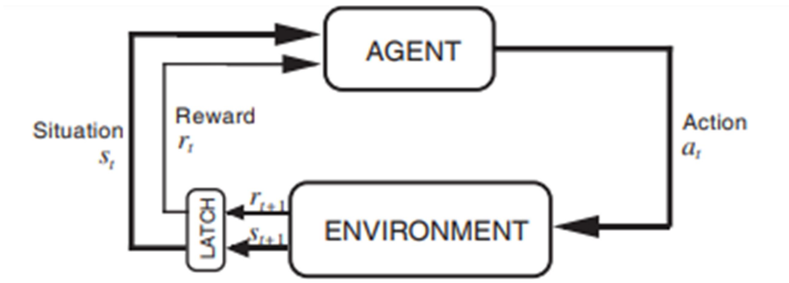

Figure 0-9 Reinforcement Learning Model

Reinforcement learning is the branch of machine learning in sequential decision-making settings. Mathematically is it modeled by the Markov decision process (MDP). Reinforcement learning (RL) methods such as Q-Learning are closely related to a long line of research on the Dynamic Programming approach to solving Markov Decision Processes.

2.6.5 Multitask learning

Sharing what is learned by different tasks while tasks are trained in parallel is the central idea in MultiTask Learning (MTL)(Caruana, 1997). MTL trains many tasks in parallel on one learner. These tasks or a subset of tasks are related to each other and it helps to improve the learning performance of a model. A good example of an MTL net is a shared hidden layer trained in parallel on all the tasks, implying that learning for one task can help other tasks share the learning experience as a result of multi-hidden layers.

Multitask learning is defined as follows: Given m learning tasks where all the tasks or a subset of them are related, multitask learning aims to help improve the learning of a model for Ti by using the knowledge contained in all or some of the m tasks(Zhang & Yang, 2018).

MTL can be modeled by neural or non-neural methods. Neural models are represented by deep neural networks whereas the non-neural models are represented by linear models, kernel methods (KNN), and Bayesian algorithms.

2.6.6 Ensemble learning

Ensemble learning is to mean the field of combining classifiers(Kuncheva, 2004).Objects have some specific characteristics or what are called features (attributes). For n featured spaces and C classes, the classifiers are defined as any function D that maps Rn→C. Decisions of different classifiers are combined, and the class that receives the majority of 'votes' (i.e., the class predicted by the majority of the learning machines) is the class predicted by the overall ensemble(Re & Valentini, 2012). Ensemble algorithms can be categorized into bagging and boosting algorithms.

Bagging

Bootstrap AGGregation, or "bagging" was introduced by Breiman (Breiman, 1996) to improve the performance of a predictor through bootstrap resampling from the empirical distribution. In the Bagging algorithm, each member of the ensemble is constructed from a different training dataset, and the predictions are combined either by uniform averaging or voting over class labels.

Boosting

Boosting refers to a family of algorithms that combines multiple weak learners sequentially into a single strong learner, each trying to correct its predecessors. It was the idea of Kearns(Kearns, 1988),Valiant (Kearns & Valiant, 1989) Robert Schapire (Schapire, 1990) affirms the answer.

Boosting algorithms are categorized into the following three groups. These are Adaboost (Adaptive boosting), Gradient boosting, and XGbosting (Extreme gradient boosting). In the case of Adaboost, in its sequence of iteration, the weights of every incorrectly classified observation are changed. The gradient Boosting method tries to fit the new predictor to the residual errors made by the previous predictor. XGBoosting is an implementation of gradient-boosted decision trees designed for speed and performance.

2.6.7 Neural network

A neural network is the well-developed and most applicable method of artificial intelligence. Researchers classify the study of NN into the following three broad categories: Supervised NN, Unsupervised NN, and reinforced NN.

Supervised neural network

In a supervised neural network, the actual outputs of the input data are already known (Dey, 2016). A familiar example of supervised neural networks is feedforward networks with backpropagation of errors. The multilayer perceptron is the most common feed-forward network, a well-known and widely used model of supervised neural networks.

Unsupervised neural network

The self-organizing map (SOM) introduced by Kohnon is a very popular ANN algorithm based on unsupervised learning(Kohonen, 1982)(Kohonen, 2001)(Kohonen, 2013). Kohnon described the method as every input data item shall select the model that matches best with the input item, and this model, as well as a subset of its spatial neighbors in the two-dimensional grid, shall be modified for better matching.

More similar models will be associated with nodes that are closer in the grid, whereas less similar models will be situated gradually far away in the grid. SOM applies the concept of Vector quantization concept introduced in scaler form by(Lloyd, 1982) and in vector form by Forgy(1965).

Adaptive Resonance Theory (ART) developed by Grosberg (Grossberg, 1987) is another competitive learning model that is categorized into the unsupervised neural network. ART has various architectures like ART1, ART2, ART3, and fuzzy ART. ART1 is an architecture that can be used for the clustering of binary inputs only. ART2 improved upon the ART1 architecture to support continuous input. Fuzzy ART incorporates fuzzy set theory into the pattern recognition process.

Reinforced neural network

Several types of research have been done that integrate reinforcement learning and Neural Networks. Miljković (Miljković et al., 2013)(2013) used Q- learning, and SARSA coupled with neural networks for visual control of a robot manipulator. Also, Ghanbari (Ghanbari et al., 2014) survey the literatures that considers the reinforcement learning in neural network.

Recently, research on the application of NN to multivariate process control is conducted by Bersimis(Bersimis et al., 2021), Shao & Lin(Shao & Lin, 2019), and Boran & Diren(Boran & Diren, 2017). Bersimis introduced a meta-method that combines the results of four well-known analytical methods. These are Doganaksoy, Faltin, and Tucker's (DFT) algorithm, Murphy's (MUR) out-of-control algorithm, Mason, Young, Tracy (MYT) decomposition method, and Maravelakis and Bersimis' (MAB) algorithm. Then transform the output of these analytical methods so that they can be used as input to train the multilayer perceptron Artificial Neural Network (MLP ANN).

Shao & Lin (Shao & Lin, 2019) proposed Time Delay Neural Network (TDNN) model and compare the results with Artificial Neural Networks(ANN), Support Vector Machines(SVM), and Multivariate Adaptive Regression Splines(MARS). The result shows that TDNN outperforms the other three techniques in classification accuracy. Boran & Diren (Boran & Diren, 2017) proposed a model by forming an individual control chart for every variable and determining out-of-control conditions for every control chart. Then the multilayer NN model is to be developed based on the values of individual control charts.

2.6.8 Instance-based learning

Instance-based learning (IBL) algorithm is an extension of the k-nearest neighbor algorithm that maps instances to categories. Hence it is a classification function algorithm. Its basic assumption is similar instances have similar classifications. It applies mainly distance functions and probabilistic concepts to its analysis(Aha et al., 1991).

K-Nearest Neighbor (K-NN) is one of the widely used classification techniques. It was first introduced by Skellam (Skellam, 1952) then Cover & Hart(Cover & Hart, 1967) proposed the K-Nearest Neighbor algorithm, and it is the foundation of most of the classification techniques. It is a nonparametric type. K-Nearest Neighbor (K-NN) algorithm uses a database in which the data points are separated into several separate classes, and then the classification of new sample points is predicted based on distance measures of nearest neighbors.

The earliest method of instance-based learning is Condensed Nearest Neighbour (CNN) proposed by Hart(Hart, 1968). Then several extensions and invariants of instance-based learning are introduced. Verma(Ranjan et al., 2019) has identified the following invariants: Locally adaptive KNN, Weight adjusted KNN, Improved KNN for text categorization, Adaptive KNN, KNN with shared Nearest Neighbors, KNN with K-Means, SVM KNN, KNN with Mahalanobis Metric, Generalized KNN, Informative KNN, and Bayesian KNN. There are also other types like Selective NN (Ritter et al., 1975) Generalized condensed NN (Chou et al., 2006), Edited NN(D. L. Wilson, 1972), IB1, IB2, and IB3(Aha et al., 1991) DROP1,…DROP5 (D. R. Wilson & Martinez, 2000), and Iterative case filtering algorithm(ICF) (Brighton & Mellish, 2002).

2.7 Summary of literature review

The trend of publications in the field of industrial process control has evolved, progressing from univariate statistical process control to multivariate statistical process control, followed by data mining-based process control, and ultimately machine learning-based process control. In the domains of univariate and multivariate statistical process control, there is a strong emphasis on rigorous statistical and mathematical analysis. The integration of data mining techniques with traditional process control methods has allowed for the extraction of valuable information from large databases using algorithms developed by computer scientists. Meanwhile, machine learning techniques have introduced a new paradigm of learning algorithms.

Each group of techniques possesses its strengths and weaknesses. For example, traditional statistical process control methods leverage mathematical and statistical theories, such as distribution theories, which cannot be replicated by machine learning algorithms. Conversely, conventional statistical process control methods can be time-consuming and require a solid understanding of statistical concepts to be effectively applied in the production process. Moreover, data mining and machine learning algorithms excel at extracting features from big data, a task that is not easily accomplished by conventional SPC methods. Thus, it is important to note that the trend of publications over time does not imply that earlier research methods on SPC have been replaced by more recent ones.

Research on multivariate statistical process control techniques continues to address unresolved theoretical and empirical challenges, such as identifying which variable or group of variables triggers an out-of-control signal. Another characteristic of recent machine learning-based process control techniques is the reduced emphasis on rigorous mathematical and statistical analysis in the algorithms. This shift may influence the research direction in this field, moving away from purely mathematical and statistical approaches and towards the utilization of machine learning and artificial intelligence algorithms.

The literature review of statistical process control methods reveals a research trajectory that has transitioned from statistical-based methods to data mining and machine learning algorithms. Some data mining techniques integrate statistical and information extraction algorithms to leverage insights from large datasets. In contrast, machine learning methods rely on learning algorithms and may not require extensive datasets like data mining techniques do. These methods, such as decision trees, reinforcement learning, and artificial neural networks, depend on the logical relationships present within the data rather than intensive statistical theories.

As a result, a research gap emerges in terms of comparing the performance of data mining and machine learning techniques, which has not been adequately covered in the existing literature.

To summarize, after conducting a comprehensive literature review in the field of statistical process control, the researcher has developed a conceptual model to address the research problem outlined in this study, as depicted in Figure 0-10.

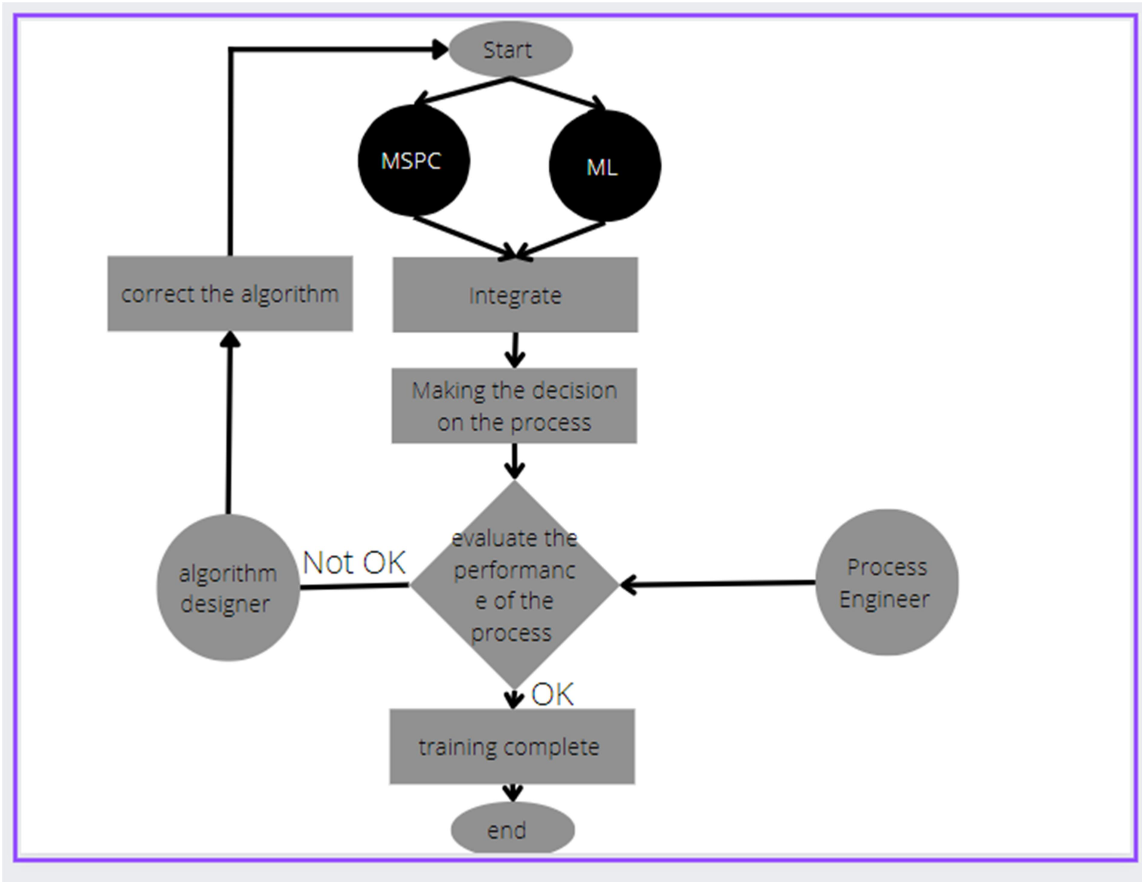

Figure 0-10 A Conceptual model to address the research problem

CHAPTER THREE
RESEARCH DESIGN AND METHODOLOGY

Chapter three of this dissertation is the roadmap of the research which is designed to accomplish the objective of the study. It identifies the type of research that shall be conducted, the data collection sources, and the methods and the tools to be used.

3.1 Introduction

Since the initiation of industrialization in human history, the scientific study of process control has been undertaken. However, as industrialization has become more complex and technological advancements have emerged, the approach to process control has evolved over time. This research acknowledges the changing landscape of industrialization and aims to explore and investigate the concept and nature of process control in the contemporary context.

The research method employed in this study aims to provide a clear direction in answering the research question, which revolves around process control techniques. The central theme of this research is to gain a comprehensive understanding of the concept of "process" and how it is controlled and monitored in various industries. By delving into the intricacies of process control, this study seeks to acquire in-depth knowledge about industrial operations and the effective management of manufacturing or service-oriented processes.

Process in the production system has multiple inputs and outputs variables. Understanding the relationship between these variables is a formidable task. Statistical and mathematical theories and philosophies are intensively applied to design the algorithms. However, research findings and articles disclose that this research is still underway to describe the nature of the relationship between process parameters.

Scientific research is done to understand the nature of the phenomena. There are two types of research. These are Basic and Applied research. Basic research collects and uses empirical data to formulate, expand or evaluate theory and finally discover knowledge. Basic research has the following three forms: Discovery, Invention, and Reflection(Zegeye et al., 2009). Discovery refers to a very new idea or explanation that emerges from empirical research which may revolutionize thinking on that particular topic. The invention refers to a new technique or method created. Reflection refers to an existing theory or technique, or that group of ideas is re-examined possibly in a different organizational or social context. This research follows the invention form of the basic research.

Applied research is conducted to solve practical problems in society. Its primary goal is to solve the immediate societal problem(Y. K. Singh, 2006). Adding to the body of knowledge is a secondary objective. In this research 'Process' is referring to a generic term. It can be applied to every industry, either the production or service industry. But the immediate problem can be easily identified, for example in the case of cement production. Furthermore, the research gap is described in statistical and mathematical terms.

This research is also having the characteristics of the explanatory type of research. Explanatory research looks for causes and reasons. It uses facts or information already available and analyzes these to make a critical evaluation of the material (Kothari, 2004). It answers the question "Why". Explore the cause and effect of the variables. This means extracting knowledge from a reliable data source i.e. from the database.

This research goes from inception to implementation of a new multivariate statistical process control method. The inception phase deals with the nature of basic research. It requires the introduction of an innovative idea. The validation and implementation phase deals with the nature of applied research where empirical data collection and analysis are undertaken. Hence, the research direction will have the characteristics of a mixed research type.

The first objective of this research is to examine the fundamental nature of "process control." By delving into the theoretical foundations and historical development of process control techniques, the study seeks to uncover the underlying principles that govern effective process control. This exploration will shed light on how the concept of process control has evolved over time and how it has been shaped by technological advancements and industrial complexities.

Furthermore, the research aims to investigate how patterns and sequence of operations within processes can be used to optimize resources. By analyzing and understanding the patterns that emerge in industrial processes, researchers can gain insights into the factors that contribute to variations, inefficiencies, or deviations from desired outcomes. This analysis will enable a comprehensive understanding of the challenges and opportunities associated with process control.

Based on the findings from the exploration of process control concepts and the identification of patterns, this research will develop a novel method for monitoring, controlling, and improving industrial processes. By leveraging the latest advancements in statistical process control techniques, multivariate analysis, neural networks, and graph theory, the study aims to propose an innovative approach to process control. This method will consider the complex interplay of variables, the correlation effects, and the historical sequences of successful operations to enhance decision-making processes and provide valuable support to process engineers and operators.

Through the implementation of this proposed method, which will be validated in the context of the cement industry, the research endeavors to bridge the existing theoretical and technical gaps in statistical process control techniques. The ultimate goal is to provide a practical and effective solution that can be applied to real-world industrial settings. By addressing these limitations and offering a comprehensive framework for process control, this research seeks to contribute to the advancement of the field and pave the way for future research endeavors.

3.2 Research Framework

This research adopts a systematic research framework, as illustrated in Figure 0-1, to address critical issues in industrial process control. The framework encompasses four distinct stages, each contributing to the overall objective of bridging the identified research gap. The first stage involves conducting an extensive literature review to identify gaps in existing knowledge and ascertain whether specific problems in industrial process control have been adequately addressed. Statistical process control chart techniques serve as the focal point of the literature review, leading to the identification of the research gap. The second stage involves the simulation of human decision-making processes, the design and implementation of an algorithm merging graph theory, Neural Networks, and Hotelling T2. The third stage is validation of the model using primary data collected from a real-world case company—a cement factory. Finally, the fourth stage deals with summarize and report findings. The detail of each stage is discussed below.

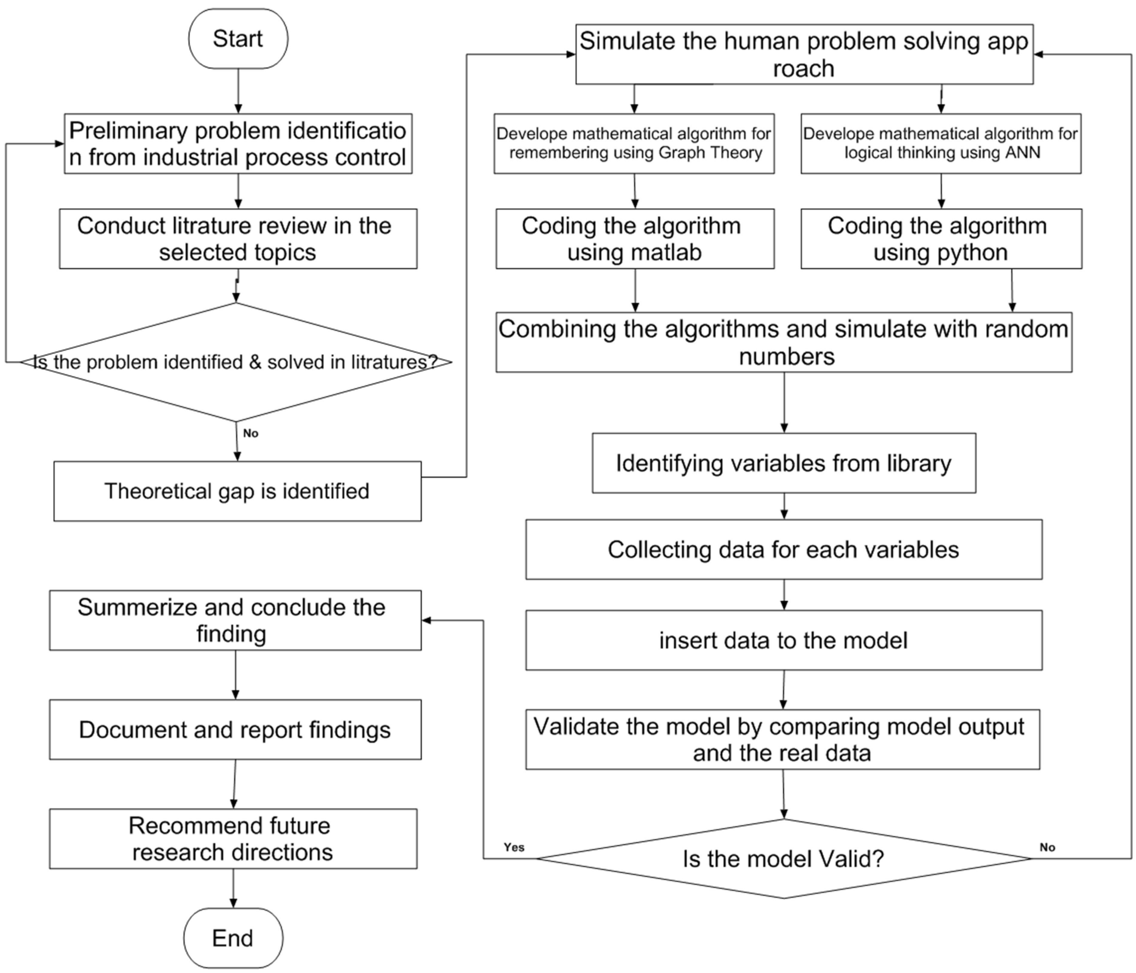

Figure 0-1 Research Framework

Stage 1: Identifying the Research Gap through Literature Review

The research framework commences with a meticulous literature review aimed at identifying gaps in existing knowledge pertaining to industrial process control. By thoroughly examining relevant literature, the researcher identifies problems and challenges prevalent in this domain. A comprehensive review of statistical process control chart techniques is conducted, enabling the researchers to understand the theoretical context surrounding these techniques. Subsequently, the identified gaps in the literature serve as the foundation for the subsequent stages of the research framework.

To meet the research objective of this study, a comprehensive and rigorous literature review (LR) technique was employed. The LR method outlined in Figure 0-2 was developed after conducting a preliminary survey of articles and proceedings of statistical process control chart design techniques. The chosen search strategy entailed a literature survey coupled with meta-analysis, ensuring a systematic, reliable, and exhaustive examination of the existing body of knowledge in the field.

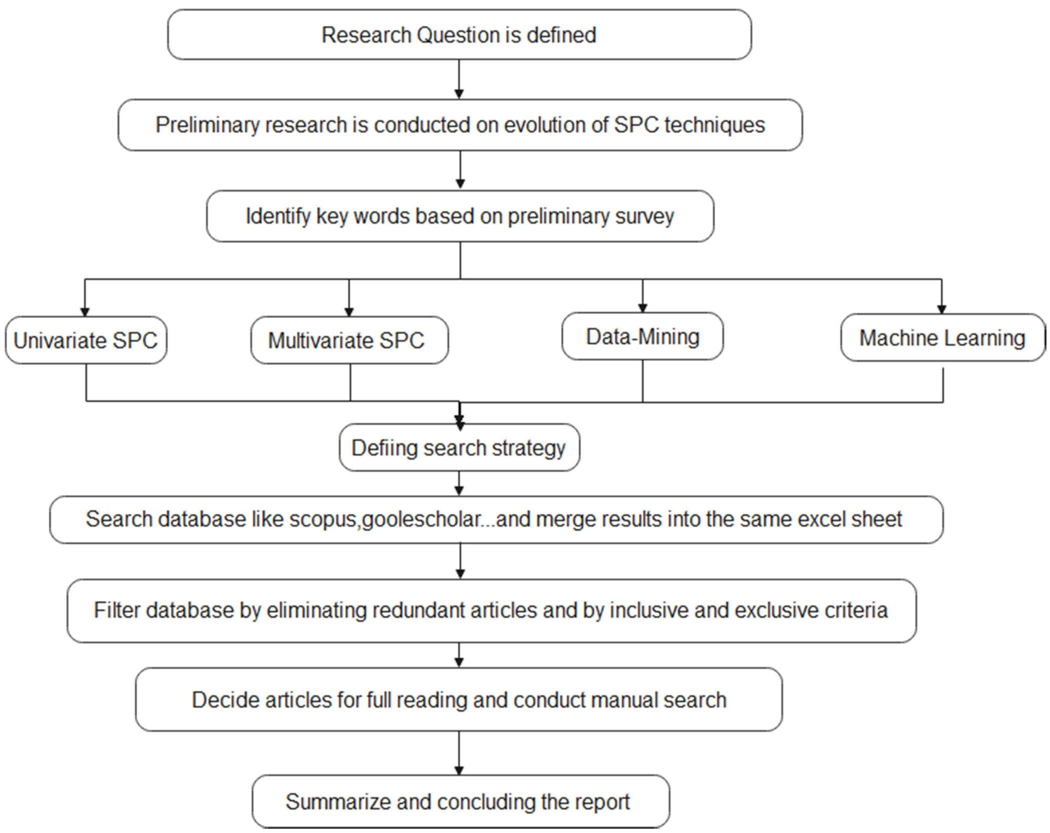

Figure 0-2 Literature review flowchart

Preliminary Survey: Statistical control chart methods have a rich and long-standing history dating back over a century, leading to an abundance of literature being created in this field. Extracting relevant articles that address research questions can be an extremely daunting task; VOSviewer software serves as an indispensable resource in aiding this search work and is therefore invaluable in this endeavor. It offers researchers powerful features to organize, visualize, and interpret collected literature in meaningful ways that highlight relationships among articles, authors, keywords, research topics, etc. These visualizations assist researchers in quickly locating relevant articles while uncovering key themes or trends within literature collections.

VOSviewer's advanced search capabilities enable researchers to filter and refine search results, facilitating the extraction of pertinent information. Utilizing VOSviewer, researchers can efficiently navigate through statistical control chart literature while making sure their articles directly align with their research question at hand. VOSviewer streamlines this search process while still meeting the rigor and comprehensiveness required for an effective literature review. Discussion to each literature review topics are discussed below.

As part of this LR work an initial preliminary survey was undertaken with the help of VOSviwer to gain an in-depth knowledge of statistical process control chart design concepts and existing literature on the case company which is cement production system and its process control. This survey is a basis for subsequent stages.

Formulating Search Criteria: Utilizing preliminary survey findings, specific search criteria were designed to locate relevant scholarly articles, research papers, conference proceedings, and other sources of information related to multivariate process control, neural networks, graph theory Hotelling T2, as well as statistical process control techniques applied within the cement industry.

Database Selection: Well-established academic databases like Scopus, IEEE Xplore, ScienceDirect, ACM Digital Library, and Web of Science were chosen to gain access to a diverse collection of high-quality research publications.

Search Execution: Formulated search criteria were applied to databases selected to retrieve relevant articles, using keywords and Boolean operators strategically to narrow the search and gain the most pertinent literature.

Screening and Selection: Search results were put through a systematic screening process to identify relevant articles for our research topic. Titles and abstracts were closely evaluated to ascertain their relevancy; those meeting inclusion criteria were selected further for examination; irrelevant or duplicate publications were rejected outright.

This paper provides a thorough literature review conducted to identify gaps in statistical process control techniques. The review spans four topics, including early univariate and multivariate process control techniques, data mining techniques, machine learning algorithms and machine learning models. By exploring literature across these subjects this review offers an in-depth examination of development stages within statistical process control while investigating temporal dynamics that reveal periods of blooming or decline over time in various techniques.

Topic 1: Earliest Univariate Statistical Process Control Techniques:

This literature review begins by investigating the early univariate statistical process control techniques. This topic explores foundational techniques that established statistical process control as a field; by looking at pioneering works and seminal articles published during its early days, this review offers insight into its initial advancements as a field. Furthermore, its temporal progression will also be studied to illustrate their relevance and impact at specific periods in history. The study of univariate statistical process control methods is found to be important to the study of multivariate process control because almost all multivariate techniques are extension of univariate statistical process control where the basic concepts of statistical process control is designed.

Topic 2: Multivariate Statistical Process Control Techniques

As a second part to its literature review, multivariate statistical process control techniques become the focus. This area covers an expansive array of methodologies designed to monitor and control multiple variables simultaneously. Taking an in-depth look into their evolution over time through surveying literature as well as researchers' contributions, this analysis uncovers key stages and advancements which have contributed towards shaping contemporary understanding and application of multivariate statistical process control.

Topic 3: Data Mining Techniques:

This literature review goes beyond multivariate statistical process control by exploring how data mining techniques fit into process control methodologies. This topic investigates clustering, classification and association rule mining approaches used for statistical process control as part of statistical process mining approaches. By looking at existing literature reviews regarding advancements made in data mining techniques and their application in analyzing process data; providing valuable insight into where statistical process control meets data mining - uncovering research gaps as well as opportunities for further development.

Topic 4: Machine Learning Techniques

This literature review's final topic centers around machine learning techniques in statistical process control. This emerging field explores how neural networks, support vector machines, decision trees, and similar techniques are being utilized by machine learning algorithms to monitor and control processes. Furthermore, this review investigates their integration within statistical process control frameworks as well as their ability to handle complex and high-dimensional data sets efficiently. Through analysis of existing literature and reviewing its stages of development for machine learning techniques as well as potential contributions they might bring towards improving process control methodologies.

Full-Text Review: Selected articles underwent a detailed full-text review, in which each was carefully considered to assess whether or not it met our inclusion criteria for the LR. The content, methodology, findings, and contributions of each article were carefully assessed to gather as much relevant information as possible from each.

Data Extraction and Analysis: Key data and information were extracted and organized from each article selected, with key findings, methodologies, theoretical frameworks, and practical insights documented for further examination. This step enabled the identification of common themes, patterns, or gaps within existing literature.

Meta-Analysis: To synthesize and integrate findings from selected articles, a meta-analysis was conducted. By combining and comparing the results of multiple studies, this analysis provided an in-depth view of current research into statistical process control techniques in cement industry settings.

Summarization and Synthesis: After compiling and analyzing all available data and analyses, key concepts, theoretical frameworks, methodological approaches, and empirical findings were summarized and synthesized into an effective narrative to form the basis for subsequent stages of research.

Gap Identification: Through the LR process, gaps were identified within existing literature that encompass both theoretical and practical elements – this allowed researchers and innovators to focus their research and innovation efforts in these areas of need.

The rigorous LR methodology outlined above ensured that research was founded upon solid existing knowledge, providing an in-depth view of its current state. Through taking an analytical, systematic, and exhaustive approach, this LR contributed to credibility and validity in subsequent research activities resulting in more robust, insightful results.

Temporal Dynamics and Development Stages: One of the key aspects of literature reviews of this dessertation is their analysis of temporal dynamics and development stages in statistical process control techniques. By surveying literature across all four topics, researchers can gain valuable insight into their evolution as well as identify research gaps for further investigation. By taking this approach to studying statistics process control techniques, one of their greatest contributions may be their revealing of temporal dynamics and development stages over time. By tracking advancements or decline of different techniques over time - such as advancement or decline over time or dominance or decline - providing invaluable information about advancement or decline over time as well as interplay among methodologies in this way illustrating periods when techniques dominated or lost relevance over time - researchers gain invaluable insights into its evolution allowing further investigation while uncovering research gaps which require further investigation.

Stage 2: Simulating Human Decision-Making Processes to design Algorithm

The second stage of the research framework focuses on developing the algorithm that simulates the decision-making processes of process engineers and operators. A human being is a smart and intelligent creature in this world. Process engineers apply their cognitive skills to run the process efficiently based on their capability. While designing the algorithm for this research, the human problem-solving methods are simulated. Two elements of the human mind are considered in this research. These are remembering and logical thinking. Some decisions needed previous experience. During this time, remembering the events and the best decisions are considered, and accordingly, the decision is performed. Remembering events and sequence of operations are simulated by graph theory. Logical reasoning of the human being is simulated by artificial neural network. In this case, simulating the biological nervous system by an artificial neural network was one of the research areas to assist the human decision support system.

To bridge the identified research gap, an algorithm is designed by merging three prominent theories: graph theory, Neural Networks, and Hotelling T2. These theories provide a comprehensive framework for effectively addressing the challenges in industrial process control. The algorithm is implemented by writing code in both Matlab and Python, leveraging the strengths of these programming languages. The output of the algorithm is then displayed using Gephi software, which provides intuitive and visually appealing representations of the results.

To effectively translate the algorithm into coding, the researcher carefully selected appropriate data analysis tools. The algorithm, being a sequence of rules and procedures, required a solid foundation based on classified knowledge. In the context of process control, this knowledge was extracted from the process database. Graph theory, Neural Networks, and Hotelling T2 emerged as the theoretical frameworks underpinning the algorithm, providing valuable insights into data analysis methodologies.

Data collected for analysis was organized and represented in matrix form. Matlab software was chosen as the ideal tool for matrix manipulation due to its powerful capabilities in handling large data sets and conducting complex mathematical operations. The researchers utilized Matlab's specialized coding language to carry out matrix calculations efficiently. Specifically, the Hotelling T2 calculations were implemented using Matlab's coding language, leveraging its mathematical prowess to derive meaningful insights from the data.

Stage 3: Validation of the Model

In order to validate the effectiveness of the model developed in the previous stages, primary data is collected from a real-world case company—a cement factory. This primary data serves as a crucial component in assessing the practical applicability and relevance of the research findings. The collection of primary data enables the research team to evaluate the performance of the algorithm and assess its impact on the industrial process control within the case company. This validation stage ensures that the proposed model aligns with the real-world scenario, bolstering its credibility and practical significance.

Obtaining reliable and accurate data is crucial to ensuring the quality and validity of research findings. In this study, the collection of primary data from a case company's production site was carefully planned to enhance the research outcomes. This data was collected through the implementation of sensors and laboratory analysis, allowing for a comprehensive understanding of the processes at hand. Additionally, it was assumed that the collected data would be efficiently stored within the database of the control system. This paper aims to delve into the importance of primary data and highlight the accessibility and usability of such data, which is stored in the control system's database in the easily accessible format of an Excel spreadsheet.

The significance of reliable primary data cannot be overstated. It serves as the foundation for sound research and aids in the development of accurate findings and conclusions. By collecting primary data directly from the case company's production site, researchers can gather firsthand information about the processes, variables, and phenomena under investigation. This ensures a higher level of accuracy compared to relying solely on secondary data sources, which may not be tailored to the specific research objectives.

To ensure the integrity of the primary data, the researchers employed a well-designed data collection strategy. This involved the utilization of records from sensors placed strategically throughout the production site, enabling the continuous monitoring and recording of various parameters. The sensors provided real-time data, allowing for a comprehensive understanding of the production processes, including factors such as temperature, pressure, humidity, and flow rates.

Furthermore, laboratory analysis played a vital role in supplementing the sensor data. By conducting thorough laboratory tests on collected samples, researchers could gain insights into the chemical composition, quality, and other relevant characteristics of the materials being processed. This combination of sensor data and laboratory analysis ensured a multidimensional and accurate representation of the production site, providing a solid foundation for research conclusions.

After collecting the primary data, it was imperative to store it in a reliable and accessible manner. The chosen approach involved storing the captured data within the database of the control system. This database was designed to handle large volumes of data, accommodating the extensive information collected from the production site. The secure storage of data within the control system's database minimized the risk of data loss or corruption, ensuring the reliability and integrity of the primary data.

In terms of accessibility, the stored data was made available to authorized operators within the company. This allowed them to access and analyze the data for internal purposes, such as process optimization, quality control, and decision-making. External parties, such as research collaborators or regulatory agencies, could also gain authorized access to the data, ensuring transparency and facilitating collaborative research efforts.

The data stored within the control system's database was presented in an easily accessible format, specifically an Excel spreadsheet. This choice of format provided a user-friendly interface for researchers, operators, and external parties to navigate and analyze the data. The spreadsheet format enabled efficient data manipulation, sorting, and visualization, further enhancing the usability and interpretability of the primary data.

Stage 4: Summarize and report findings

Summarize and report the findings are important part of research work. The literature review has reported the reviews of more than 200 articles including research gap and conceptual model to solve the research question. The algorithm is developed conceptually from the theories of graph, ANN and Hotelling T2 and the code is written using MATLAB and python. Primary data collection is conducted in the plant site and laboratory analysis of the case company. Conclusions are drawn from the result of the analysis. The result is summarized and reported in proper format.

CHAPTER FOUR

MODEL DEVELOPMENT TO MULTIVARIATE PROCESS ANALYSIS

This chapter propose new assumptions to design the model that answers the research question, then the model is designed. Ideal process which has 10 variables with 100 records is randomly generated from computer to show the outcome of the Algorithm. From these 10 variables, 2 are input variables, 3 are controllable variables, 2 are uncontrollable variables, and 3 are output variables. Then the Model is trained and tested. Finally, conclusion is drawn from its findings.

4.1 Introduction

This chapter's primary goal is to create a model which enhances traditional multivariate statistical process control methods, specifically Hotelling T^2 technique improvements. Hotelling T^2 can identify anomalies within production systems with remarkable accuracy; however, it lacks the capability of proposing strategies for correcting them and returning the system to normal functioning. Hotelling T^2 only predicts the statistical distance of the vector at one moment in time, failing to take account of how different operations within a production system interact with each other. Therefore, this research seeks to develop a technique which not only provides variable values for subsequent steps taken towards normal operation of a system but also considers their sequence in order to make more informed decisions.

This chapter is divided into four sections to explore and present the proposed model thoroughly. The first serves as an introduction, providing an overview of graph theory that forms its core. Next, understanding principles and assumptions governing and constraining platform is paramount - therefore the second section examines these core assumptions which shape its design.

In the third section, step-by-step procedures of the model are defined and explained in great detail. This section details a systematic approach for using graph theory and advanced machine learning techniques to enhance Hotelling T^2 method. These procedures aim to address limitations inherent in traditional techniques while providing operators with valuable insights for corrective actions and strategies for corrective actions.

Moving forward, the fourth section focuses on translating proposed procedures into code and practical implementation of them through computerized simulation models. In particular, this section details how random sample data generated by computer systems is utilized to assess effectiveness and accuracy of implemented codes as well as ensure their suitability for real world applications.

Finalizing this chapter's findings and conclusions with a reflection on their significance within statistical process control is its final section, providing a summary of findings and conclusions drawn from this chapter. It offers an in-depth account of advancements achieved through the proposed model as well as potential ramifications for multivariate process control techniques. By integrating graph theory and machine learning techniques into this model's anomaly prediction capabilities as well as guidance for corrective actions taking into account sequence of operations within production systems - this final chapter concludes by reflecting upon these findings as well as their possible impact within statistical process control research itself.

4.2 Introduction to Graph Theory

Graph-based model is a popular model-building tool for both engineering and social science. It has a long history of origin. The concept of Graph was introduced by Euler in 1736. Since then it is applied to solve several practical problems.

In this research, graph theory is the heart of the model and is mainly applied to design a multivariate statistical process control chart. Graph theory has many important features that can be used to solve the process control problem. As discussed in the literature review in chapter two, research has been conducted on the application of graph theory to SPC as a data mining tool. However, in this research graph is integrated into Artificial Neural Network to develop an algorithm that can identify the cause of the fault and propose a diagnosis to return the out-of-control process to controlled states.

Graph $G = (V, E)$ is defined as a pair of sets: a set of vertices, V, which denotes the entities, and a set of edges, E, which represent relationships or connections between the entities. Graph G is also defined as an ordered triple $(V(G), E(G), \psi G)$ consisting of a nonempty set $V(\ G)$of vertices, a set $E(G)$, disjoint from $V(G)$, of edges, and an incidence function ψG that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G (Bondy & Murty, 1976). If e is an edge and u and v are vertices such that $(e) = uv$, then e is said to join u and v; the Vertices u and v are called the ends of e.

Two vertices a and b are adjacent if there exists an edge (a, b) that connects them. A sequence of adjacent vertices is known as a walk. A walk, in which no vertex occurs more than once, is known as a path. A walk, in which no edge occurs more than once, is known as a trail and a closed trail is a closed walk that is also a trail. A path is a trail in which all the vertices in the sequence are distinct. A cycle is a closed trail in which all the vertices are distinct, except for the first and last, which are identical. A graph is connected if there exists a walk between every pair of its vertices. A graph that is not connected is called disconnected. A tree is a connected graph that contains no cycles. The subgraphs which are maximal for the property of being connected are called the components of G.

A graph can be categorized as a directed or undirected graph. A directed graph is a pair (V, A), where V is a nonempty set whose elements are called the vertices and A is the subset of the set of ordered pairs of distinct elements of V. Each edge,e, of graph G is associated with an ordered pair of vertices. Hence edges are usually represented by arrows pointing in the direction the graph can be traversed. In an undirected graph, the edges are bidirectional, with no direction associated with them. Hence, the graph can be traversed in either direction. The absence of an arrow tells us that the graph is undirected.

A graph can be represented in matrix form. Adjacency matrix, A(G), of a directed graph G is an mxm matrix consisting of only entries 0 and 1, where m is the number of vertices of G. The entry a_{ij} is equal to 1 if there exists a directed edge from vertex v_i to vertex v_j, otherwise, it is equal to 0. All vertex Incidence matrix of a non-empty and loopless directed graph G is $A = (a_{ij})$, where $a_{ij} = 1$ if v_i is the initial vertex of e_j, -1 if v_j is the terminal vertex of, 0 otherwise.

Coloring is an important concept in graph theory. It is defined as the coloring of vertices and edges with a minimal number of colors such that no two adjacent vertices or edges should have the same color. The minimal number of colors is called a chromatic number and the graph is called a properly colored graph.

A graph invariant is a function $I(G)$ that associates numerical values to each graph $G = (V, E)$ regardless of the way vertices or edges are labeled. The number of vertices $|V| = n$, the number of edges $|E| = m$, the maximum distance between two vertices (diameter), and the chromatic number are some of the examples of invariants.

4.3 Basic Principles and Assumptions of the model development

Mathematical and statistical concepts are the foundation of this model. This section of the dissertation discusses the refined final assumptions and the principles focusing on the development of an alternative model design to monitor and control processes in the continuous production system. The assumptions consider the different possible scenarios that could encounter during the production process.

Assumptions

To develop the model the following five assumptions are established.

Assumption 1. Variables in the production system are grouped into four sets.

The production system is a complex environment comprising a multitude of variables. Classify variables helps to simplify the system. A system is a collection of processes either in a closed or open environment. The process is a collection of activities that transform inputs into outputs. Consider now the process in the production system having inputs like raw materials, uncontrollable variables like environmental variables, controllable variables in the process like motor control, and the desired output variables like product quality characteristics. In general, there are four groups of variables that will need to be controlled and monitored in the process of the production system. These are input variables [I], controllable process variables [C], uncontrollable process variables [U], and output variables [Y]. Diagrammatically it is represented by the following figure, Figure 0-1.

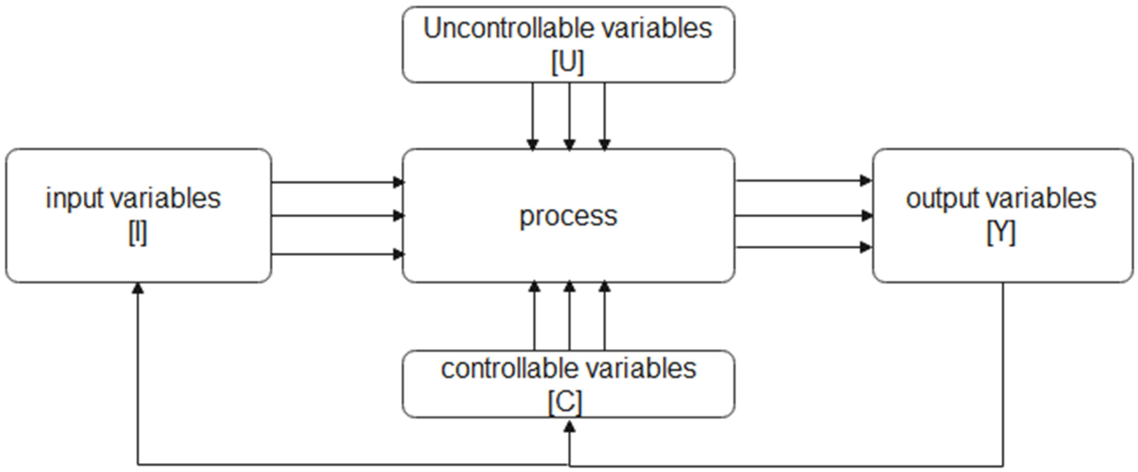

Figure 0-1 Model for the system

Assumption 1 defines the scope and demonstrates the general overview of the model. The scope is to mean the model is suitable for an open system, which means external uncontrollable variables are considered as long as it has an effect on the production system. There is no limitation to the number of variables to be considered in each group of variables in the model.

Input variables are variables that represent the physical components that enter to the system. These input variables are monitored by the process engineer. Output variables are variables that represent product characteristics of the production system. Output variables are controlled as a function of input variables, controllable variables, and uncontrollable variables.

Controllable variables are variables in the production system that can be monitored and controlled by the process engineer. These are process parameters that decide the product characteristics.

On the other hand, uncontrollable variables encompass those that originate from the external environment or are generated within the internal production system but remain unmonitored and uncontrolled by the process engineer. The model periodically reads the values of these uncontrollable variables, giving rise to multivariate time series (MTS) characteristics. MTS involves the presence of multiple time-dependent variables, each depending on its past values while also being influenced by other variables. The most suitable model for this group of variables is the Vector Auto-Regressive Integrated Moving Average (VARIMA) model. Determining the appropriate order of Autoregression (AR) value (p), order of moving average (MA) (q), and degree of differencing (d) is achieved through a grid search that minimizes the information criteria. Furthermore, the value of d is determined to ensure stationarity across the all-time series.

Mathematically it is expressed as:

$$Y_t = [c] + [\beta_1][Y_{t-1}] + \dots \dots + [\beta_p][Y_{t-p}] + [\alpha_1][e_{t-1}] + \cdots . + [\alpha_q][e_{t-q}] \dots\dots\dots (24)$$

Where: $[c]$ is the intercept with k x 1 matrix

$[\beta_i]$ is the coefficient of k x k matrix of

$[Y_{t-i}]$ is k x 1 matrix of time lag for variable i

$[\alpha_1]$is the coefficient of k x k matrix of error r terms

$[e_{t-i}]$ is k x 1 matrix for error e terms

Assumption 2. A two-dimensional matrix can be used to summarize and define terms and parameters

Data set K: is the data collected sequentially in the production process that contains all four sets of variables i.e. [I], [U], [C], and [Y]. Mathematically it is represented as follows, Where I represent input, c controlled output, u uncontrolled output, and o stands for output.

$$K_{r,v} = \left[\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_r \end{bmatrix} \begin{bmatrix} i_{1,1} & i_{1,2} & \cdots & i_{1,p} \\ i_{2,1} & i_{2,2} & \cdots & i_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ i_{r,1} & i_{r,2} & \cdots & i_{r,p} \end{bmatrix} \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{r,1} & c_{r,2} & \cdots & c_{r,m} \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{r,1} & u_{r,2} & \cdots & u_{r,n} \end{bmatrix} \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,q} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{r,1} & y_{r,2} & \cdots & y_{r,q} \end{bmatrix}\right] . (25)$$

Where R is sequence number (Record no), i represents p number of input variables, c represents m number of controllable variables, u represents n number of uncontrollable variables, y represents q number of output variables, r represents the number of records and v represent the sum of the columns.

Assumption 3. The data matrix (K) is to be translated into nodes of the graph without losing its structure and information

Variables in a vector within dataset matrix K can be divided into two groups. The first group is formed from the input variables, controllable variables, and uncontrollable variables and can be represented by state S. The second group is formed from output variables and is represented by state O.

Mathematically node S is represented as follows:

$$S = \sum_{i=1}^{r} S_i \quad \text{(26)}$$

$$or\ S = \{S_1, S_2, \dots, S_r\}$$

Where $\sum$ represents collection in set theory

S can also be written as Si is expanded

$$S = \sum_{j=1}^{r} \left[\sum_{k=1}^{p} i_{jk} + \sum_{k=1}^{m} c_{jk} + \sum_{k=1}^{n} u_{jk} \right] \quad \text{(27)}$$

$$or\ S = \{((i_{11}, \dots (i_{1p}), (c_{11}, \dots (c_{1m}), (u_{11}, \dots (u_{1n})), \dots ..((i_{r1}, \dots (i_{rp}), (c_{r1}, \dots (c_{rm}), (u_{r1}, \dots (u_{rn}))\}$$

Similarly, node O is represented as follows:

$$O = \sum_{i=1}^{r} O_i \quad \text{(28)}$$

$$or\ O = \{O_1, O_2, \dots, O_r\}$$

O can also be written as O_i is expanded

$$O = \sum_{j=1}^{r} \left[\sum_{k=1}^{q} o_{jk} \right] \quad \text{(29)}$$

$$or\ O = \{(i_{11}, \dots (i_{1q}), \dots .., (i_{r1}, \dots (i_{rq})\}$$

Each multivariate state of these groups is represented by a single node as follows, Figure 0-2

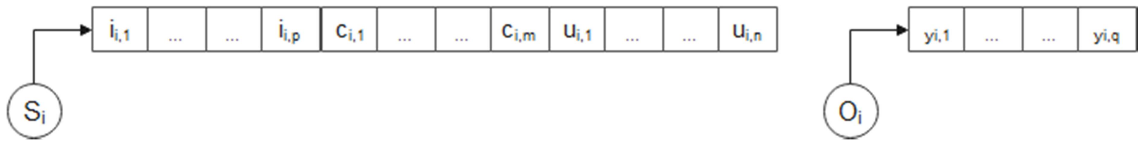

Figure 0-2 Representation of S and O nodes

The matrix K can be represented by :

$$K_{r,v}=\begin{bmatrix}\begin{vmatrix}R_1\\R_2\\\vdots\\R_r\end{vmatrix} & \begin{vmatrix}S_1\\S_2\\\vdots\\S_r\end{vmatrix} & \begin{vmatrix}O_1\\O_2\\\vdots\\O_r\end{vmatrix}\end{bmatrix}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(30)$$

*Hence **Node** is defined by*: a set of Records of dataset K at a specified boundary and identified sequence.

Assumption 4. To simplify and analyze the complex system, the building block of the model must be identified.

For each record at the given moment, two nodes are created. As time elapsed, the number of nodes is increasing and a more complex graph will be created. Hence, it is important to define the building block of the graph.

The building block of the graph to the model is represented by Figure 0-3. Si is the node that represents the state of the input variables to the system at some point in the graph. If the values of some variables are changed and the change variables are represented by G2, the resulting node will be P2. With the given input node, P1, the resulting output node will be Y1. Similarly, the output of P2 will be Y2. The change in output between Y1 and Y2 is represented by node B2. P2 is influenced by B2 because product outputs Y1 and Y2 cannot be dramatically changed. Similarly, Y2 is influenced by G2 because P1 and P2 cannot be dramatically changed. This step will be repeated for each period where the record is added and finally form the complex network.

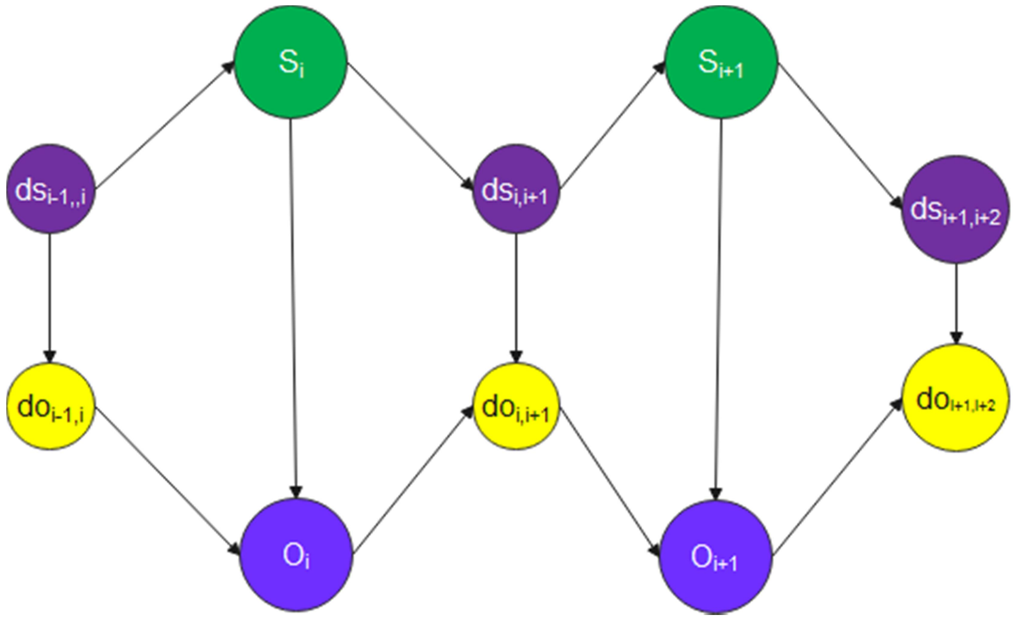

Figure 0-3 Network of nodes

Assumption 5. The process graph is to be defined based on the identified building block

Assuming that the number of records, denoted as R, is represented by individual nodes, the formation of a graph network occurs when these nodes are interconnected through edges based on their sequential arrangement. This concept aligns with assumption 2, which establishes R as the parameter defining the number of records within the network.

By visualizing each record as a distinct node within the network, the interconnectivity between these nodes is established through edges that reflect their sequential order. As a result, a comprehensive graph network takes shape, encapsulating the relationships and dependencies between the records. The edges serve as conduits, linking the nodes in a structured manner that corresponds to the sequential arrangement of the records.

This graph network representation provides a powerful framework for analyzing and understanding the relationships between the records within the dataset. It allows for efficient traversal and exploration of the sequential data, enabling researchers and analysts to uncover patterns, dependencies, and trends that may exist within the network. By capturing the sequential order of the records and establishing connections between them, the graph network provides valuable insights into the underlying structure and dynamics of the dataset. Here the following scenarios are considered.

What if the record appears again?

In the context of a sequence, a noteworthy scenario arises when encountering a duplicate record. In such cases, the system does not generate a new node but rather establishes a connection, or edge, leading to the existing node that represents the identical record. This approach ensures that the sequence remains streamlined and efficient by avoiding unnecessary redundancy. By directing the edge towards the already existing node, the system effectively consolidates and organizes the information, maintaining a cohesive structure while minimizing unnecessary data duplication. This optimized approach allows for more efficient processing, retrieval, and manipulation of the sequence, as it eliminates the need to create additional nodes for duplicate records.

What if the state is returned to the previous state?

In the realm of system operation, the interplay between input and output states is characterized by a direct and interconnected relationship. This connection is paramount in understanding how the system responds to different inputs and generates the corresponding outputs. As the input states undergo a transition from state1 (S1) to state2 (S2), a synchronized transformation unfolds in the output states, shifting from output1 (Y1) to output2 (Y2). This dynamic correlation between the input and output states serves as evidence of the system's adaptability and responsiveness, as it promptly adjusts its behavior to align with the given inputs.

Moreover, when the input state progresses from state2 (S2) back to state1 (S1), an intriguing reversal transpires within the output states. Consequently, the output values undergo a distinct metamorphosis, transitioning from output2 (Y2) back to output1 (Y1). This reciprocal relationship between the input and output states endows the system with a consistent and reversible behavior, wherein alterations in the input states directly correspond to corresponding modifications in the output states, and vice versa.

This intricate interdependence between the input and output states guarantees the system's reliability and predictability. By closely tracking and adapting to changes in the input states, the system harmoniously adjusts its output states, ensuring a harmonious response to evolving conditions. The ability to seamlessly reverse the transformation when the input states revert underscores the system's stability and capacity to maintain a consistent behavior.

Understanding the intimate connection between input and output states is crucial in comprehending the system's behavior and predicting its responses. This dynamic relationship facilitates effective control and management of the system, allowing for informed decision-making and optimized performance in various operational contexts.

This interdependence between input and output states enables the system to effectively respond to varying inputs and produce desired outputs, exhibiting a dynamic and adaptable nature. The ability to transition between different states and revert back to previous states allows for flexibility and versatility in the system's functionality, catering to different operational requirements and ensuring the desired outcomes are achieved.

Where do NN and Graph integrate?

Both Neural Networks and Graph Theory share the fundamental concept of nodes as fundamental building blocks. However, their roles and definitions diverge significantly. In this particular model, the utilization of graph theory introduces two distinct types of nodes. These nodes can be categorized as input nodes and output nodes, each serving a specific function within the system.

From a graph theory perspective, the input nodes (Si) are created by combining three essential vectors: input variables, controllable variables, and uncontrollable variables. These input nodes serve as the entry points for data into the system, providing a comprehensive representation of the various factors that influence the system's behavior. On the other hand, the output nodes (Oi) are generated exclusively from the output variables, encapsulating the desired results or outcomes of the system.

The three vectors of the input nodes traverse through the nodes present in the hidden and output layers of the neural network. The neural network performs complex computations and transformations on the input data, ultimately generating an output vector that corresponds to the output nodes within the graph representation. This integration between the neural network and the graph is visually depicted in Figure 0-4, showcasing the interconnectedness of these two components.

Interestingly, the parameters of the neural network derived from the model play a crucial role in determining the characteristics of the edges within the graph representation. As a result, Figure 0-4 evolves into a more condensed representation, depicted in Figure 0-5. This condensed figure represents the fusion of the neural network and the graph theory, highlighting the harmonious integration of these two fundamental components.

This model, combining elements from Neural Network and Graph Theory, demonstrates a novel approach to capturing and analyzing complex systems. By leveraging the strengths of both methodologies, a comprehensive understanding of the system's behavior and its output characteristics can be achieved. The interplay between nodes, edges, and the computations performed by the neural network paves the way for insightful analyses and informed decision-making within diverse domains.

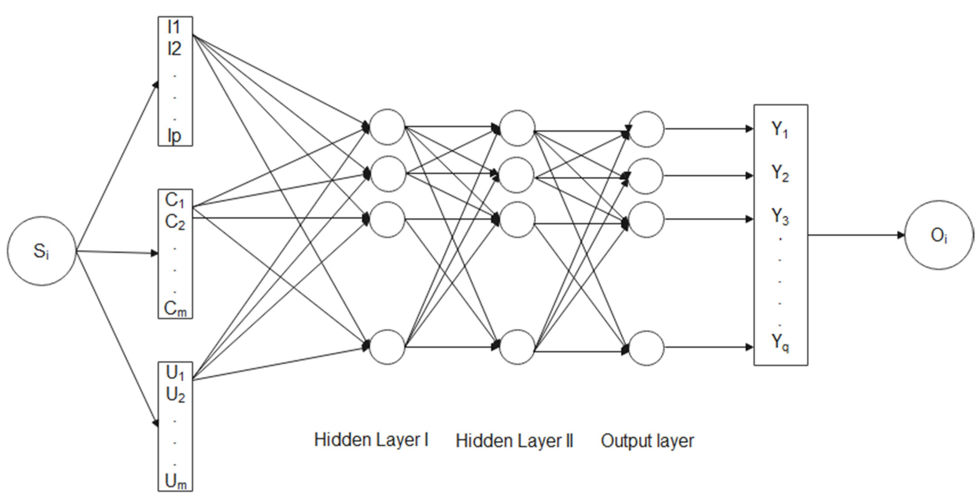

Figure 0-4 Integration of Neural Network and Graph

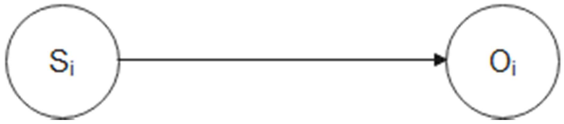

Figure 0-5 The condensed form of Fig 4-4

4.4 Algorithm Development

The following steps are designed to develop the algorithm.

1. **define and collect the dataset K**

Identifying important input variables (I), controllable variables(C), uncontrollable variables (U), and output variables (O) for the model, collecting data periodically to form a vector, and defining the dataset K, or matrix from the collection of these row of vectors. Automated Industries collect data using sensors and accumulate these data on their database engines. Almost all automated industries can transform and deliver their data in Excel format. Hence, in this case, the first step of the algorithm is to communicate with the user to identify the variables, arrange data from Excel and transfer to it the dataset K.

2. ***Discretize the dataset K***

Discretizing the data helps to sort, classify and find data quickly, hence it is an important step for the algorithm. When recorded data in one variable is discretized, first the upper and lower bound of the data is decided then this bound is divided into the desired number of intervals or classes. Finally, every data record is assigned to which class it belongs. Discretizing data is important because the model is based on graph theory which is part of discrete mathematics. The algorithm is represented by Figure 0-6

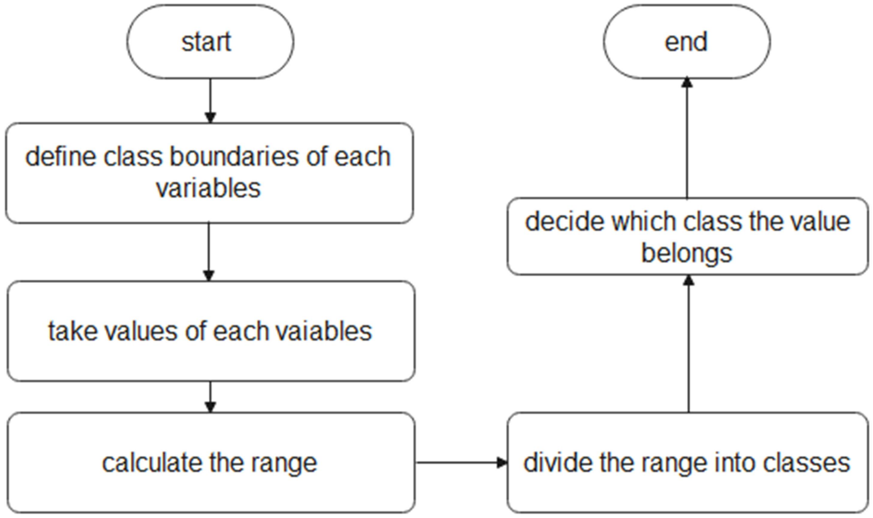

Figure 0-6 Flow sheet for discretizing data

3. ***Create a graph for the dataset K***

Each row of the dataset K form two nodes, the first one, node S, is defined from the I, U, and C variables, and the second node, node O, is defined from the O variable. These two nodes are connected by one edge. Sequences of two vectors are again connected by edges. Then this collection of nodes and edges forms the network.

To define the process of creating the graph from the given database, the following steps are stated, and it is represented by Figure 0-7.

Say R is the number of records and i is the nth number of records in the database.

Step 1: Check whether i is less than or equal to R. If it is greater than R, stop iteration, else go to step 2

Step 2: Separate the vector into two vectors, a1 and a2, where the first vector is formed from the collection of I, C, and U variables and the second vector is formed from the collection of output variables.

Step 3: Check whether that record for a1 is unique, and if it is assign node Si to that record else go to the next record. Calculate the change between Si and Si-1, and create the node dSi. Similarly, check whether that record for a2 is unique, and if it is assign node Oi to that record else go to the next record. Calculate the change between Oi and Oi-1, and create the node dOi.

Step 4: If a new node is created for the record, add an edge between the last and the new nodes as the sequence for S and O. Figure 21 shows the pictorial representation of the above steps

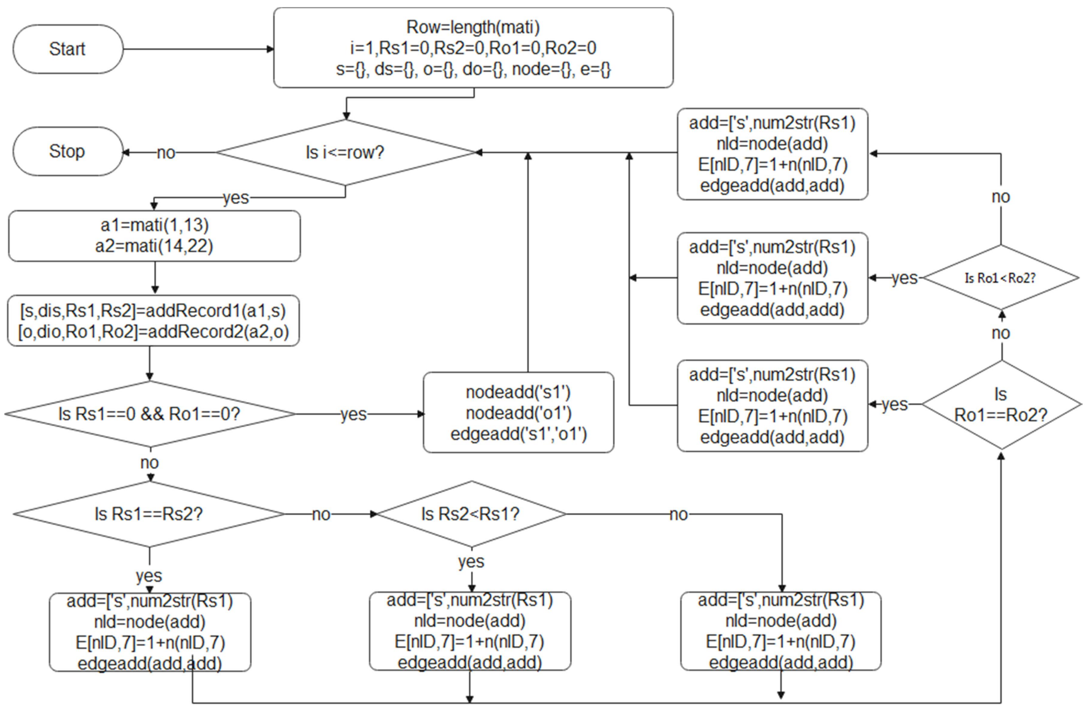

Figure 0-7 Algorithm to develop a graph

4. ***Design the Neural Network for the dataset K***

Two neural networks are designed to help the decision-making process. The first one takes values I, C, and U values as an input And O values as the output of the Neural Network. The second one is formed from O and U variables as input and I and C as output. The feed forward-Neural Network (FFNN) with gradient descent algorithm is designed from the relationship of input nodes and output nodes variables. The feed-forward NN helps to predict the output from the given input node. To minimize the prediction error, the number of layers is increased and the deep neural network (DNN) design techniques are implemented.

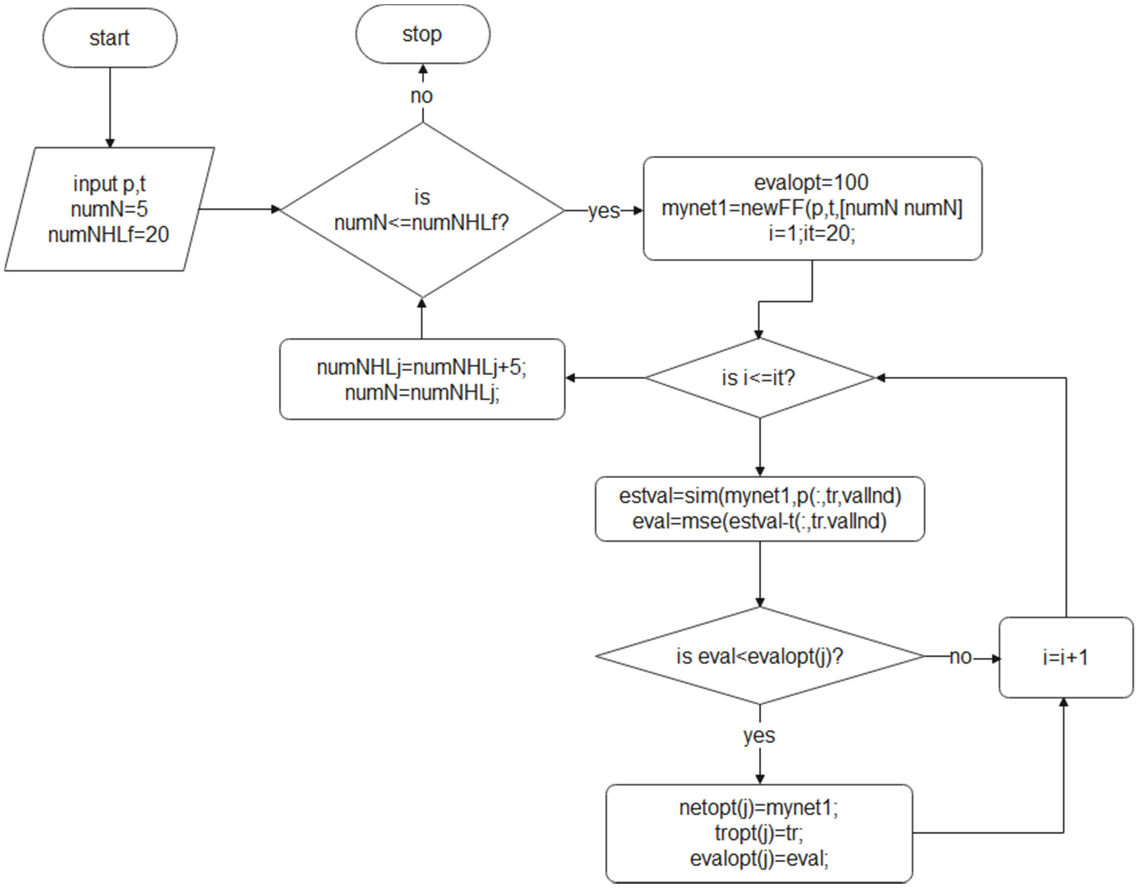

Figure 0-8 algorithm of Neural Network

The steps to be followed to develop the Neural Networks are as follows,

Step 1: Start from one hidden layer having 5 nodes for the network

Step 2: Increase the number of node steps to 5 up to 100 and calculate the error for each iteration. Calculate the error term for each iteration and if the error term is within the acceptable range stop and save the network parameters as the model else go to step 3.

Step 3: increase the number of hidden layer step by one then go to step 2 until the number of hidden layers is reached where the error value has reached the acceptable range.

Figure 0-8 shows the pictorial representation of the above steps

5. ***Display the result using a graph displaying applications***

The proposed graph indicates each node at the statistical distance from the centerline. The size of the nodes varies according to the frequency of happening, i.e. the less frequently happening nodes diminish, the frequently happening node is enlarging, and finally, it indicates that connecting the larger nodes indicates the pattern of the plant. The colors of the nodes indicate the node type. The algorithm is represented by Figure 0-9.

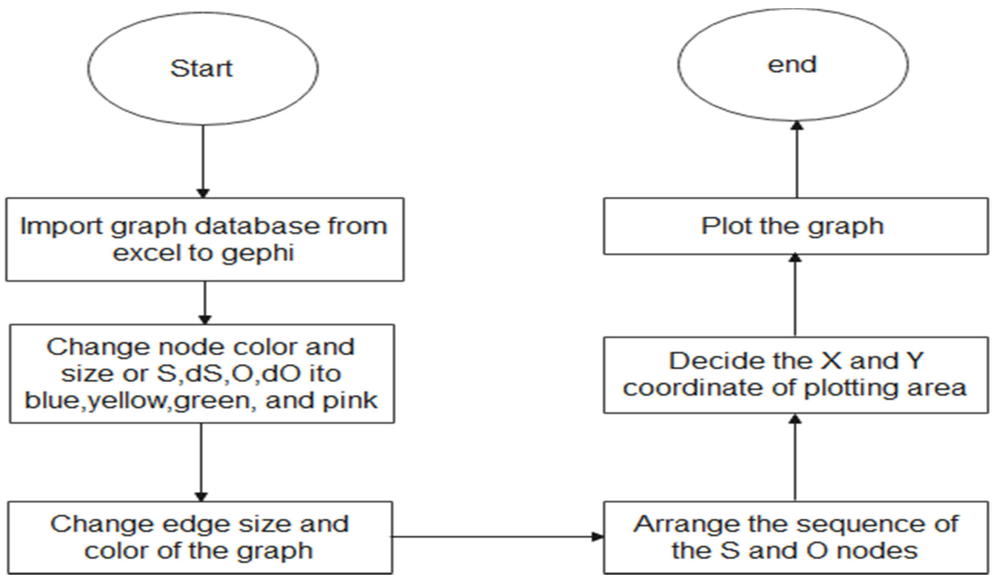

Figure 0-9 Steps for plotting the graph

4.5 Translate the algorithm into coding

The algorithm proposed in section 4.2 is translated into code using two Programming languages, which are Matlab and Python. Matlab is chosen for its suitability for matrix manipulation and designing of NN. The second part of the coding, which is Python, is used to analyze and visualize the graph network, which is going to be displayed by network displaying software, gephi. Gython scripting is used as coding in gephi software.

4.6 Study the algorithm with sample data

For modeling purposes, the ideal process which has 10 variables with 100 records is randomly generated to show the outcome of the Algorithm. From these 10 variables, 2 are input variables, 3 are controllable variables, 2 are uncontrollable variables, and 3 are output variables. The algorithm takes into consideration that the Input parameter (Dataset K) is stored in an Excel file in a proper format. Then follows the following steps

1. **Define the dataset K**

Communicate with the user, find the Excel data file, change to matrix form, and save to Matlab for further analysis.

Here in this script Matlab communicate with the user to ask about the number of input variables, controlled variables, uncontrolled variables, and output variables. Assuming that 100 records are saved for all variables by the user in Excel format. Matlab copies these records in matrix form and calls this matrix mat_values. These mat_values will be used throughout the analysis to define and forecast the relationship between the four groups of variables. Furthermore, to keep the original data two files are created, one is for the original file and the other is a copy file that we can modify and update during the program run. This file is saved in the same folder named "c:\\process-data". The original process data is given the name "c:\\process-data\O_processdata.xlsx" and the copy data will be given the name "c:\\process-data\X_processdata.xlsx". In order to not mix values of variables, titles are defined for both original and copy files. Then discretized Excel file is saved as shown in Table 0-1.

Table 0-1 Data generated for analysis

Seq_no	I1	I2	C1	C2	C3	U1	U2	O1	O2	O3
1	65	22	174	20	48	95	20	202	94	35
2	57	27	172	25	58	113	16	197	98	39
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
99	50	21	165	29	47	92	17	200	89	39
100	63	21	177	21	56	98	13	202	86	38

2. **Discretize the data**

The upper control limit and the lower control limit of each variable are decided, class size is fixed and each value of data is represented in discrete form, Table 0-2. This step is important because graph theory is based on discrete data types.

Table 0-2 Upper and lower control limits

	I1	I2	C1	C2	C3	U1	U2	O1	O2	O3
Lower control limit (LCL)	50	20	160	20	40	90	10	195	70	30
Upper control limit(UCL)	75	30	180	30	60	120	20	205	100	40
No of class	25	10	20	10	20	30	10	10	30	10

As a result, the discretized data will be as shown in Table 0-3.

Table 0-3 Discretized data

Seq_no	I1	I2	C1	C2	C3	U1	U2	O1	O2	O3
1	61	21	71	1	41	17	100	71	81	51
2	29	71	61	51	91	77	61	21	94	91
3	49	71	56	71	31	1	41	91	74	100
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
98	21	41	31	11	71	1	61	91	91	41
99	1	11	26	91	36	7	71	51	64	91
100	53	11	86	11	81	27	31	71	54	81

3. **Create the graph**

Based on the above discretized data, the following nodes table of the graph is developed, Table 0-4. Total of 397 nodes are created from the 100 discretized data.

Table 0-4 Nodes table

seq_no	Id	label	I1	I2	C1	C2	C3	U1	U2	O1	O2	O3	MD
1	1	s1	61	21	71	1	41	17	100				11.03718
2	2	o1								71	81	51	11.03718
3	3	s2	29	71	61	51	91	77	61				4.831258
4	4	ds1	-32	50	-10	50	50	60	-39				11.03718
5	5	o2								21	94	91	4.831258
6	6	do1								-50	13	40	11.03718
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
394	394	s99	53	11	86	11	81	27	31				8.883541
395	395	ds99	52	0	60	-80	45	20	-40				8.883541
396	396	o100								71	54	81	6.474286
397	397	do99								20	-10	-10	8.883541

Similarly, the edges will have the following data, Table 0-5.

Table 0-5. Edges table

seq_no	Source	Target	Type	Id	label	Interval	Weight
1	1	2	Directed	1	e (s1-o1)		1
2	1	4	Directed	2	e (s1-ds1)		1
3	4	3	Directed	3	e (ds1-s2)		1
.	.	.	.	.	.	.	.
.	.	.	.	.	.		.
.	.	.	.	.	.		.
493	395	394	Directed	493	e (ds99-s99)		1
494	392	397	Directed	494	e (o99-do99)		1
495	397	396	Directed	495	e (do99-o100)		1
496	394	396	Directed	496	e (s99-o100)		1

Finally, the graph will be plotted as shown in Figure 0-10. It is known that the graph plot has no defined and identified x and y coordinates. It simply shows the connection of nodes by edges. The length of edges and position of nodes doesn't have meaning. However, the plot shows the number of nodes, the size of nodes, and the presence of connections between nodes.

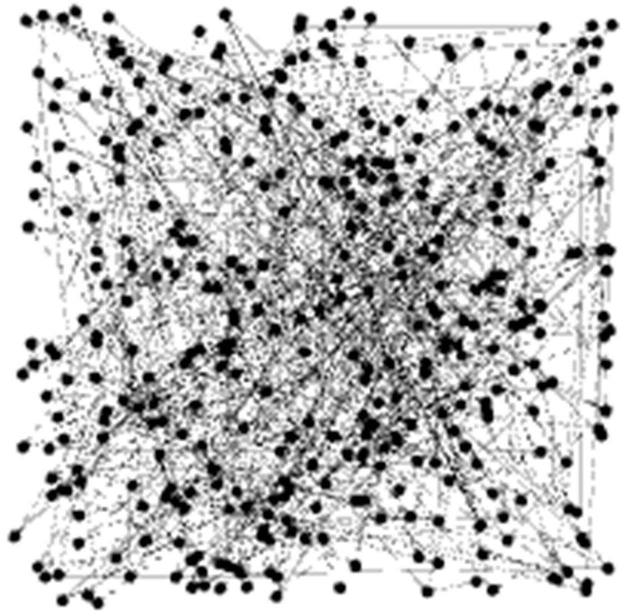

Figure 0-10:Gephi plots the graph

4. **Develop the Neural Network**

Data is continuously collected in the production system. The record collected during the production time predicts the condition of the process and it is important to make decisions to get smooth operation. The bulk database formed as a result of the continuous collection of data over a long period of time is the history of that production system. This database is used to predict the relationship between variables and helps to study the pattern of the plant. A neural network is one of the machine learning techniques that learn from the database and explores the relationship between variables.

In this research, two NN models are designed as shown in Figure 0-11. The first one, Deep Neural Network1 (DNN1), is used to predict the output of the system from the given input variables, controllable variables, and uncontrollable variables. The purpose of this NN is to monitor the output quality characteristics. The second one, Deep Neural Network 2(DNN2), is designed to predict the next move by deciding the input variables and controllable variables. The input to DNN2 is the desired output characteristic of the production system. To meet this objective, DNN2 proposes the required input and controllable variables. The uncontrollable variables are forecasted from the Multivariate Time Series (MTS) model and are given as input to the DNN2. The detail of the two DNNs is discussed in the next paragraphs.

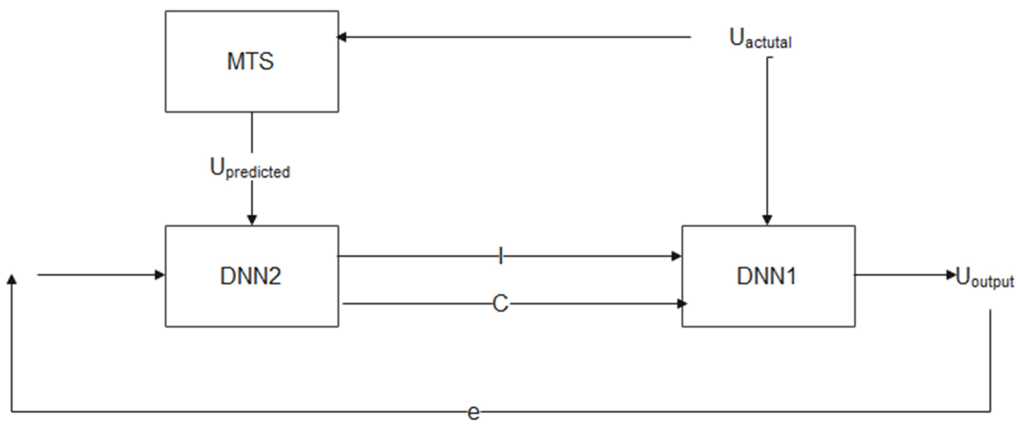

Figure 0-11:Framework to display the connection between DNNs and MTS

DNN1 is constructed in such a way that the neural network consists of dense layers for each input vector that are connected then concatenate these signals into one layer, and then two hidden layers and one output layer are followed. The nodes in each layer are fully connected to the nodes in the next layer. The number of nodes for each layer is decided between 15 to 100 nodes using search algorithms. The number of trainable parameters comes to be 49,385 i.e., the number of weights and biased constructs in the neural network is 49,385. Sufficient Training samples that contain input signals and the corresponding output values are required to train the Neural Network. Then, training samples are divided into three groups, training, testing, and validation data. The ReLU activation function is used in the nodes of hidden layers and the Softmax activation function is used in the nodes of the output layer. Mean Squared Error (MSE) is used to minimize the error between predicted and actual errors during the forward and backward flow of signals in the Neural Network and finally, the optimized weight and bias coefficients are determined, Figure 0-12.

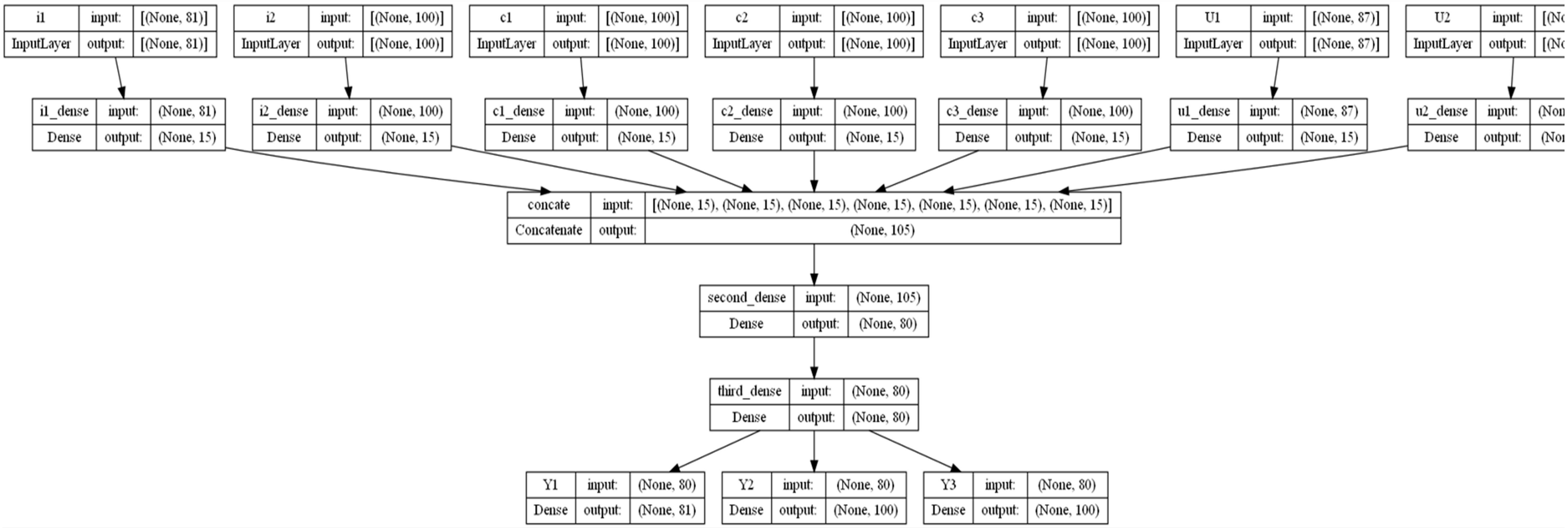

Figure 0-12: Model for DNN1

The performance of the DNN1 Neural Network is predicted by the loss function. The loss functions of the Y1, Y2, and Y3 output variables are found to be below 2.7% which is the acceptable range as shown Figure 0-13.

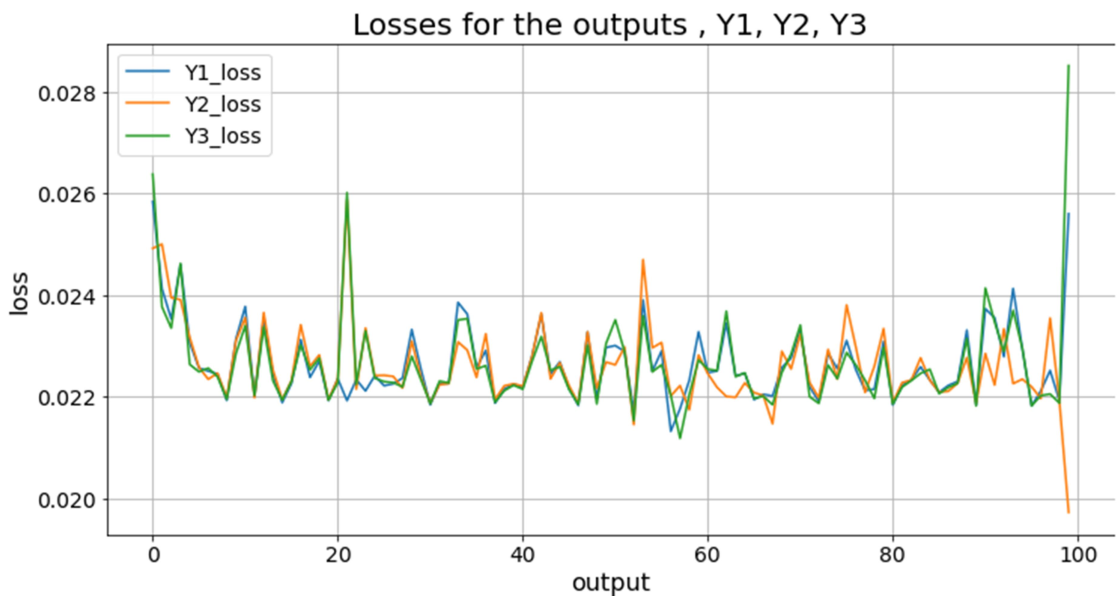

Figure 0-13: Losses for the outputs

In the same way, the deep neural network2 (DNN2) decides what should be the next move to get the desired output. Here the input is the desired output characteristics of the product and the values of the uncontrolled variables U.

The U values are forecasted from the MTS model. U in this case is composed of only two variables, U1 and U2, which is the simplest case of multivariate time series. Vector-Auto-Regression-Moving-Average (VARMA) with auto_ARIMA prediction of individual variables are used to predict the next values. Auto_ARIMA decides the values of p, d, and q for each U1 and U2 variable, where p is a parameter for Auto-Regression, d is the difference to make it stationery and q is a parameter for Moving-Average.

The Auto-ARIMA values for U1 are found from iteration and select the minimum AIC values and it is found to be (4,1,1). The AR parameters are [-0.03295221 -0.00982854 -0.01855821 -0.07053032], the MA parameter is [-0.99211253], and the minimum test RMSE is 9.507. For the window size of 50, the plot of the last window is shown in Figure 0-14.

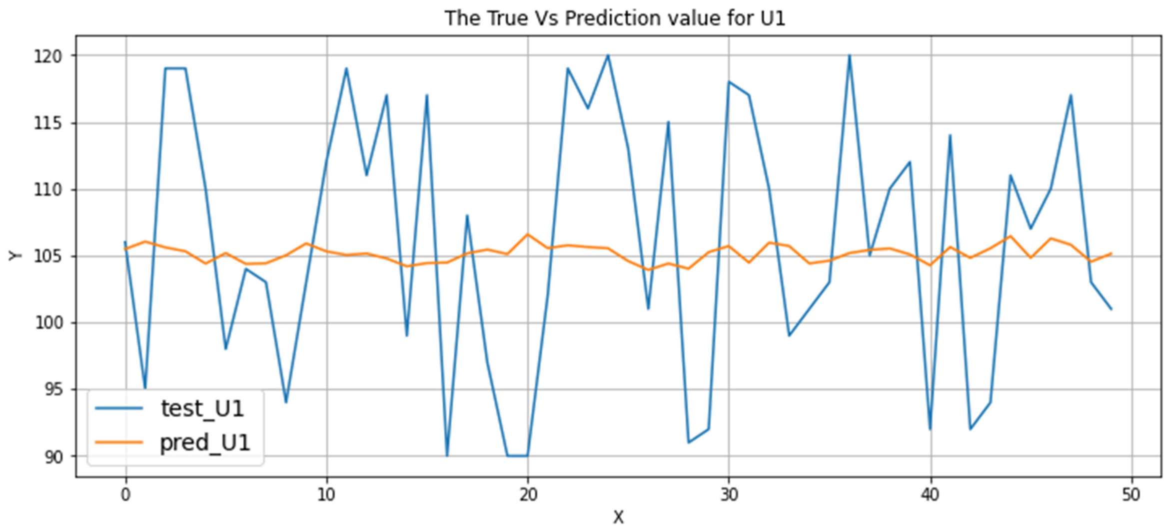

Figure 0-14:The true Vs predicted values for U1

Similarly, the Auto-ARIMA values for U2 are found from iteration and select the minimum AIC values and it is found to be (5,1,0). The AR parameters are [-0.90795008 -0.74053796 -0.55120317 -0.291019 -0.15919071], the MA parameter is nill because q parameter is zero, and the minimum test RMSE: 3.365. For the window size of 50, the plot of the last window is shown in the Figure 0-15.

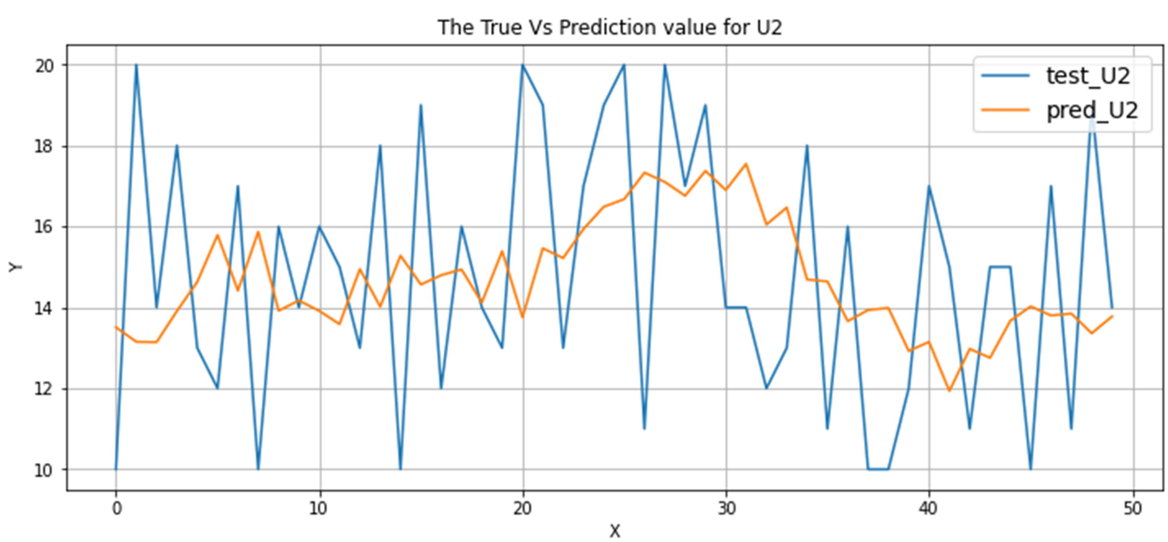

Figure 0-15: The true Vs predicted values for U2

Figure 0-16 shows the U values for a window size of 100 over the time sequence. From this series, the next value of u1 and u2 is to be forecasted from the MTS model.

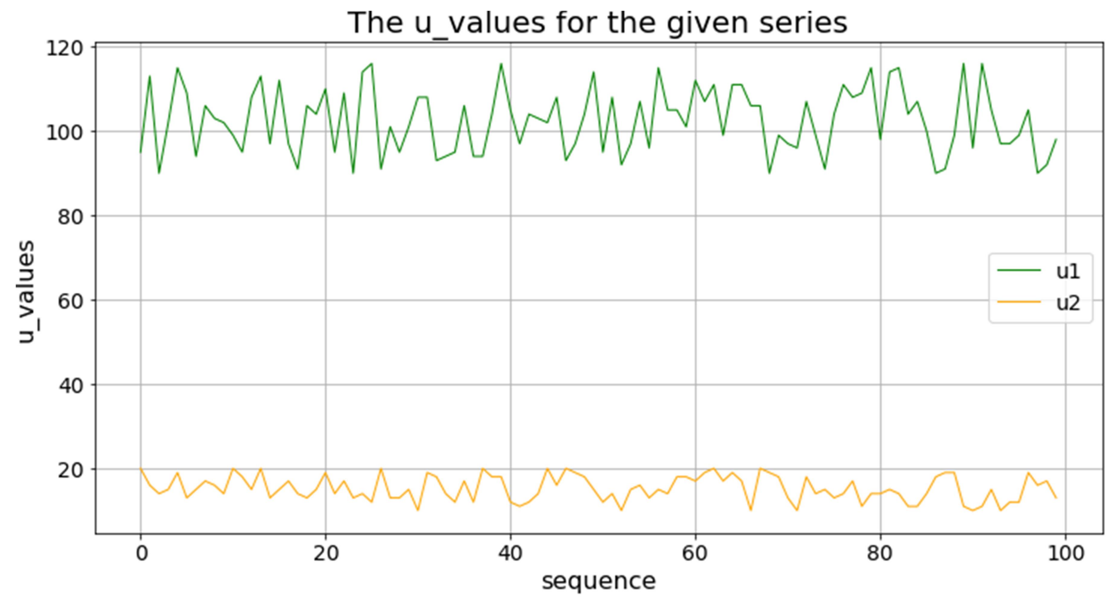

Figure 0-16: U1 and U2 values in the sequence

Now, the u values are forecasted and the desired output variables are set, then DNN2 will decide what should be the input(I) and control (C) variables. The output of the DNN2 model is the values of the input variable and the necessary change in the controlled variables. The model is simplified and represented by Figure 0-17.

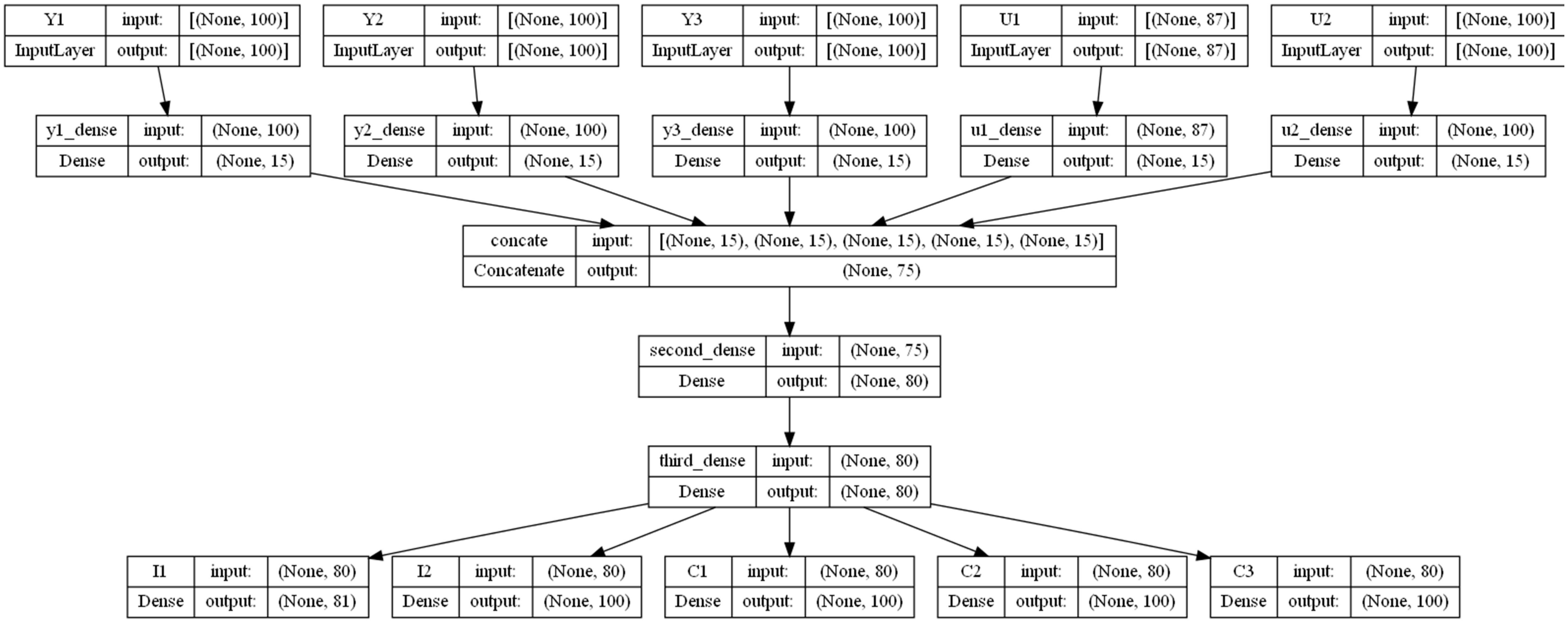

Figure 0-17: Model to DNN2

DNN2 is the deep neural network designed from the two inputs (U and desired output O) and the two outputs (input I and controllable variables U). DNN2 is important to decide the next decision point for the desired output of the system. It decides the values of I and C. Finally, the system needs to minimize the error term e between O output and O Desired. The degree of the error term e is determined from the GANNT control chart displayed in Figure 0-18.

5. **Display the result using the graph displaying application**

In the graphical representation, the nodes in the chart are distinguished by different colors, namely pink, green, yellow, and blue, which correspond to the dS, S, dO, and O nodes, respectively. Each node is strategically positioned in the chart based on its statistical distance from the center line, indicating its relationship to the desired control limits. This positioning allows for a visual assessment of the node's proximity to the control limits and provides valuable insights into the process performance.

Furthermore, the size of each node in the graph varies in accordance with its weight, which is determined by the frequency of its occurrence. Nodes with higher weights, indicating more frequent occurrences, will be depicted with larger sizes, while nodes with lower weights will appear smaller. This variation in node size serves as a visual representation of the relative importance or impact of each node within the process.

Combining these elements, the graph provides a comprehensive and intuitive visualization of the statistical process control data. It allows observers to easily identify and analyze the distribution of nodes, their proximity to the control limits, and their respective weights. This graphical representation enhances the understanding of the process behavior, facilitating effective decision-making and action planning to maintain the desired control and improve overall process performance.

The final display of the graph is as follows in Figure 0-18, showcasing the interplay of colors, node positions, and varying sizes to present a clear and informative representation of the statistical process control data.

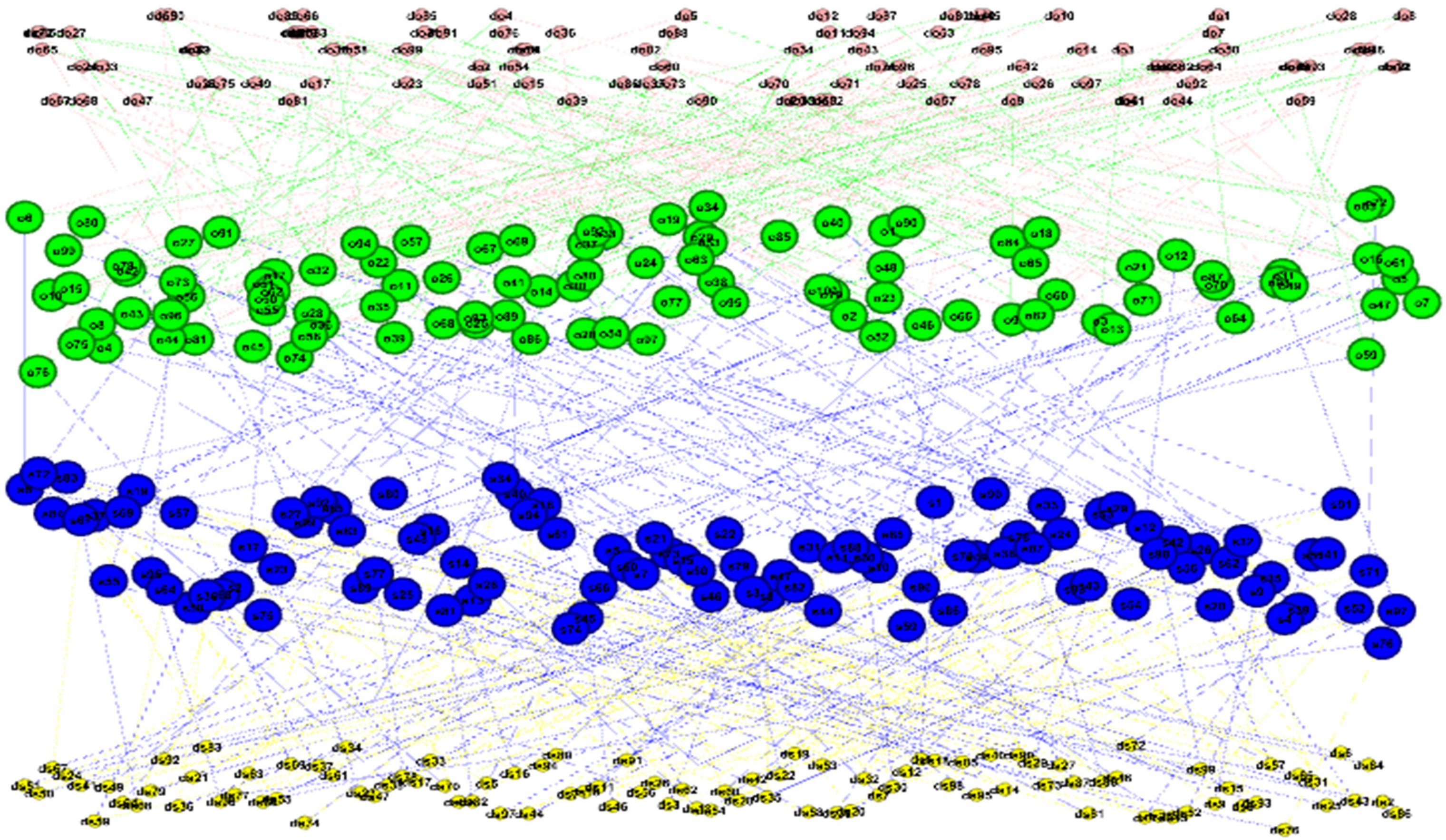

Figure 0-18: GANNT chart for the process in the condensed form

4.7 Summary of the Chapter

Statistical Process Control (SPC) plays a pivotal role in reaching desired levels of product quality by closely monitoring and controlling process and product variables. Multivariate statistical control techniques have played an instrumental role in recent advances, leading to improved production efficiency and performance. Progress has been enabled by data collection devices and processors capable of processing large volumes of information, thus revolutionizing traditional statistical control methods into more sophisticated machine-learning approaches. As a result, product quality has shown tremendous improvements and companies increasingly recognize advanced process control methodologies as key competitive advantages.

Traditional multivariate statistical control techniques have undergone rigorous statistical validation, but their primary purpose remains alerting operators of current operations status as well as any possible out-of-control signals that might emerge soon. However, their greatest shortcoming lies in their inability to pinpoint the source of process variation that exceeds acceptable limits. Therefore, these techniques do not offer operators clear guidance for efficiently and quickly returning the process back to its normal operating state. Current statistical methodologies provide tools that enable prediction of possible outcomes and investigation into future events; however, research which integrates such techniques with advanced machine-learning methods remains scarce.

In order to address the limitations associated with traditional multivariate statistical control techniques, a model was devised that not only investigates relationships among variables but also considers their simultaneous decisions and any subsequent ramifications of operation sequences. This model integrates graph theory and artificial neural networks to address some of the gaps encountered when employing traditional approaches. Graph theory is used to dissect intricate relationships among operation sequences, while Deep Neural Networks intelligently suggest the best next step for reaching desired operational results. When integrated into existing Hotelling T2 models, both techniques allow the Hotelling T2 model to control and predict operations efficiently.

This research seeks to advance the Hotelling T2 technique, an established multivariate statistical control method. Through developing a novel technique which not only predicts anomalies but also provides guidance for corrective actions while taking into account sequence of operations, its limitations can be addressed. Furthermore, its model categorizes variables into distinct groups - input, output, controllable and uncontrollable variables with VARIMA being utilized to capture multivariate time series data dynamics.

This model takes an integrated process control approach, enabling an assessment of all four variables within one control chart - effectively serving as an executive summary of operations for any given window size. Not only can it display process trends like traditional multivariate control charts do, but also predicts next best steps with minimal error. To meet this objective, graph theory and neural networks have proven an ideal combination. Both graph theory and neural networks possess similar building blocks; nodes and edges for graphs and neurons and connections for neural networks respectively. Graph theory examines the order of operations while neural networks act as memory storage units by recalling decisions previously made by an operating system. When combined together, graph theory and neural networks serve a common goal while fulfilling distinct tasks; data collection over an extended period increases synchronization while Hotelling T2 analyzes variance/covariance relationships of variables to produce results plotted on Hotelling control charts.

CHAPTER FIVE

MODEL IMPLEMENTATION

In this chapter, the model developed from ideal process in chapter four is implemented to real cement industry. The cement production process is discussed and the quality parameters are identified. Then the model is trained and tested based on the data collected from the case company.

5.1 Introduction

There are similarities between all manufacturing industries. Hence, the model developed in Chapter 4 is intended to be implemented for any manufacturing industry. They all need raw material inputs with the desired specification range. The raw materials are processed into some intermediate or final product by different operations. These operations are monitored by a specific range of process parameters set based on the target quality constraints of the final product. The production system is affected by the external environment to some extent; hence it needs to identify these variables and adjusts the process accordingly. The model is designed to incorporate all these parameters.

The purpose of this chapter is to implement the model to a specific industry and as a result, show how MSPC is important to improve the quality of the product to the identified sector. Here in this research, the cement industry is selected as a case to demonstrate the implementation of the model. A preliminary survey discloses that one of the challenges in Ethiopian cement Industries is the lack of technical tools to monitor and control processes. It is hard to find research that identifies this problem and gives technical support. It is observed that the quality control methods in the Ethiopian cement industry are dependent on the seven statistical quality control tools. The statistical process control tool is the Shewhart control chart. Shewhart's control chart is univariate control chart. These methods are important to run the process, but the current technology is now demanding online and simultaneous analysis of many variables. The global trend urges industries to train their process control engineers to advance their production process control methods. The significance of this research is in line with the interest in introducing a new concept to industrial process control techniques.

The current trend in the cement industry shows that cement quality is highly dependent on the skill and knowledge of the operators and process engineers. The effort of operators and process engineers to monitor and control process disturbance is appreciated. However, any process by its nature is Multivariate. The skill and experience of operators can never solve multivariate problems because human capability has limitations to analyze many problems simultaneously. Unfortunately, multivariate process control techniques are not introduced and exercised in Ethiopian industries. As a result, the process can be interrupted unexpectedly, the production plan fails, the quality cost is high and as a consequence the production cost is high and profit is reduced. Industries are less competitive in the market by cost and quality. As a result, this research has proposed the application of multivariate process control techniques.

The strategy of implementing the model to the case company is that there is already a statistical process control technique applied to the cement industry. Hence, there is known process variables and data collection system and for each process variable, there is an upper and lower control limit that needs to be kept. Now, the model uses the data collected and the control limit as input to its MSPC chart.

5.2 Overview of Cement Production Technology

Some "Natural rocks" contained all ingredients like CaO, $Si0_2$, Al_20_3, and Fe_20_3 in approximately the right proportions so that they did not need any additions, and when ground calcined, and sintered in a kiln produced clinker which when ground with 5% gypsum produced what has come to be known as "Portland Cement". S.P.Deolalkar

Cement is a hydraulic binder, i.e. a finely ground inorganic material that, when mixed with water, forms a paste that sets and hardens using hydration reactions and processes and which, after hardening, retains its strength and stability even under water (EN197-1). Cement is also defined as a man-made gray powder that consists essentially of compounds of lime and silica, with a smaller amount of iron and alumina, when proportionally mixed with water and aggregates, produces concrete. The cement-producing process by its nature is complex. This is due to a large number of factors involved in the production system, which are interrelated and influenced by each other. The process must smoothly flow starting from raw material selection until the delivery of the product to the customer. The production process passes several stages with the intermediate product. The output of each stage is the input for the next stage; hence, the quality of the final product and the performance of the whole production process is dependent on every stage. The Ordinary Portland Cement(OPC) has composition represented by Table 0-1.

Table 0-1 . Target chemical composition of OPC

No	*Chemical*	*Composition*
1	CaO	65-68%
2	SiO_2	20-23%
3	Al_2O_3	4-6%
4	Fe_2O_3	2-4%
5	MgO	1-5%
6	Mn_2O_3	0.1-3%
7	TiO_2	0.1-2%
8	K_2O	0.1-1%
9	Na_2O	0.1-0.5%

In general, the cement production process can be classified into five stages. These are raw material preparation, raw mix and milling, clinkerization, cement milling, and cement packing subsystems, Figure 0-1. Each subsystem consists of a continuous process that transforms raw materials input into the intermediate or final product.

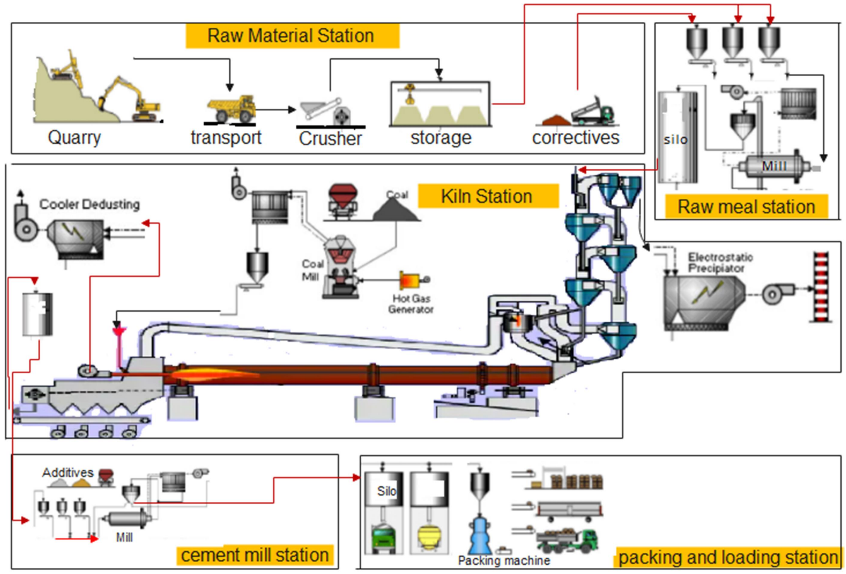

Figure 0-1:Cement plant process flow

5.2.1 Raw Material Preparation

The major chemicals needed for the production of cement are calcium oxide, silica, iron oxide, and aluminum oxide. These chemicals are extracted from different types of cement rocks such as ***Calcareous materials*** (containing calcium carbonates) such as limestone, chalk, marl, and marble; ***Argillaceous materials*** (containing alumina and silica) such as clay, shale, and slate; ***Siliceous materials*** (containing silica) such as sand and sandstone.

There is a desired proportion of these chemicals for the production of cement. Depending on the content of these rocks 2, 3, 4, or 5 component mixes can comprise the proportions. Hence, some cement plants are designed for a four-component mix and some other cement plants can be designed for a five-component mix.

The major activity to be done at this stage is size reduction or crushing of raw materials, transport to the raw material storage sites, and then mixing and storing to their identified location or station. Limestone is crushed to 15-20mm (max of 25mm) size, which is suitable to be transported to the main plant site and convenient for raw mill feed. The process of raw material preparations is discussed below.

Quarrying: -Quarrying is extracting raw material from mineral-containing sites. The most common raw materials used for cement production are limestone, sand, and clay. The major component of the raw material is limestone usually extracted from the quarry adjacent to or very close to the plant. Limestone provides calcium oxide for cement production while sand, clay, shale, and other material provide most of the silicon, aluminum, and iron oxide. The raw material extracted from the quarry site is then sent to a crusher to reduce the size to about 25mm.

5.2.2 Raw mix and milling

Here in this stage, raw materials are mixed proportionally depending on the composition of their chemicals and finely ground by the raw mill then transport to the raw meal silo. Raw mix design is based on the following parameters: LSF (Lime saturation factor), SM (Silica Module), and AM (alumina Module). These parameters affect fuel consumption, burn ability, clinker nodule formation, and early strength of cement and ensure that the finished product meets the specification. Specific power consumption, fineness, moisture content of the raw meal, inlet, and outlet gas temperature, and outlet gas flow of the mill are important parameters to be controlled in this station.

The Lime saturation factor (LSF) is the ratio of the actual amount of lime to the theoretical lime required by the other major oxides in the raw mix or clinker.

$$LSF = \frac{100(CaO+0.75\ MgO)}{2.85SiO2+1.18Al2O3+0.65Fe2O3} \qquad (1)$$

Typical LSF values in modern clinkers are 0.92-0.98 or 92%-98%. If it is above 100 % it implies that there is free lime not combined with other chemicals. The free lime target is about 0.5-1.5% CaO-free.

Silica Module (SM) also known as Silica ratio (SR) is defined as the ratio of silica to alumina and iron oxides.

$$SM = \frac{SiO2}{Al_2O_3+Fe_2O_3} \qquad (2)$$

SR is typically between 2.0 and 3.0.

When this method is used as a calculation of a raw mix for three raw materials, the required LSF is 0.92 and the required SR is 2.70.

The amount of melt phase in the burning zone is a function of SM. When SM is high, the amount of melt is low, and vice versa. When the SM is too high, the formation of nodules and the chemical reactions may be too slow making it difficult to operate. The high the SM the harder it is to burn. When SM is too low there may be too much melt phase and the sulfur coating can become too thick.

Alumina Ratio(AR) is defined as the ratio of alumina to iron oxide.

$$AR = \frac{Al_2O_3}{Fe_2O_3} \qquad (3)$$

The temperature by which the melt form depends on the AM. The lowest temperature is obtained when the AM is approximately 1.6, which is optimum for the formation of the clinker minerals and nodule formation. The AM also affects the color of clinker and cement. The higher the AM the lighter the color of the cement.

Proportionally mixing and milling raw material: Raw materials are mixed in proportion according to the desired chemical composition of the mix. Then it will be finely grounded into powder which is called raw meal.

Storing and homogenizing raw meal:

Storing and homogenizing the raw meal in the raw meal silo is the major operation of this station. Differential pneumatic blending is carried out in blending silos and the blended raw meal is stored in storage silos.

5.2.3 Clinker formation

In this stage, raw meal is burned into clinker. Burning (Pyro processing) is the most important operation in the manufacturing of cement because (i) fuel consumption is the major expense in the process. (ii) Capacity of a plant is measured by kiln output and (iii) strength and other properties of cement depend on the quality of clinker produced.

Kiln, cooler, preheater, and pre-calciner are the major equipment in the production of clinker. The kiln system can be considered a chemical reactor in which flue gas and raw material flow in opposite directions transfer heat and change raw material to clinker.

The general overview of the transformation process is presented in Figure 0-2.

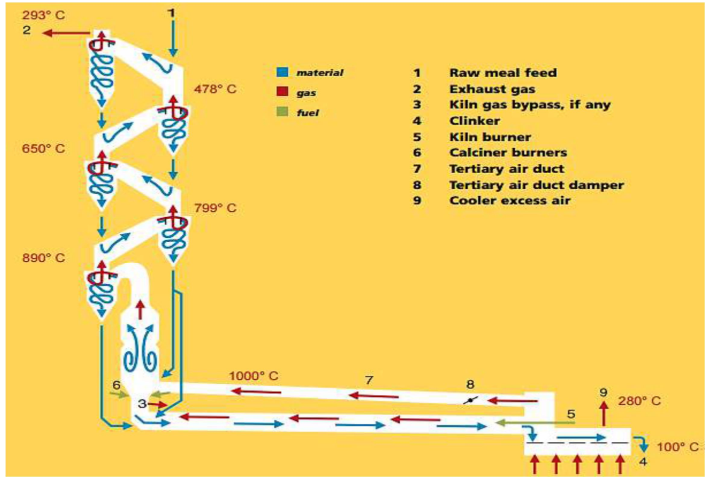

Figure 0-2: Six-stage preheater and kiln for the production of clinker (FLS Smidth,2010)

Table 0-2. Material and gas flow of cement production process

Material flow side		***Gas flow side***	
1	Evaporating free water, at temperatures up to 100°C	1	Ambient air preheated by hot clinker from kiln 20°C up to 600°C to 1100°C
2	Removal of adsorbed water in clay materials 100°C - 300°C	2	Fuel burns in preheated combustion air in a kiln from 2000°C to 2400°C
3	Removal of chemically bound water 450°C - 900°C	3	Combustion gases and excess air travel along the kiln, transferring heat to kiln charge and kiln refractories 2400°C down to 1000°C
4	Calcination of carbonate materials 700°C - 850°C	4	Preheating system for further recovery of heat from kiln gases into the material charge in the kiln system 1000°C down to 350°C to 100°C
5	Formation of C_2S, aluminates and ferrites 800°C - 1250°C	5	Further heat recovery from gases for drying of raw materials or coal
6	Formation of liquid-phase melt >1250°C		
7	Formation of C_3S 1330°C - 1450°C		
8	Cooling of clinker to solidify liquid phase 1300°C - 1240°C		
9	Final clinker microstructure froze in clinker <1200°C		
10	Clinker cooled in cooler 1250°C - 100°C		

The raw meal is burned inside the kiln to produce clinker. Clinker is made in the kiln at temperatures of 1,450°C. This hot clinker is cooled by a cooler to a temperature about a little higher than the ambient temperature. Table 0-2 represent overall heat exchange between the material and flue gas.

Storing the clinker:

Clinker should be stored in such a way that it does not contaminate with moisture otherwise it will make concrete.

All kiln systems aspire to optimize heat exchange between the gas streams and material streams at various stages to minimize waste heat and maximize thermal efficiency. The following table summarizes the theoretical heat balance in the kiln system

Table 0-3. Energy consumption for endothermic and exothermic reactions

Endothermic Reaction	KJ/Kg	Kcal/Kg
Dehydration of clays	170	40
Calcination	1990	475
Heat of Melting	105	25
Heating Raw materials	2050	490
Sub Total	**4315**	**1050**
Exothermic reactions	**KJ/Kg**	**Kcal/Kg**
Crystallization of dehydrated Clay	40	10
Hear of formation, clnker minerals	420	100
Crystallization of melt	105	25
Cooling of clinker	1400	335
Cooling of CO_2	500	120
Cooling of water	85	20
Sub Total	**2550**	**610**
Net heat for clinker formation (Endothermic heat -exothermic heat)	**1765**	**420**

Add inefficiencies, indicated in Table 0-3, to Table 0-4 will give the complete picture of energy flow in cement production.

Table 0-4. Energy loss related to inefficiencies in the production

Heat losses related to inefficiencies during the kiln operation	KJ/Kg		Kcal/Kg	
	Dry process	**Wet process**	**Dry process**	**Wet process**
Evaporation of water	20	2100	5	502
Heat losses, gas, clinker, dust	840	1250	201	299
Radiation, convection loss	650	360	156	86
Total heat consumption= (endothermic heat-exothermic heat) + heat losses due to	**3275**	**5475**	**782**	**1307**

inefficiencies				

The kiln operation is smooth if the following parameters are controlled

Kiln feed chemical composition, Kiln feed chemical uniformity, Kiln feed fineness, Kiln feed rate, fuel heating value, fuel fineness and volatile content, fuel feed rate, Clinker cooler operation.

The distribution of main minerals in a neat Portland cement (i.e without any other mineral additions that calcium sulfate) may typically be of the order 60% C3S, 20% C2S, 10%C4AF, 5% C3A, and 5% CSH2.

Bogue Calculation is useful formula to calculate the C3S, C2S, C3A, and C4AF concentration in the clinker

1. C3S = $4.0710CaO-7.6024SiO_2-1.4297Fe_2O_3-6.7187Al_2O_3$C2S = $8.6024SiO_2+1.0785Fe_2O_3+5.0683Al_2O_3-3.0710CaO$
2. C3A = $2.6504Al_2O_3-1.6920Fe_2O_3$
3. C4AF = $3.0432Fe_2O_3$

5.2.4 Cement milling

Cement milling is the last stage of making cement. These are Ordinary Portland Cement (OPC) which is the mixture of clinker and gypsum and Pozzolanic Portland Cement (PPC) which is the mixture of clinker, gypsum ($CaSO_4.2H_2O$), and Pozzolanic material like pumice.

The proportioned mixture is fed to a ball mill, then finely ground. The milled cement is sucked by the high-pressure exhaust fan and transported to the classifier. Here, the course is separated from the fine. And the course returned to the mill. The fine powder is further separated from the air by cyclones and air bag filters, and the final product is transported to a cement silo by a bucket elevator. Such a type of milling system is called a closed-circuit mill system. The circuit is represented in Figure 5.3.

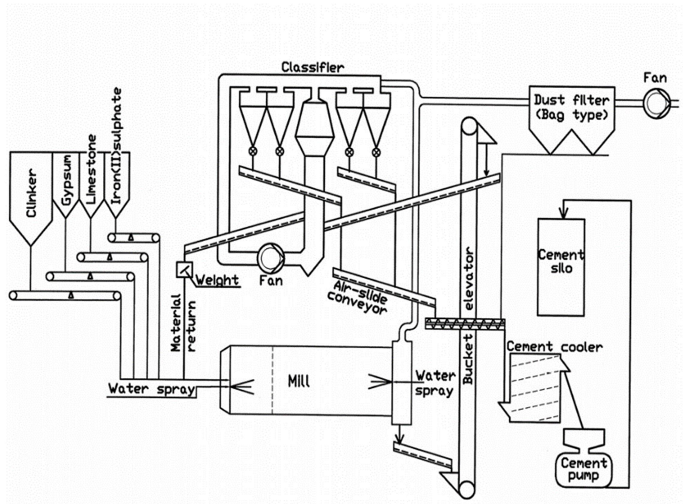

Figure 0-3: Cement mill circulation diagram

Milling operation is monitored by controlling the following parameters:

Proportionally mixing the clinker and additives and milling the mixture: Clinker is then ground with gypsum and pumice to produce the powder known as cement.

Storing the cement: milled cement is stored in the silos keeping the moisture suitable for cement and mixing and homogenizing operation takes place.

5.2.5 Packing and loading the cement

Delivering cement to the customer on the plant site takes place either in packed bags or in the bulk load. In the case of packed cement, the cement is extracted from cement silos through a rotary feeder and after passing through a rotary screen, it is packed in bags by packing machines. In the case of bulk loading, the cement is extracted from the rotary feeder and directly feed to bulk-loading trucks.

5.3 Process Control in Cement Production

The offline laboratory is the traditional quality control setup to cement production systems. The process control laboratory, analytical laboratory, instrumental laboratory, and physical laboratory measure the quality parameters and maintain the cement quality within the allowable quality range. The process control lab measures the Free CaO, sulphate concentration, pozzolanicity test, and the determination of fineness by the air permeability method.

The analytical laboratory determines the complete analysis of raw materials, Clinker, and cement using analytical methods. The determination of the chemical composition of matter is important for the design of cement composition. Complete analysis of raw material determines the percentage composition of Los on Ignition(LOI), SiO2, Al2O3, Fe2O3, CaO, MgO, SO3, K2O, Na2O , TiO2, Cr2O3, Mn2O3, P2O5 ,Cl and F. Complete analysis on clinker determines the insoluble residue and chemical composition of clinker. In general, the Clinker quality parameters are Line Saturation Factor (LSF), Silica Module (SM), Alumina Module (AM), dicalcium silicate or belite (C2S) concentration, tricalcium silicate or alite (C3S) concentration, tricalcium aluminate (C3A) concentration, and tetracalcium aluminoferrite(C4AF) concentration. Determination of chloride is performed in this laboratory.

The instrumental laboratory is another important laboratory that measures Na2O and k2O concentration.

Lastly, we have a physical laboratory that measures the fineness, setting time, soundness, strength, and hydration of cement. The fineness of raw meals and cement is determined in this laboratory.

Two classes of early strength are included for each class of standard strength, a class with ordinary early strength, indicated by N, and a class with high early strength, indicated by R (see Table 0-5).

Table 0-5. Strength class for standard cement

Strength class	**Compressive strength MPa**				**Initial setting**	**Soundness**
	Early strength		**Standard strength**		**Time (min)**	**(Expansion) mm)**
	2 days	**7 days**	**28 days**			
32,5 N	**-**	**≥ 16.0**	**≥ 32.5**	**≤ 52.5**	**≥ 75**	
32,5 R	**≥ 10.0**	**-**				
42,5 N	**≥ 10.0**	**-**	**≥ 42.5**	**≤ 62.5**	**≥ 60**	

42,5 R	**≥ 20.0**	**-**				
52,5 N	**≥ 20.0**	**-**	**≥ 52.5**	**-**	**≥ 45**	
52,5 R	**≥ 30.0**	**-**				**≤ 10**

Cement is produced in a continuous fashion and under different constraints. Raw materials have to be proportionally mixed; clinker quality parameters must be kept in the standard; the final product, i.e. cement, should meet the quality requirements; electrical and thermal energy consumption should be optimized; equipment life is to be prolonged, and environmental pollution should be minimized. The research aimed to review process control literature and investigate how far process control techniques are applied to the cement industry, and then further scrutinize the gap the process control techniques have not fulfilled the demand of the cement industry. Several statistical process control methods are developed since the introduction of process control techniques in the 1920s. The promising techniques that simultaneously search for optimal solutions are conducted by MSPC, data mining, and machine learning. However, needs further research in this sector.

5.4 General overview of the case company

The case company is an integrated cement plant located at a distance of about 50 km from Addis Ababa, Ethiopia. This five-stage dry kiln system plant has a production capacity of 1.2MTPD clinker. Two types of cement are produced in the plant. These are Ordinary Portland Cement (OPC) and Pozzolanic Portland Cement (PPC). The major chemicals needed for the production of cement, which are calcium oxide, silica, iron oxide, and aluminum oxide, are extracted from limestone, clay, and sand. Coal imported from South Africa is mixed with local coal for fuel consumption. However, based on the direction of the government, the plant is planning to substitute imported coal with local coal.

The Cement production system for the case company has five stages. These are raw material preparation, raw meal preparation, Clinkerization, Cement milling and finally packing and Delivering. Each stage has its process quality parameters, which affect the next stage.

In the first stage, rocks are exploded in the quarries, transported to the crusher, and their lump size is reduced to about 25mm. Then it will be transported from the crusher site to the storage area in the plant location. In the second stage, raw materials are mixed proportionally depending on the content of their chemicals and finely ground by the raw mill then transport to the raw meal silo. Raw mix design is based on the following parameters: LSF (Lime saturation factor), SM (Silica Module), and AM (alumina Module). These parameters affect fuel consumption, burn ability, clinker nodule formation, and early compressive strength of cement and ensure that the finished product meets the specification. Specific power consumption, fineness, moisture content of the raw meal, inlet, and outlet gas temperature, and outlet gas flow of the mill are important parameters to be controlled in this station.

The third stage is the clinker-forming stage. In this stage, the raw meal is burnt into clinker. The suspension preheater, Calciner, Rotary Kiln, and grate cooler are the major equipment in the kiln system. The kiln system can be considered a chemical reactor in which flue gas and raw material flow in opposite directions to transfer heat and change raw material to clinker. Burning (Pyro processing) is the most important operation in the manufacturing of cement because (i) fuel consumption is the major expense in the process. (ii) Capacity of a plant is measured by kiln output and (iii) strength and other properties of cement depend on the quality of clinker produced.

Clinker is the composition of four chemicals. These are dicalcium silicate or belite (C2S), tricalcium silicate or alelite(C3S), tricalcium aluminate (C3A), and tetetra calciumluminoferrite(C4AF). The proportion of these chemicals in the clinker determines the quality of the clinker. The other important parameters that need control and monitoring in the production operation are Free lime (F-CaO), liter weight, liquid phase, insoluble residue (IR), coating indexes, burnability indexes, Gas analyzer results, preheater gas temperature at 1st and 5th stage, and other equipment capacities indicators like amperages and temperatures.

Cement milling is the fourth stage of cement making process. According to the product type, clinker is mixed in different proportions with gypsum and additives, like pumice and limestone. Then the mixture is finely grounded by a ball mill. Strength, Setting time, soundness, fineness, and SO3 content are important cement quality parameters. Cement packing and delivering is the last stage where the product is delivered to the customer. Here cement is transferred from the silos either directly into bulk load or to a bagging station.

5.5 Applying the algorithm to the kiln system

The process control model developed in chapter four uses 10 ideal variables which are composed of 2 input variables, 3 controllable variables, 2 uncontrollable variables, and 3 output variables. Now, the model is implemented in the real industrial setup in the cement clinker section. This model needs four groups of variables. These are input variables, controllable variables, uncontrollable variables, and output variables. These variables are identified by senior process engineers from the case company and listed in Table 0-6 and represented by Figure 0-4.

Three-month process data is collected from the case company. The model needs the data collected simultaneously at a specified interval. However, data are mostly collected by the company using offline laboratory analysis or online analyzers. There is a measurement delay associated with these methods. Offline laboratory analysis may need 4hrs to generate results and online analyzers can show results within a few minutes. Hence, the model assumes the result of offline laboratory analysis is consistent with the measurement period; hence the requirement of simultaneous measurement of all variables at that specific period in the model is achieved. As a result, the right format for the data matrix K is obtained.

Table 0-6. Important variables in the kiln system

Input variables		***Uncontrollable variables***		***Controllable variables***		***Output variables***	
I1	*Kiln feed*	*U1*	*% CaO of raw meal*	*C1*	*LSF*	*O1*	Lt Wt
I2	*Local Coal feed*	*U2*	*% SiO_2 of raw meal*	*C2*	*SM*	*O2*	*Free CaO*
I3	*SA coal feed*	*U3*	*% Al_2O_3 of raw meal*	*C3*	*AM*	*O3*	*C_3S*
		U4	*% Fe_2O_3 of raw meal*			*O4*	*C_2S*
		U5	*% MgO of raw meal*			*O5*	C_3A
		U6	*% by mass of SO_3*			*O6*	C_4AF
		U7	*Loss on Ignition(LOI)*			*O7*	*Liquid Phase*
						O8	*B.I (Burnability Index)*
						O9	*C.I (coating index)*

To discretize the data, the upper control limit (UCL) and lower control limit (LCL) of each variable are to be specified, shown by Table 0-7. Then, these boundaries are divided into classes and each data is assigned to the respective class.

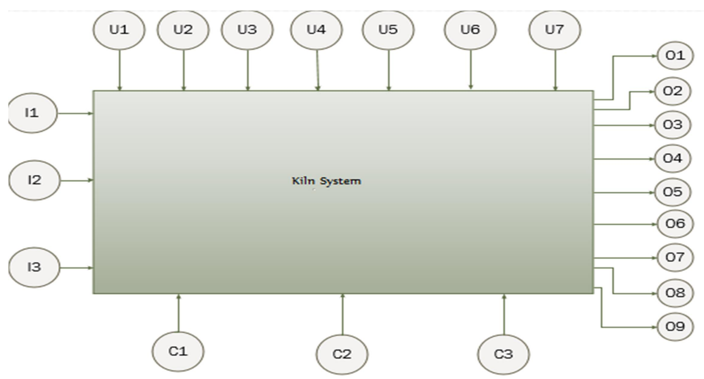

Figure 0-4: Model for kiln system

Table 0-7. UCL and LCL boundaries of kiln variables

Input variables			***Uncontrollable variables***			***Controllable variables***			***Output variables***		
	LCL	*UCL*		*LCL*	*UCL*		*LCL*	*UCL*		*LCL*	*UCL*
I1	90	105	*U1*	64	67	*C1*	90	105	*O1*	1200	1400
I2	8	12	*U2*	19	22	*C2*	2	3	*O2*	0.5	3
I3	12	18	*U3*	4	7	*C3*	1	2	*O3*	45	65
			U4	2	4				*O4*	3	21
			U5	1	3				*O5*	7	11
			U6	0	2				*O6*	8	13
			U7	0	1				*O7*	24	31
									O8	2	4
									O9	23	35

The data collected in the plant is now split into three sets. These are the train, test, and validated data. The train and test data are used for this chapter to train deep neural networks 1 and 2. The validation data is saved for the validation purpose for chapter six.

The data collected is then checked for missing data and unfit number format. Then, the data is discretized based on the upper and lower range for each variable. The number of data collected is 880. The data is then finally displayed in Table 0-8.

Table 0-8. Discretized data

Seq_no	I1	I2	I3	C1	C2	C3	U1	.	.	.	U7	O1	O2	O3	O4	O5	O6	O7	O8	O9
1	7	1	51	56	8	5	65	.	.		56	1	85	79	67	85	30	88	17	62
2	87	76	17	67	37	61	69	.	.		35	38	74	74	100	9	1	86	67	36
3	94	1	51	82	97	30	40	.	.		22	56	40	32	3	39	81	13	97	25
4	54	100	1	10	100	95	56	.	.		30	6	96	90	23	58	89	18	31	21
.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.													
.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.
876	14	26	100	55	80	48	99	.	.		46	59	82	68	84	76	63	44	12	96
877	54	100	100	90	62	19	39	.	.		6	42	19	53	82	58	36	5	68	52
878	94	51	67	56	18	46	60	.	.		95	5	28	84	1	62	2	27	39	93
879	74	76	100	66	67	93	3	.	.	.	19	79	33	77	87	72	83	4	8	75
880	41	100	67	46	69	100	91	.	.		98	47	81	25	73	84	63	89	1	92

Now, based on the data collected, the nodes and edges of the graph are created. The nodes for each record and the edges for each sequence are displayed in Table 0-9 and Table 0-10 respectively.

Table 0-9. Nodes and their characteristics data

seq_no	Id	label	I1	I2	I3	C1	C2	C3	U1	.	.	O6	O7	O8	O9	MD
1	1	s1	7	1	51	56	8	5	65	.	.					12.89798
2	2	o1								.	.	30	88	17	62	12.89798
3	3	s2	87	76	76	17	67	37	61	.	.					11.3495
4	4	ds1	80	75	-34	11	29	56	4	.	.					12.89798
5	5	o2								.	.	1	86	67	36	11.3495
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
			-							.	.					
3516	3516	ds879	33	24	-33	-20	2	7	88							10.97235
3517	3517	o880								.	.	63	89	1	92	16.32503
3518	3518	do879								.	.	-20	85	-7	17	10.97235

A total of 3581 nodes and 4396 edges are created from the given data

Table 0-10. edges and their characteristics for the given data

seq_no	Source	Target	type	Id	Label	Interval	Weight
1	1	2	Directed	1	e (s1-o1)	1hr	1
2	1	4	Directed	2	e (s1-ds1)	1hr	1
3	4	3	Directed	3	e (ds1-s2)	1hr	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
4393	3516	3515	Directed	4393	e (ds879-s880)	.	1
4394	3513	3518	Directed	4394	e (o879-do879)	.	1
4395	3518	3517	Directed	4395	e (do879-o880)	.	1
4396	3515	3517	Directed	4396	e (s880-o880)	.	1

The graph is then plotted as shown in Figure 0-5. The gephi software displays the nodes and edges connecting the nodes as shown in the fissure. Here we can observe that the nodes are plotted randomly to the plotting area and as the number of nodes are increasing, the graph will be dense as well. There is no clear information in this graph except that it shows nodes are connected by edges. It is also difficult to add labels for each node to identify where each node is placed because it will be condensed.

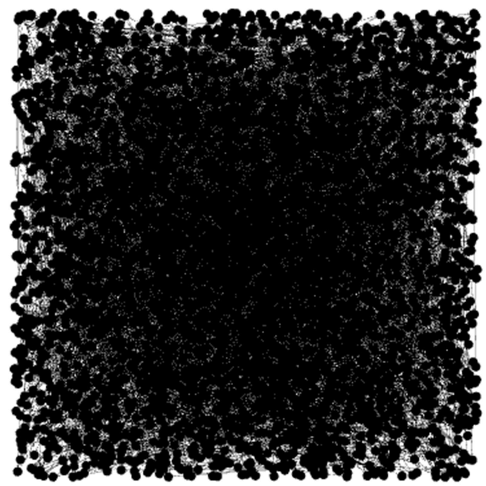

Figure 0-5: Gephi plot of the kiln data

The next step of the algorithm is to design the DNN1. The total number of data collected for analysis is 880. These data are split first into the train_test and validation data in the ratio of 90: 10. This gives 792 data for train_test and 88 for validation purposes. The train_test data is then further split into train and test in the ratio of 90:10, giving 712 and 80 records for train and test respectively. For the convenience of analysis in the deep neural network, the data collected is advised to be multiple of two. Hence, the nearest number of records for the train, test, and validation respectively will be 512, 64, and 64. The extra records will be truncated.

After several iteration tests, it is observed that three hidden layer network gives the best performance for the given data. The structure of the deep neural network 1 (DNN1) is then shown in Figure 0-6. Changing the number of neurons, learning rate, epochs, and batch size are the parameters important to obtain the best neural network. Hence accordingly the model allows varying these parameters and finally, the best model is to be achieved. As a result, 64 neurons for each i, u, and c dense layers,256 neurons for hidden layer 2, 128 neurons for hidden layer 3, 64 neurons for hidden layer 4, the learning rate of 0.00001, epochs 100, and batch size 16 are found to be the model parameters for best performance.

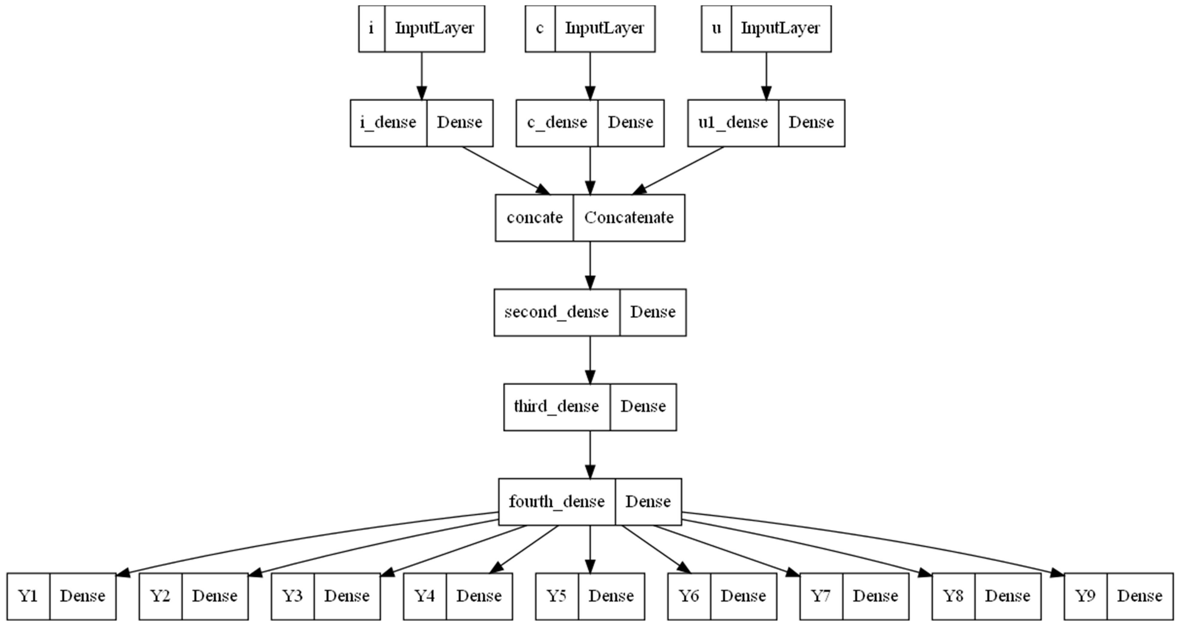

Figure 0-6: DNN1 structure

The DNN1 loss is observed in the following figure. This result is obtained after 92,169 parameters are optimized. Figure 0-7 has shown that when epochs are greater than 60, both the train and test losses are observed to be below 0.25. Hence, the Neural Network performance is improved well.

The square_root of each mean_squared_error for each output variable for the training data is as follows; y1= 62.98, y2= 0.80, y3= 5.94, y4= 5.14, y5= 46.39,y6= 8.66, y7= 28.21, y8= 1.79, y9= 26.33

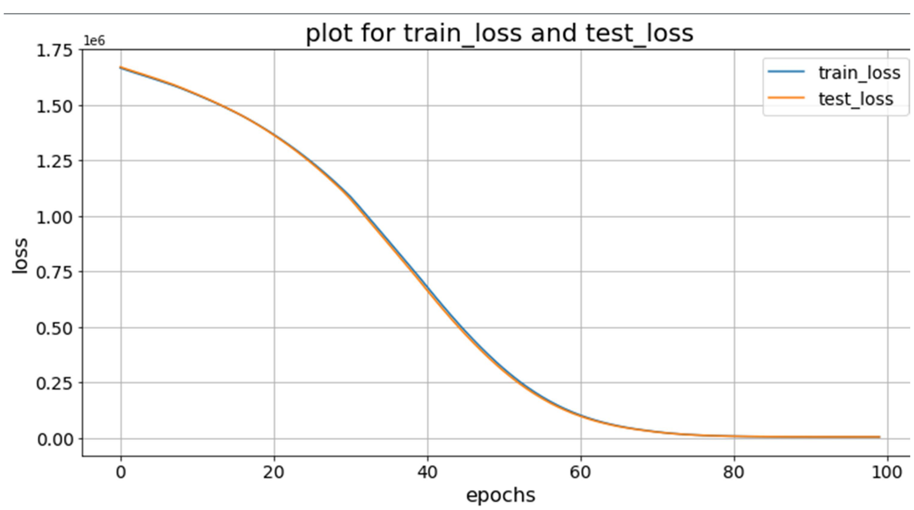

Figure 0-7: losses Vs epochs for training and test data

The model needs further prediction of the U variables because these data are not directly measurable in the production processes. The VARIMA model forecasts these values. There are seven U variables. These are u1, u2, u3, u4, u5, u6, and u7. The values are indicated in Table 0-11.

Table 0-11. The U variables data

	U1	U2	U3	U4	U5	U6	U7
1	65.94	19.54	4.29	2.38	1.86	0.7	0.55
2	66.05	21.19	5.05	3.19	2.97	1.87	0.34
3	65.19	19.67	5.09	3.35	2.95	0.52	0.21
.	65.65	21.46	6.99	2.52	1.83	0.72	0.29
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
877	.	.	.	.	.	.	.
878	65.79	21.5	5.03	2.06	1.57	0.7	0.94
879	64.06	20.31	5.47	3.41	2.35	1.15	0.18
880	66.7	20.08	4.23	3.74	2.86	0.77	0.97

The VARIMA model split the dataset into train and test in the ratio of 80% and 20 % which gives 704, and 176 records respectively. The plot for these data is indicated in Figure 0-8. Then, the test result and the VARIMA predicted result will be minimized in order to forecast future values.

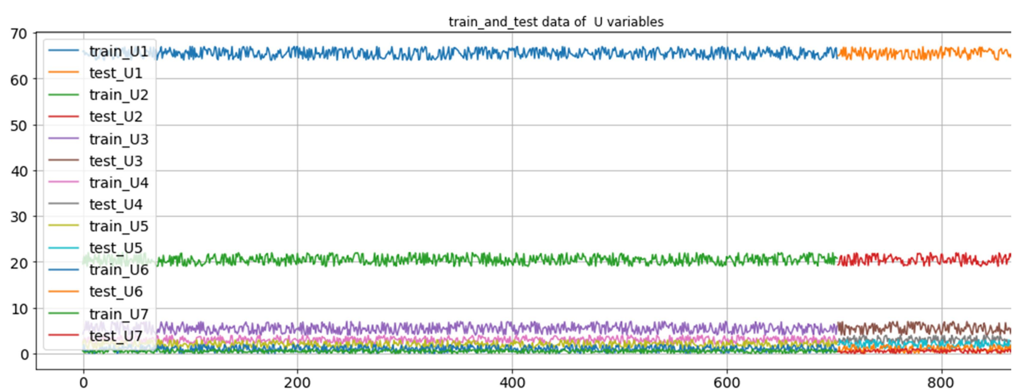

Figure 0-8:Train and test data of U variables

To begin with, each u variable is checked for stationarity using the Adfuller test (ADF), and the following result is obtained.

Augmented Dickey-Fuller Test on "U1"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.

Augmented Dickey-Fuller Test on "U2"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.

Augmented Dickey-Fuller Test on "U3"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.

Augmented Dickey-Fuller Test on "U4"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
Augmented Dickey-Fuller Test on "U5"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.

Augmented Dickey-Fuller Test on "U6"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
Augmented Dickey-Fuller Test on "U7"

Null Hypothesis: Data has a unit root. Non-Stationary.
Significance Level = 0.05
=> P-Value = 0.0. Rejecting Null Hypothesis.

=> Series is Stationary.

All U variables are stationary. Hence, the value of differencing, d, is zero in the VARIMA model. As a result, VARIMA can be equivalently done by VARMA. VARMA has two components, Vector Auto Regression (VAR) and Vector Moving Average (VMA). The order of lag order(p) for vector auto_regresion and the order of vector moving average (q) is then should be determined. These terms can be obtained from the grid search of different combinations of p and q values that will result in minimum mean squared error between test and forecasted values.

Once the stationarity of the U variables is tested, the next step is to test for cointegration between variables. The result of cointegration is found as follows in Table 0-12. Hence the U variables are cointegrated.

Table 0-12. Significance of the cointegration

Column Name	>Test Stat	>C(95%)	==>signif
U1	> 2015.73	> 111.7797	=> True
U2	> 1672.16	> 83.9383	=> True
U3	> 1339.75	> 60.0627	=> True
U4	> 1038.83	> 40.1749	=> True
U5	> 756.9	> 24.2761	=> True
U6	> 485.65	> 12.3212	=> True
U7	> 238.56	> 4.1296	=> True

To the maximum of p=10 and q=10, each U variable is searched for the optimum p and q values. For each variable of u, the p and q values with the minimum AIC value are calculated as follows using the auto ARIMA model.

Searching order of p and q for: U1
Performing stepwise search to minimize aic
ARIMA(1,0,1)(0,0,0)[0] : AIC=inf, Time=0.11 sec
ARIMA(0,0,0)(0,0,0)[0] : AIC=2285.770, Time=0.02 sec
ARIMA(1,0,0)(0,0,0)[0] : AIC=2079.709, Time=0.02 sec
ARIMA(0,0,1)(0,0,0)[0] : AIC=inf, Time=0.09 sec
ARIMA(2,0,0)(0,0,0)[0] : AIC=1997.503, Time=0.03 sec
ARIMA(3,0,0)(0,0,0)[0] : AIC=1960.575, Time=0.05 sec
ARIMA(4,0,0)(0,0,0)[0] : AIC=1944.330, Time=0.06 sec
ARIMA(5,0,0)(0,0,0)[0] : AIC=1930.553, Time=0.08 sec
ARIMA(6,0,0)(0,0,0)[0] : AIC=1908.888, Time=0.08 sec
ARIMA(7,0,0)(0,0,0)[0] : AIC=1902.600, Time=0.11 sec
ARIMA(8,0,0)(0,0,0)[0] : AIC=1892.699, Time=0.13 sec
ARIMA(9,0,0)(0,0,0)[0] : AIC=1893.688, Time=0.15 sec
ARIMA(8,0,1)(0,0,0)[0] : AIC=inf, Time=1.10 sec
ARIMA(7,0,1)(0,0,0)[0] : AIC=inf, Time=0.89 sec
ARIMA(9,0,1)(0,0,0)[0] : AIC=inf, Time=1.14 sec
ARIMA(8,0,0)(0,0,0)[0] intercept : AIC=1894.698, Time=0.17 sec

Best model: ARIMA(8,0,0)(0,0,0)[0]
Total fit time: 4.219 seconds
optimal order for:U1 is: (8, 0, 0)

similarly for other variables the following result is obtained.

Searching order of p and q for: U2
Best model: ARIMA(1,0,1)(0,0,0)[0]
Total fit time: 1.230 seconds
optimal order for:U2 is: (1, 0, 1)

Searching order of p and q for: U3
Best model: ARIMA(10,0,0)(0,0,0)[0]
Total fit time: 4.071 seconds
optimal order for:U3 is: (10, 0, 0)

Searching order of p and q for: U4
Best model: ARIMA(10,0,0)(0,0,0)[0]
Total fit time: 4.339 seconds
optimal order for:U4 is: (10, 0, 0)

Searching order of p and q for : U5
Best model: ARIMA(6,0,0)(0,0,0)[0]
Total fit time: 3.113 seconds
optimal order for:U5 is: (6, 0, 0)

Searching order of p and q for : U6
Best model: ARIMA(9,0,0)(0,0,0)[0]
Total fit time: 5.771 seconds
optimal order for:U6 is: (9, 0, 0)

Searching order of p and q for : U7
Best model: ARIMA(10,0,0)(0,0,0)[0]
Total fit time: 5.757 seconds
optimal order for:U7 is: (10, 0, 0)

the grid search for VARMAX is calculated and found as shown in Table 0-13. Searching is so rigorous that to run the grid-search of VARMAX it needs 3691 seconds.

Table 0-13. Grid search result of VARMAX

	P	q	RMSE U1	RMSE U2	RMSE U3	RMSE U4	RMSE U5	RMSE U6	RMSE U7
0	8	0	0.863626	0.910986	0.877282	0.587905	0.637263	0.582464	0.311819
1	1	1	0.854236	0.891592	0.881243	0.590685	0.599063	0.586665	0.306488
2	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868
3	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868
4	6	0	0.858605	0.924878	0.869575	0.587464	0.64398	0.579781	0.319799
5	9	0	1.62039	1.003359	1.173502	0.67292	0.797214	0.783085	0.330502
6	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868

The sorted value of Table 0-13 based on root-mean-squared-error (RMSE) is shown in Table 0-14. Hence the optimum value of (p,q) is (1,1). Hence the VARIMA parameters are set.

Table 0-14. Sorted grid search result of VARIMAX

	P	q	RMSE U1	RMSE U2	RMSE U3	RMSE U4	RMSE U5	RMSE U6	RMSE U7
1	1	1	0.854236	0.891592	0.881243	0.590685	0.599063	0.586665	0.306488
4	6	0	0.858605	0.924878	0.869575	0.587464	0.64398	0.579781	0.319799
0	8	0	0.863626	0.910986	0.877282	0.587905	0.637263	0.582464	0.311819
2	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868
3	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868
6	10	0	0.869752	0.916909	0.87461	0.593207	0.618497	0.581592	0.309868
5	9	0	1.62039	1.003359	1.173502	0.67292	0.797214	0.783085	0.330502

The next step of the model is designing Deep Neural Network 2(DNN2) as shown in Figure 0-9. This neural network is important to decide the next input node of the process based on the learning experience of the relationship between variables. From the predicted U variables and the desired values of the output, Y, variables, the objective is to determine the input variables, I, and controlled variables, c. Similar to DNN1, the 880 collected data are used to find the DNN2 parameters.

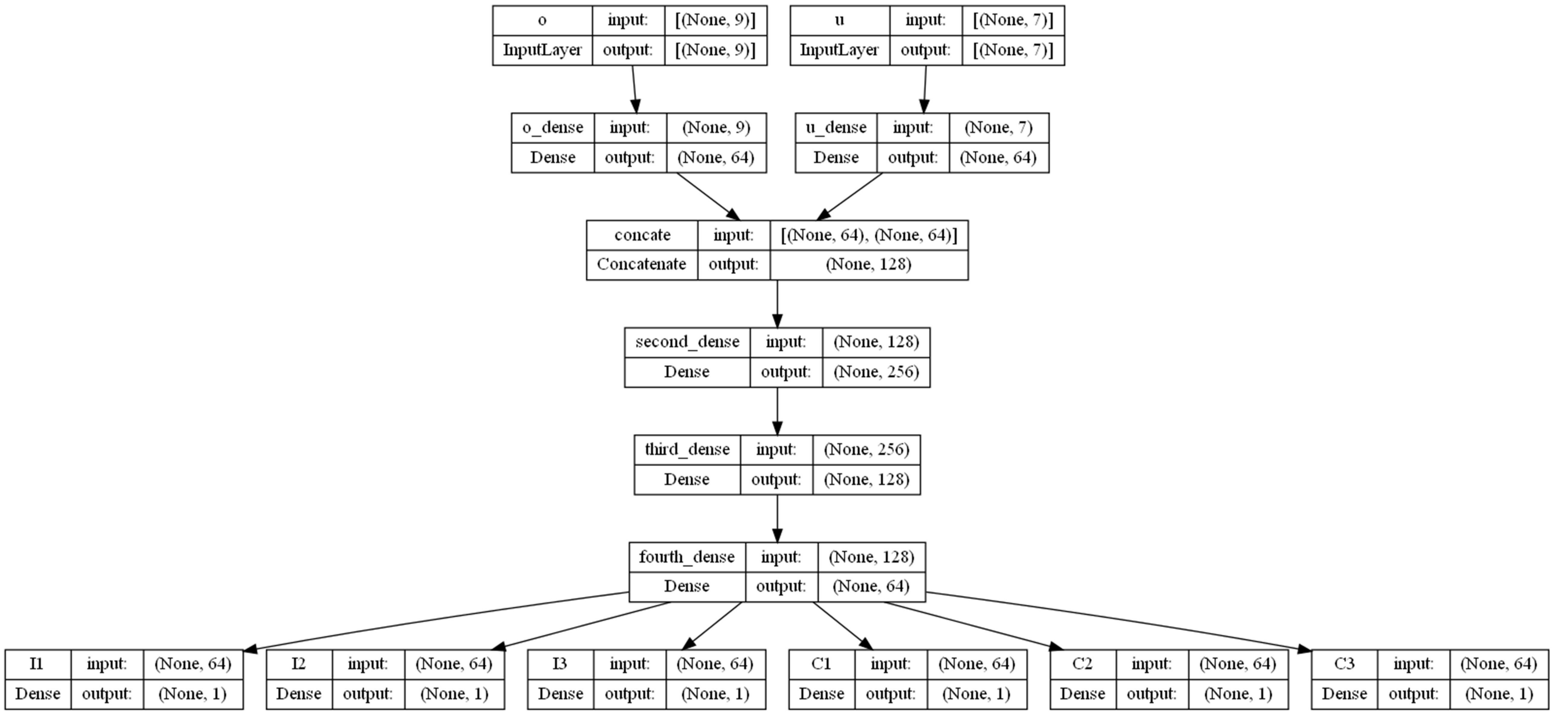

Figure 0-9:DNN2 structure

The same as the DNN1 design, DNN2 also has 512 training data, 64 test data, and 64 validation data extracted from a total of 880 collected data. However, input data for DNN2 is composed of u and o variables, whereas input data for DNN1 is composed of i, c, and u variables. The output of DNN2 is composed of i, and c variables whereas the output of DNN1 is composed of o variables.

Then graph will be formed from the discretized dataset for input and output of the DNN2 dataset as shown in Figure 0-10.

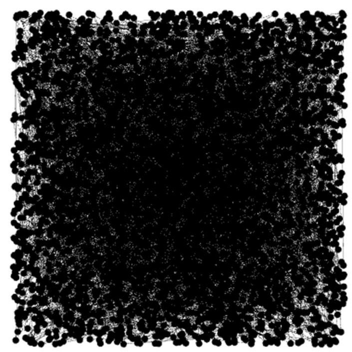

Figure 0-10: Gephi plot of the graph

Now, iterating the model parameters, like the number of hidden layers, number of neurons in each layer, batch size, learning rate, and epochs, and then it is found that when learning rate =0.00001,epochs=40, batch size=32, three hidden layers, and each having 256,128, and 64 neurons consecutively gives the best performing Neural Network And it is pictured as shown in Figure 0-11.

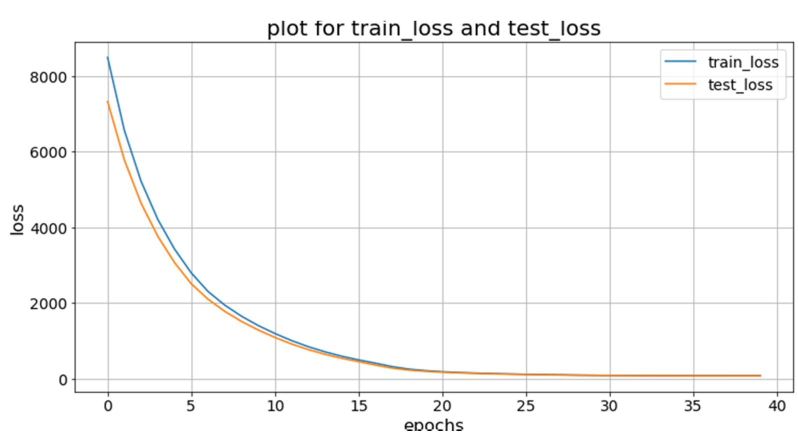

Figure 0-11:Plot for train_loss and test_loss

The total number of iterable parameters is 75,718, Table 0-15 shows where these parameters are found

Table 0-15: DNN2 parameters

Layer	Output Shape	# of parameters	Connected to
o (InputLayer)	[(None, 9)]	0	[]
u (InputLayer)	[(None, 7)]	0	[]
o_dense (Dense)	(None, 64)	640	['o[0][0]']
u_dense (Dense)	(None, 64)	512	['u[0][0]']
concate (Concatenate)	(None, 128)	0	['o_dense[0][0]', 'u_dense[0][0]']
second_dense (Dense)	(None, 256)	33024	['concate[0][0]']
third_dense (Dense)	(None, 128)	32896	['second_dense[0][0]']
fourth_dense (Dense)	(None, 64)	8256	['third_dense[0][0]']
I1 (Dense)	(None, 1)	65	['fourth_dense[0][0]']
I2 (Dense)	(None, 1)	65	['fourth_dense[0][0]']
I3 (Dense)	(None, 1)	65	['fourth_dense[0][0]']
C1 (Dense)	(None, 1)	65	['fourth_dense[0][0]']
C2 (Dense)	(None, 1)	65	['fourth_dense[0][0]']
C3 (Dense)	(None, 1)	65	['fourth_dense[0][0]']

The Neural Networks are designed; then, the next step is creating the graph and then plots the nodes and edges in the control chart. The total number of nodes formed from the known records is 3518 and the number of edges connecting these nodes is 4396. When the graph is formed for the first time, the nodes are positioned randomly in the plotting area, the size is in the default size, and the edges are formed with no weights, as indicated in Fig 44, the graph is formed and then plotted using gephi software. However, no information is displayed about the process.

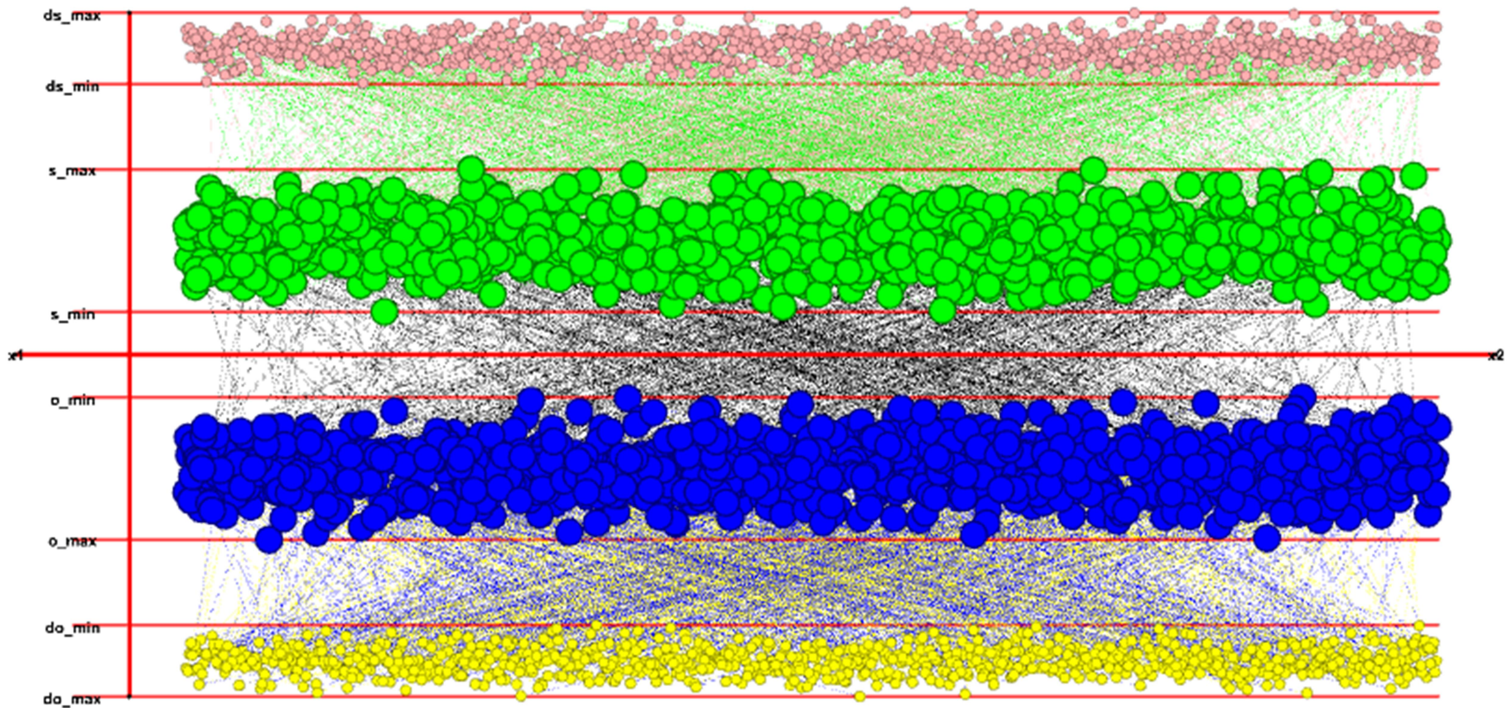

Figure 0-12: Condensed picture for GANNT diagram for kiln process

Figure 0-12 for the clinker production process indicates several important results. The main objective of the paper was to identify which combination of quality characteristics and how much should be changed to produce a product having the desired combinations of product quality characteristics. To show how this basic question is answered, the first ten sequences of nodes in Figure 0-12 are expanded and plotted again in Figure 0-13. Gephi is a flexible software that can be expanded to see the detail of the graph and there is a database that shows the detailed variables of each node.

Now say production is at the state of output node 10 (O10), and it is required the product characteristics to be changed to state O6. There are two paths to travel from State 10 to State 6. The first alternative is O10→do12→O7→do7 →O6, and the second alternative is O10→O9→O8→O4→O7→O6. Here the first alternative needs 4 steps and the second alternative needs 5 steps, hence the decision is the shortest path which is the first alternative. In this case, the path and alternatives are easily observable and simple. There can be conditions where more than two alternatives can appear and the path is long and complex. In that case, the Gephi graph software identifies the shortest path.

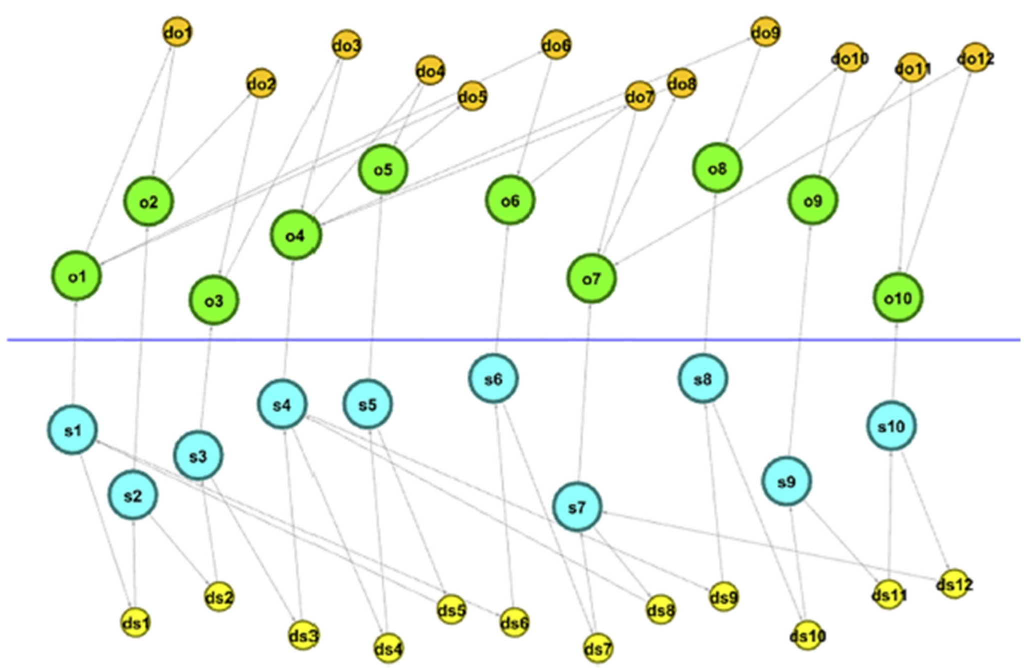

Figure 0-13: Expand the graph for the sequence of the first 10 operations

The statistical distance of each node is extracted from the gephi database and shown in the following table, Table 0-16.

Table 0-16. Statistical distance of the first 10 nodes of the sequence

Node	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
MD	2.98	13.08	11.92	4.10	9.60	2.98	10.50	1.67	4.10	10.13
Node	o1	o2	o3	o4	o5	o6	o7	o8	o9	o10
MD	2.98	13.08	11.92	4.10	9.60	2.98	10.50	1.67	4.10	10.13

5.6 Summary

Local cement industries use statistical process control to monitor process variables. Mainly shewart control chart is the frequently used technique that helps to keep each process variable within the allowable ranges. However, the cement production system is a long process that is divided into separate sections and the number of variables to be monitored for each section is large in number. Hence, it is observed that the cement industries have limitations in controlling process variables simultaneously. As a result, the multivariate statistical process control technique is proposed.

In the implementation of the model to the cement company, it is identified that the cement kiln system has 3 input variables, 3 control variables, 7 uncontrolled variables, and 9 output Variables. The operation needs to control the 9 output variables based on the 3 input variables, 3 control variables, and 7 uncontrolled variables. This decision is made by Deep Neural Network 1 and 92,169 parameters are optimized to make the best decision. There are 880 records collected from the plant for developing the model. From these 880 records 512 are used for training the model, 64 are used for test purposes, and 64 are saved for validation purposes.

A graph is created from the nodes and edges emanating from each record. There are 3518 nodes and 4396 edges formed to form the graph. Each node is plotted in the control chart based on Hotelling T2. The edges of the graph are important to trace the sequence of operation and the position of the nodes in the control chart describes the presence of the anomalies in the operation or the deviation of the operation from the acceptable range.

Deep Neural Network 2 (DNN2) is used to predict the next decision to achieve the targeted output and uncontrollable variables. This network takes the values of the uncontrollable variables from the VARMA model. Hence the performance of prediction is improved well. DNN2 optimizes 75,169 parameters to get the best performance of the Network.

In general, it is possible to say that the model extracts to the maximum all the information available among variables in the operation.

CHAPTER SIX

MODEL VALIDATION

The model developed in the previous chapters is needed to be validated to generalize and conclude that the model is performing well. There are a number of statistical methods used to validate the models. Here in this chapter, the specific tool is selected and the model is validated.

6.1 Introduction

The implementation of the model developed in Chapter Four took place in a cement company, as discussed in Chapter Five. During the implementation phase, a total of 880 data points were collected from the case company. These data points underwent filtering and were then split into three subsets for training, testing, and validating the model.

As previously discussed, Chapter Five outlined that 512 data points were extracted for training the model, 64 data points were designated for testing the model, and an additional 64 data points were set aside for validation purposes. In this chapter, the focus is on utilizing the 64 validation records to validate the model's performance.

The model comprises five key components: Deep Neural Network 1 (DNN1), Deep Neural Network 2 (DNN2), Vector Auto Regressive Moving Average (VARMA), Euler Graph, and Hotelling T2. The Mahalanobis distance (MD) determines the positioning of each node from the Euler graph on the chart. In this chapter, the validation data will be employed to individually validate each component of the model, followed by concluding the overall model's performance.

In practice, cement industries typically control their production systems based on Shewhart control charts. In this chapter, a comparison is made between the operator's skill and experience in running the plant based on the available data and the decisions made using the model developed in Chapter Four. The comparison assesses the model's decision-making capabilities in predicting uncontrollable variables, determining the desired process output based on input (i), controllable variables (c), and uncontrollable variables (u), as well as predicting the next move based on the previous outputs of the kiln system.

The model initiates its process by forecasting the uncontrollable variables, which are variables that directly impact the production system but cannot be directly measured. The model employs forecasting techniques to predict these values. Operators often struggle with accurately forecasting uncontrollable variables, as it requires the utilization of mathematical models for prediction. The effect of operators' lack of forecasting for uncontrollable variables on the process output will be examined in this chapter.

Next, based on the predicted uncontrollable variables and desired output variables, the model makes decisions regarding the input and controllable variables. Using the determined input, controllable, and uncontrollable variables, actual output variable values are obtained. A comparison is then made between the real output and the model output, and the system updates its parameters accordingly, allowing the process to continue its iterative loop.

Overall, this chapter will focus on validating the model by assessing its components, analyzing the impact of operators' forecasting capabilities for uncontrollable variables, comparing the desired process output, and evaluating the model's predictive capabilities based on the previous outputs of the kiln system.

6.2 Validating DNN1

In Chapter Five, the Deep Neural Network-1 (DNN1) was developed using the train and test data. In this section, the focus is on validating the performance of DNN1 using the 64 records specifically reserved for validation purposes. The validation process holds significant importance in the design of DNNs as it assesses the model's ability to generalize beyond the data used for training and testing.

Validation serves as a crucial step in evaluating the effectiveness and reliability of DNN1. It assures that the model can accurately predict and generalize patterns in new and unseen data. By employing the reserved validation dataset, the model's performance can be thoroughly assessed against real-world scenarios.

The primary objective of validating DNN1 is to ensure that it can effectively handle data that was not part of the training or testing phases. This enables the model to demonstrate its ability to make accurate predictions and generate reliable results for new inputs. By validating DNN1, any potential limitations or weaknesses in the model can be identified and addressed, leading to improvements and enhanced generalization capabilities.

The validation process involves feeding the 64 records into DNN1 and analyzing the outputs it produces. Comparing the predicted results with the ground truth values allows for a comprehensive evaluation of the model's performance. Through this validation procedure, it can be determined whether DNN1 has successfully learned the underlying patterns and relationships in the data and whether it can effectively generalize its predictions to unseen instances.

By validating DNN1 using the reserved validation dataset, its robustness and reliability can be assessed. This validation step is crucial in ensuring that DNN1 performs accurately and consistently in real-world scenarios, providing confidence in its ability to generalize and make reliable predictions beyond the training and testing datasets.

In the process of evaluating the model's acceptability to the given data, several statistical techniques can be employed. One such technique is the calculation of the R^2 value, also known as the coefficient of determination. This measure quantifies the proportion of variance in the dependent variable that can be explained by the independent variables included in the regression model. The R^2 value typically ranges between 0 and 1, with higher values indicating a better fit of the model to the data.

The R^2 value serves as an important indicator of the model's performance and its ability to capture the relationships between the independent and dependent variables. Assessing the proportion of variance explained provides insights into the predictive power and goodness of fit of the regression model. A higher R^2 value suggests that a larger proportion of the variability in the dependent variable can be accounted for by the independent variables included in the model.

Typically, an R^2 value of 0 indicates that the model fails to explain any variance in the dependent variable, while an R^2 value of 1 signifies that the model perfectly captures all the variance. However, in practical scenarios, achieving a perfect R^2 value of 1 is rare, and it is more common to observe values below 1. Nevertheless, the closer the R^2 value is to 1, the better the model's fit to the data and its ability to explain the observed variations.

By calculating the R^2 value, researchers and analysts can assess the extent to which the independent variables contribute to the variability in the dependent variable. This statistical technique provides a quantitative measure of the model's performance and offers insights into the quality and reliability of its predictions. A higher R^2 value indicates a stronger relationship between the variables and suggests that the model is more effective in capturing the underlying patterns and dynamics present in the data.

To sum up, calculating the R^2 value is a valuable technique for evaluating the acceptability of a model to the given data. It provides a quantitative assessment of the model's fit, indicating the proportion of variance in the dependent variable that can be explained by the independent variables. By aiming for higher R^2 values, researchers can strive for models that better capture the relationships and variability present in the data, thereby enhancing the model's predictive capabilities.

To calculate the R^2 value of a model, the following steps can be followed:

1. Fit the regression model to the data: In this case, the regression model used is DNN1, which was designed based on the train and test data. The model utilizes the actual values of the dependent variables to make predictions.
2. Calculate the mean of the dependent variable: Determine the mean of the dependent variables by summing up all the values of each dependent variable and dividing the sum by the total number of data points.
3. Calculate the total sum of squares (SST): Start by calculating the total sum of squares (SST). This is done by taking each observed dependent variable and subtracting the mean of the corresponding dependent variable. Then, square the differences obtained for each data point and sum them up. The resulting sum represents the SST.

4. Calculate the residual sum of squares (SRES): Next, calculate the residual sum of squares (SRES). This involves subtracting each predicted dependent variable from the corresponding observed dependent variable. Square the differences obtained for each data point and sum them up to get the SRES.
5. Calculate the R2 value: The R2 value can be calculated by dividing the explained variance (SST - SRES) by the total variance (SST) and then subtracting the result from 1. This provides a measure of how well the independent variables in the model explain the variability in the dependent variable.

By following these steps, the R^2 value of the model can be determined. The R^2 value indicates the proportion of variance in the dependent variable that is accounted for by the independent variables included in the regression model. A higher R^2 value suggests a better fit of the model to the data and a greater ability to explain the observed variations.

The designed Neural network is acceptable if the R^2 value of the coefficient of determination test is in the acceptable range. The R^2 statistics are used to determine if a sample of data comes from a population with a specific distribution. The neural network predicts some value and the distribution of that value is compared to the distribution of the actual value. If the R^2 value is not in the acceptable range it is necessary to improve the neural network.

In addition to the R^2 value, there are further refining techniques that can be employed to validate the NN model. One of these techniques is hold-out validation, which involves the following steps:

1. Split the data: The variable data is randomly divided into two subsets: a training set and a validation set. The recommended ratio for this split is typically 60-80% for the training data and 20-40% for the validation data. This ensures that a sufficient amount of data is used for training while reserving a separate portion for validation.
2. Train the model: The training data set is utilized to train the machine learning model. The model learns from the input data and adjusts its parameters to optimize its performance on the training set. This involves iteratively updating

the model based on the calculated errors or discrepancies between the predicted outputs and the actual outputs.

3. Validate the model: The validation data set is then used to evaluate the performance of the trained model on new, unseen data. By feeding the validation data into the model, its predictive capabilities are tested, and the resulting outputs are compared against the known ground truth values. This step helps assess how well the model generalizes to unseen data and provides insights into its performance and potential shortcomings.
4. Refine the model: Based on the results obtained during the validation step, the model can be refined and improved. This is done by adjusting the model's hyperparameters, which are settings that control its behavior and performance. By fine-tuning the hyperparameters, such as learning rate, regularization strength, or network architecture, the model's performance can be optimized. This process involves iterating and adjusting the hyperparameters repeatedly until an acceptable level of performance is achieved, as determined by the evaluation of the validation data.

By employing hold-out validation, the model's performance is assessed on separate data that it has not been trained on. This helps validate the model's ability to generalize and make accurate predictions on new, unseen instances. The iterative refinement process allows for the optimization of the model's performance, enhancing its predictive capabilities and ensuring its suitability for real-world applications.

A neural network has the following parameters to be adjusted so that the desired neural network is achieved. These are:

1. Number of neurons: The number of neurons in the hidden layers of the neural network is to be adjusted to control the complexity of the model. One important feature of increasing the number of neurons is it allows the NN to learn more complex relationships in the data.
2. Learning rate: The learning rate controls how much the weights and biases of the neural network be adjusted during the forward and backward iteration process. A high learning rate can cause the weights to oscillate so that it prevents the model be converged quickly, while a low learning rate can cause the model to converge slowly.

3. Activation function: The activation function determines the output of each neuron in the ANN. There are common activation functions like sigmoid, tanh, ReLU, and softmax. Changing the activation function can increase or decrease the performance of the model and hence it is necessary to test each activation function if the performance can be increased.
6. Regularization: Regularization techniques such as L1 and L2 Regularization are used to prevent overfitting in the neural network by adding a penalty term to the loss function. Mainly noise and irrelevant features are what regularization can eliminate.
7. Dropout: Similar to Regularization, dropout can eliminate overfitting in neural networks by randomly dropping out some neurons in the layers during training.
8. Batch size: batch size determines how many samples are used to update the weights of the neural network during iteration. The large batch size can lead to more stable convergence but require more memory and computational resources.
9. The number of epochs: the number of epochs determines how many times the neural network is trained on the entire training set. The higher number of epochs allows the model to converge to a better solution but also increases the risk of over-fitting.

The validation data of DNN1 is indicated in Table 0-1.

Table 0-1. Validation data

I1	I2	I3	C1	C2	C3	U1	.	O2	O3	O4	O5	O6	O7	O8	O9
104	12	13	129.03	2.04	1.54	43.19	.	0.84	55.57	13.24	10.58	10.10	28.67	11.10	29.97
91	10	17	122.97	1.99	1.67	44.19	.	0.59	50.11	20.40	10.45	9.86	28.20	10.86	30.87
96	8	13	125.61	2.00	1.74	43.98	.	0.56	57.91	12.44	10.25	9.98	27.97	10.98	29.28
91	12	15	118.11	2.09	1.64	43.86	.	0.28	59.18	11.63	10.29	9.86	27.86	10.86	28.95
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93	9	14	126.64	1.79	1.38	43.19	.	1.31	57.90	8.34	9.45	11.41	28.86	12.41	30.03
96	8	13	125.61	2.00	1.74	43.98	.	0.56	57.91	12.44	10.25	9.98	27.97	10.98	29.28
91	12	15	118.11	2.09	1.64	43.86	.	0.28	59.18	11.63	10.29	9.86	27.86	10.86	28.95
90	10	13	122.68	2.02	1.59	43.42	.	1.64	59.22	7.95	10.31	10.10	28.21	11.10	28.64

The values of these 64 new data identified for this purpose are now inserted into the DNN1 and tested the model if it can generalize for any process data. If it is not satisfactory, the model is refined and tested iteratively until an acceptable percentage of generalization is reached.

There are two ways that the refining of the DNN1 can be performed. The first one is increasing the layer and number of neurons on each layer so that the unexplained features are extracted more. The second one is to split the DNN1 so that rather than increasing the layers, replace it with two or than two DNNs. In this chapter, the second approach is followed and the result is presented in Table 23 and pictorial it is represented by Fig. 48.

As shown in table 23, various Neural Networks can be designed for the given train and test data. However it fails when it is tested by validation data. Here in the design of DNN1 four neural networks are working together to satisfy the requirement of validation result. The four NNs are represented by NT29, NT34, NT45, and NT64. The numbers in the designation shows that more than 64 NNs were designed that satisfy the train and test requirement, but these four are found the better result to satisfy the requirement of validation.

The comparison between the prediction result generated by DNN1 and the train-test values of the data is presented in **Table 0-2**. The table displays the respective train, test, and validation result for each output variables. The red label shows the best performance for that specific variable. The variable Y1, for example, achieves its best prediction of validation data by NT34 and the value is 67%. The summary is shown in **Figure 0-1**.

Table 0-2 The outputs of the four NNs

	NT29			NT34			NT45			NT46		
	Train	Test	Val	Train	Test	Val	Train	Test	Val	Train	Test	Val
Y1	100	95	30	95	85	67	99	94	37	99	94	36
Y2	98	96	83	94	88	69	98	94	65	97	93	55
Y3	100	98	76	98	95	84	100	98	69	100	98	70
Y4	100	98	69	100	96	67	100	98	83	100	98	84
Y5	99	97	44	94	90	6	99	96	41	97	92	33
Y6	98	91	79	99	91	73	99	97	80	99	96	81

Y7	100	97	55	97	91	32	97	92	56	98	92	55
Y8	98	94	77	95	91	55	99	95	86	98	94	88
Y9	99	98	61	98	93	61	100	98	64	99	97	69

To quantitatively evaluate the accuracy and goodness of fit of the model, the R^2 values for each output of the model are calculated and presented in Table 2. The R^2 value, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables used in the regression model. Higher R^2 values indicate a better fit of the model to the data.

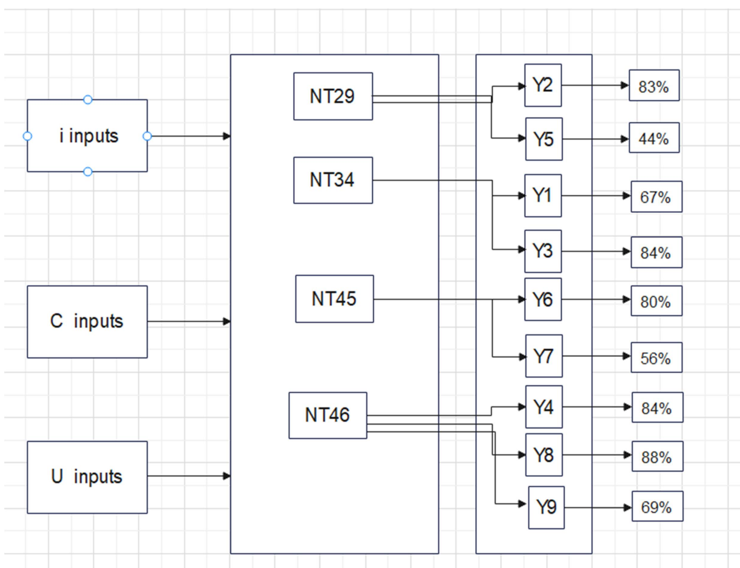

Figure 0-1: Arrangement of the four NNs

Upon analyzing the R^2 values, it is determined that approximately 72.78% of the predictions made by the model are deemed to be accurate for the validation data. This indicates a substantial level of prediction performance and suggests that the model exhibits a favorable degree of accuracy and generalization when applied to unseen data.

The R^2 values serve as a valuable metric to assess the model's ability to capture the underlying patterns and relationships within the data. By providing a quantitative measure of the model's predictive power, the R^2 values offer insights into the reliability and effectiveness of the DNN1 in producing accurate predictions for the given validation dataset.

6.3 Validating DNN2

The validation process for DNN2 follows a similar technique employed for validating DNN1. Using the same set of validation data, the inputs and outputs of both neural networks, namely DNN1 and DNN2, exhibit differences. However, it is crucial to ensure that these two neural networks are synchronized in order to maintain stability within the system during their simultaneous iterations.

Table 0-3 presents the results of the validation, showcasing the percentages of correctly predicted values for each output (i1, i2, i3,c1, c2, c3) variable obtained from DNN2. According to the table, XNT1 neural network accurately predicts 99% of i1 values, 79% of i2 values, and 89% of i3 values. Additionally, the remaining predicted data, specifically c1, c2, and c3, achieve prediction accuracies of 98%, 73%, and 61%, respectively using the XNT6 neural network.

To provide a comprehensive overview of the combined results, a diagrammatic representation (Figure 0-2) displays the predictions for each variable (i1, i2, i3, c1, c2, c3) in relation to the given uncontrollable (U) and output (O) variables. This graphical representation helps to visualize the predicted values and their corresponding relationships.

The validation process serves as a crucial step in assessing the accuracy and reliability of DNN2. By comparing the predicted values to the actual values for the validation data, the effectiveness of the neural network in capturing the underlying patterns and relationships within the dataset is evaluated. The high percentages of correctly predicted values for both input and output variables demonstrate the capability of DNN2 to effectively generalize and produce accurate predictions based on the given data.

The synchronized operation of DNN1 and DNN2 is vital to ensure stability within the system. By aligning the predictions from both neural networks, the model can achieve a harmonious balance and provide consistent and reliable results. This synchronization allows for a comprehensive understanding of the system's behavior and aids in making informed decisions based on the combined predictions of the two neural networks.

In summary, the validation of DNN2 demonstrates its ability to accurately predict the values of both input and output variables. The synchronization of DNN1 and DNN2 guarantees the stability of the system and enhances the overall performance of the model. The graphical representation further enhances the comprehension of the predicted values, highlighting the relationships between the variables. These findings contribute to the validation and reliability of the model, ensuring its suitability for controlling and monitoring the given multivariate data analysis.

Table 0-3 The outputs of the two NNs of DNN2

	XNT1			XNT6		
	Train	Test	Val	Train	Test	Val
i1	100	100	99	98	97	98
i2	92	93	79	77	85	75
i3	96	96	89	89	89	86
C1	100	100	96	100	100	98
C2	97	97	58	96	95	73
C2	91	88	39	92	91	61

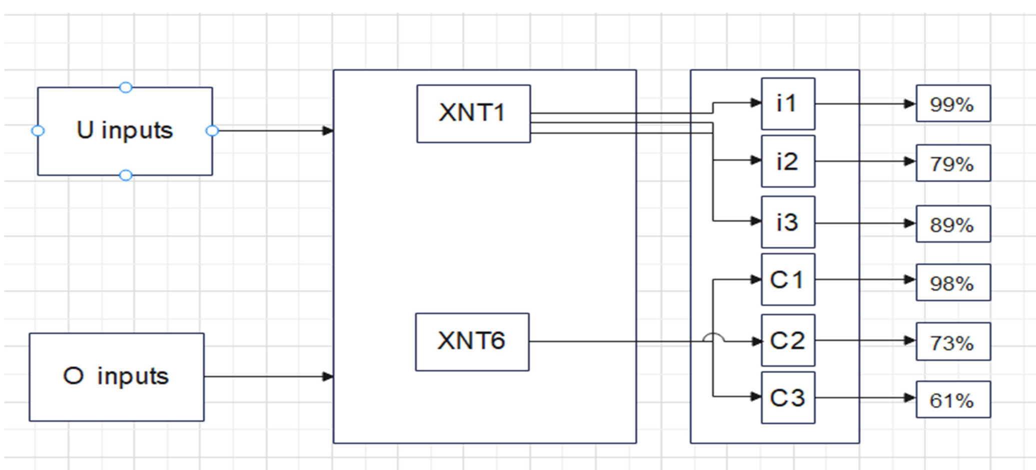

Figure 0-2: Arrangement of the two NNs of DNN2

6.4 Validating the VARIMA model

VARIMA is the time series model which is validated similarly to any other time series model. Here are the steps to be followed:

1. Split the data: here the data is split for training and testing purposes. The training data is used to fit the VARIMA model, while the test set will be used to evaluate its performance.
2. Fit the model: Training data is used to fit the model so that model parameters are adjusted.
3. Make a prediction: use the fitted VARIMA model to make a prediction of test data.
4. Evaluate the performance: evaluate the performance of the VARIMA model by comparing the predicted values to the actual values in the test data. Metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error(MAPE) are used to access the accuracy of the prediction.
5. Adjust the model: If the result of the model is not satisfactory, the model will be adjusted by changing the order of autoregressive and moving average terms.
6. Repeat steps 2-5: iterating the process of fitting the model, making predictions, and evaluating the performance until a satisfactory model is achieved.
7. Finalize the model. Once the performance of the model is satisfactory, the validation step will be completed.

The main objective of adding the VARIMA model is to predict the uncontrollable variables incorporated in the model. As it is discussed in Chapter 5, there are seven uncontrollable variables in the model represented by u1, u2, u3, u4, u5, u6, and u7. The model was trained and tested for the given data. Here in this section, the result of the VARIMA model for the validation data is observed and if it is not satisfactory, it will be retained.

Chapter 5 discuss the model development of the VARIMA model to the maximum of p=10 and q=10, now in order to get more precise result for the validation data, the p and q values are extended to p=20 and q=20. The result of the combination of p and q with the minimum AIC value is calculated and presented using the auto ARIMA model as shown below, Table 0-4.

Table 0-4 The RMSE for different combinations of p and q

	p	Q	RMSE U1	RMSE U2	RMSE U3	RMSE U4	RMSE U5	RMSE U6	RMSE U7
0	10	0	0.699018	0.94193	0.315292	0.238764	0.152477	0.055276	0.489071
1	8	0	0.711813	0.912496	0.319952	0.234642	0.152842	0.058434	0.482373
2	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485
3	12	0	0.658306	0.915313	0.310475	0.233485	0.15007	0.054067	0.502922
4	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485
5	0	1	0.592295	0.884912	0.306016	0.233475	0.149446	0.052374	0.473851
6	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485

The sorted value of Table 0-4 based on root-mean-squared-error (RMSE) is shown in Table 0-5. Hence the optimum value of (p,q) is (0,1).

Table 0-5 sorted values based on RMSEof p and q

	p	Q	RMSE U1	RMSE U2	RMSE U3	RMSE U4	RMSE U5	RMSE U6	RMSE U7
5	0	1	0.592295	0.884912	0.306016	0.233475	0.149446	0.052374	0.473851
3	12	0	0.658306	0.915313	0.310475	0.233485	0.15007	0.054067	0.502922
2	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485
4	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485
6	11	0	0.666375	0.91516	0.312186	0.237506	0.15061	0.05417	0.493485
0	10	0	0.699018	0.94193	0.315292	0.238764	0.152477	0.055276	0.489071
1	8	0	0.711813	0.912496	0.319952	0.234642	0.152842	0.058434	0.482373

Now, from the model selected with minimum RMSE value, the prediction values are generated. The result is indicated in Table 0-6.

Table 0-6 Validation and prediction values of U variables

	Validation data							Prediction data						
	U1	**U2**	**U3**	**U4**	**U5**	**U6**	**U7**	**U1_p**	**U2_p**	**U3_p**	**U4_p**	**U5_p**	**U6_p**	**U7_p**
1	43.09	9.77	3.26	2.40	1.66	0.39	36.76	42.74	8.89	3.25	1.93	1.56	0.51	35.29
2	44.34	12.01	3.52	2.53	1.44	0.40	36.23	42.73	8.89	3.25	1.93	1.56	0.51	35.29
3	44.00	12.44	3.21	2.23	1.52	0.50	36.70	42.73	8.89	3.25	1.93	1.56	0.51	35.29
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
63	43.30	9.70	3.51	2.20	1.49	0.50	35.78	42.46	9.04	3.18	1.85	1.57	0.51	35.25
64	43.32	10.37	3.26	1.97	1.66	0.47	35.58	42.46	9.05	3.18	1.85	1.57	0.51	35.25

Then the error term is calculated from the prediction and the validation data. The mean absolute percentage error (MAPE) is calculated to see the error deviation of prediction. The result is indicated as shown in the following table, Table 0-7.

Table 0-7 Mean absolute percentage error for U variables

	U1	U2	U3	U4	U5	U6	U7
MAPE	3.02	18.38	6.70	13.34	8.91	18.08	2.68

The average MAPE will be 10.15. Hence the precision is about 90 percent, which is acceptable.

6.5 Validating the calculated Hotelling T2 values

Hotelling T2 is a multivariate statistical analysis tool used to compare multiple variables simultaneously and is a need to analyze high-dimensional data. The concept is similar to Mahalanobis distance where both measure statistical distance in multivariate statistics. However, they differ in their intended purpose where Mahalanobis measures the distance between a point and a group of points in a multivariate space whereas Hotelling compares the means of two groups of observations in a multivariate space.

The validation of the statistical distance is done in the following ways:

1. Cross-validation: in this method, the Mahalanobis distance is calculated on a subset of the data and validated on the remaining data. This helps to assess the generalization performance of the distance metrics and to detect overfitting.
2. Visualization: Mahalanobis distance can be used to visualize multivariate data in two or three dimensions. By plotting the Mahalanobis distance of each observation, it is possible to identify outliers or clusters of similar observations.
3. Sensitivity analysis: This method involves assessing the impact of small changes in the covariance matrix on the Mahalanobis distance. This helps to understand the stability of the distance metric under different conditions and to detect possible sources of biases.
4. Comparison with other distance metrics: Mahalanobis distance can be compared with other distance metrics, such as Euclidian distance or cosine distance, to assess its superiority in specific contexts.
5. Statistical tests: Mahalanobis distance can be used as a dependent variable in regression analysis or as a test statistic in a hypothesis test. This helps to assess the relationship between the distance metric and other variables of interest or to test hypotheses about the underlying population.

For the given data validation data, additional 204 nodes and 309 edges are added to the graph created for the training and test graph. The statistical distance of these validation data is calculated following the same procedure as that of training and test data and it is indicated as shown in Table 0-8.

Table 0-8 The statistical distance of nodes for validation data

Seq_no	MD_S	MD_DS	MD_O	MD_DO
1	11.91179	13.06142	4.724907	2.555853
2	5.586519	4.241573	9.925148	3.383033
.	.	.	.	.
.	.	.	.	.
62	5.539864	7.113855	3.807189	0.27399
63	7.899495	5.932763	5.330582	3.356166

64	6.041319	4.034569

The validation data's statistical data (MD) is visually represented in Figure 50, providing a clear depiction of the values within the acceptable range of the calculated statistical distance from the train_test data. The plot showcases key statistical metrics that further support the evaluation of the model's performance.

Analyzing the data, we observe that the mean value of the statistical distance is calculated as 4.9358. This value serves as a central indicator, representing the average distance of the data points from the expected or target range. Additionally, we find that the minimum statistical distance recorded is 1.196, indicating the closest proximity of a data point to the desired control limits.

Furthermore, the plot highlights two extreme values, namely 18.847 and 14.442, which represent the maximum statistical distances observed. These outliers provide insights into the instances where the data points deviate the most from the desired control limits, potentially indicating anomalies or exceptional conditions within the system.

It is worth noting that the majority of the statistical distances fall below 10, indicating that the majority of the data points are well within the acceptable range. This distribution demonstrates the model's ability to effectively monitor and control the multivariate data, ensuring that the variables remain within desired boundaries and exhibit stable behavior.

In summary, the pictorial representation in **Figure 0-3** effectively showcases the statistical data (MD) of the validation data. The plot's analysis reveals a mean value of 4.9358, with the minimum value recorded at 1.196 and the two extreme values observed at 18.847 and 14.442. These insights provide valuable information regarding the distribution and proximity of the data points to the expected control limits, further validating the model's efficacy in monitoring and managing the multivariate data.

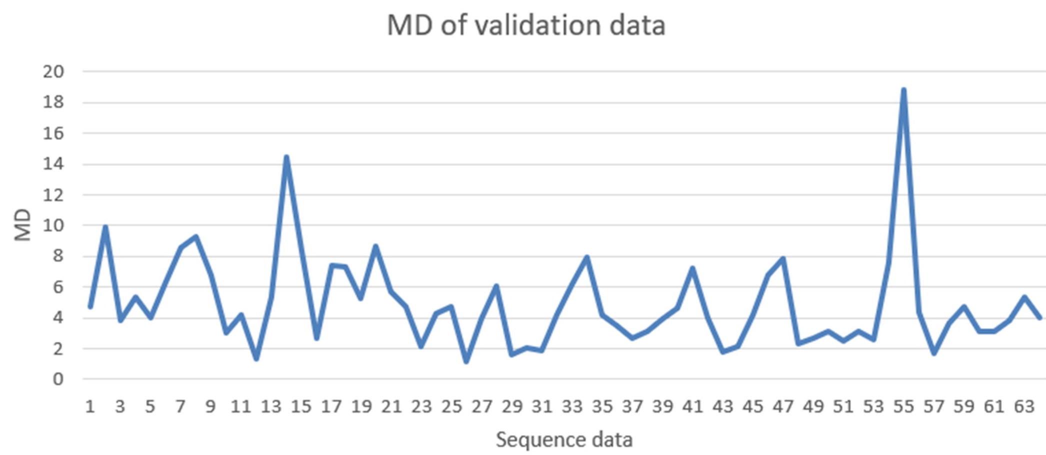

Figure 0-3: MD for validation data

6.6 Summary

Model validation plays a crucial role in assessing the accuracy and generalizability of a predictive model. In this chapter, we focus on validating the model developed in Chapter Four using statistical methods. While there are various techniques available for model validation, we adopt an approach that takes into consideration factors such as dataset size, model complexity, and computational resources available.

It is important to note that the model includes two neural networks, which require substantial computational capacity to operate efficiently. Taking into account the aforementioned constraints and considering the fundamental requirements of the model development process, we have successfully demonstrated that the model exhibits strong predictive power.

An integral aspect of the model involves predicting variables that cannot be directly measured. For this purpose, we employ the VARIMA model, which demonstrates a remarkable precision of approximately 90% in predicting these variables. This prediction capability is essential as it allows for the improvement of model outputs by incorporating not only directly measurable variables but also non-measurable ones, enhancing the overall understanding of the system dynamics.

Furthermore, the model employs the Hotelling T2 formula to predict the statistical distance (MD) of each node within the control chart. The results indicate that the model accurately predicts the values of each node, as they fall within the acceptable range of statistical distance. This range takes into account the variance and covariance between variables, ensuring a comprehensive assessment of the multivariate data analysis.

In summary, based on the validation data provided and the successful prediction of various variables, we can confidently conclude that the model has undergone effective validation. It has demonstrated its usefulness in controlling and monitoring multivariate data analysis, providing valuable insights into the underlying processes. The model's robust performance and ability to predict both directly measurable and non-measurable variables further establish its reliability and practicality in real-world applications.

CHAPTER SEVEN

CONCLUSION AND RECOMMENDATION

This chapter concludes the output of the research. Furthermore, the recommendation and future research directions are proposed. The conclusion indicates that the model developed is promising to be effective in improving the performance of the industry. But it is still open to further improvement.

7.1 Conclusion

To investigate and analyze the behavior of a production system effectively, as well as to ensure efficient monitoring and control of the production process, it is crucial to examine the relationship between various process parameters and the resulting product quality characteristics. This research sheds light on the limitations encountered when employing multivariate statistical process control (MSPC) methods for monitoring and controlling the production process. Traditional MSPC techniques are capable of detecting instances of the production system being out of control. However, they fall short of identifying the specific variable or group of variables responsible for the out-of-control signal. Consequently, it becomes the responsibility of process engineers and operators to determine which process parameters need correction to restore normal operation.

As the number of parameters requiring control increases, the challenge of effectively managing the system becomes more pronounced. Traditional MSPC methods typically represent the cumulative effect of variables as a single point on the control chart. However, this approach cannot reveal the intricate relationships among variables. This study proposes an innovative approach by integrating Neural Networks and Graph theory, where the single point representation on the MSPC chart is substituted with interconnected nodes. This novel control chart offers a visual depiction of the relationships between variables.

The findings of this research demonstrate that the integration of Neural Networks and Graph theory simplifies the complexity associated with traditional MSPC techniques. Furthermore, it yields results that are closely aligned with human intuition. By providing a more comprehensive understanding of the interdependencies among variables, this new type of control chart empowers operators and process engineers to make informed decisions regarding the correction of process parameters. Ultimately, this advancement enhances the ability to monitor, control, and optimize the production process, thereby facilitating improved product quality and operational efficiency.

The Traditional MSPC procedures rely on plotting points, which serve to convey two essential pieces of information. Firstly, each point represents the statistical distance between the multivariate values and the nominal values. This provides a measure of how far the process deviates from the desired target. Secondly, the collection of points forms a pattern that reveals the flow of the process. However, the novel control chart developed using nodes in the Graph offers a more comprehensive display of information.

Nodes in the Graph-based control chart possess several attributes, including size, color, position, and label, which can accommodate an extensive range of information. The size of a node indicates the frequency of occurrence of a particular vector in the process, shedding light on its relative importance. Color coding of nodes allows for the categorization of different nodes, providing a visual distinction between various groups or classes. The position of a node on the chart represents the statistical distance of its values from the nominal value, allowing for a clear understanding of the extent of the deviation.

Furthermore, each node is labeled with a specific name, facilitating easy identification and reference. The label serves as a quick reference point for process engineers and operators to identify and analyze specific nodes of interest. Additionally, all other attributes associated with nodes can be readily accessed from the database of the graph, ensuring that a wealth of information is readily available for analysis.

By utilizing this innovative control chart, process engineers gain access to a wealth of valuable information, significantly enhancing their decision-making process. The combination of size, color, position, and label attributes provides a comprehensive overview of the process, enabling engineers to quickly identify critical nodes, understand their significance, and take appropriate actions. The wealth of information available empowers process engineers to make informed decisions that contribute to the optimization and improvement of the production process, leading to enhanced product quality, operational efficiency, and overall process control.

Within the innovative algorithm presented in this research, a neural network runs within each graph node and leverages the power of minimized error to predict the next decision point accurately. The neural network plays a crucial role in validating the model based on collected process data, providing an additional layer of confidence in the decision-making process.

The network of nodes designed from the graph offers a comprehensive view of the entire process, allowing process engineers to observe the big picture from the control chart drawn by the nodes. The algorithm provides a clear representation of the change in variables between two consecutive process decisions, allowing engineers to identify precisely which variable or group of variables have been altered and to what extent.

The presence of nodes in the graph further enhances the decision-making process by enabling process engineers to determine precisely which variables need to be modified to reach the optimal decision point. Through careful analysis of the node attributes, including size, color, position, and label, engineers can determine which variables are most critical to the overall process and make informed decisions accordingly.

Overall, the incorporation of neural networks within the nodes of the Graph-based control chart allows for greater accuracy and validation in the decision-making process. The visualization provided by the network of nodes enables engineers to gain a comprehensive understanding of the process, enabling them to identify critical variables and make informed decisions that optimize and enhance the production process.

Further investigation is required to explore the application of integrating neural networks and graph theory for the implementation of soft sensors. If it becomes possible to predict the relationships between variables, it would facilitate the detection of inaccurate information from sensors, discrepancies in laboratory results, and sensor failures. This research has the potential to significantly improve the reliability and accuracy of soft sensors in industrial processes.

When applied to the clinker production process, the algorithm yields several noteworthy outcomes. During operation, operators must determine whether the process is running normally or not. By incorporating the procedures of Hotelling T2 and plotting the nodes on the control chart, the algorithm visually represents the pattern of the production process. Consequently, operators can utilize their knowledge of identifying control chart patterns to assess the normality of the process.

By integrating a neural network into the algorithm, the system can be trained to remember past data that has yielded successful outcomes. This capability ensures consistent operation sequences, regardless of the operator's skill or experience. As the production process heavily relies on operator expertise, having a new or inexperienced operator can lead to relatively unstable operations. However, by leveraging the neural network to memorize and utilize previous data, the system can support and stabilize the operations conducted by new operators.

The application of graph theory to the control chart provides operators with enhanced understanding and decision-making capabilities regarding the operation process. The weight of the edges, color, and size of the nodes, and node labels and positions serve as direct observations for operators. Moreover, the graph database offers the flexibility of not being constrained by the number of node attributes, empowering operators with a comprehensive view of the system.

The production of high-quality clinker presents numerous challenges due to the involvement of multiple variables within the production system. Variables such as maximum clinker output, minimum fuel consumption, optimized air-fuel ratio, reduction of free CaO, maintenance of acceptable liquid phase levels, burnability index range, coating index range, and many others significantly impact the production process. Ultimately, the interconnectedness of the edges within the graph determines the overall pattern of the plant's performance and output.

7.2 Recommendation

The GANNT model encompasses a multidisciplinary approach, leveraging the collective knowledge of industrial engineering, statistics, mathematics, and computer science. However, it is worth noting that the research was conducted solely by an individual from the industrial department. To further enhance the study's comprehensiveness and validity, it is highly recommended to undertake future research by integrating researchers from various departments.

By involving experts from diverse fields, such as industrial engineering, statistics, mathematics, computer science, and potentially other relevant disciplines, a more holistic perspective can be attained. Each discipline brings unique insights and methodologies that can contribute to a more robust analysis and a comprehensive understanding of the subject matter.

The collaboration between researchers from different departments allows for a broader range of expertise and skill sets to be applied. For example, industrial engineers can provide insights into process optimization and efficiency, statisticians can contribute their expertise in data analysis and modeling, mathematicians can offer mathematical frameworks and algorithms, and computer scientists can assist in developing advanced computational techniques and visualization tools.

Integrating researchers from various departments foster a multidimensional approach to the study, enabling a more thorough exploration of the research problem. This interdisciplinary collaboration encourages the exchange of ideas, promotes critical thinking, and ensures that different aspects of the research are adequately addressed.

Furthermore, the involvement of multiple researchers increases the likelihood of identifying potential limitations, considering alternative perspectives, and conducting rigorous peer reviews. This enhances the credibility and reliability of the research findings and allows for a more comprehensive interpretation of the results.

In conclusion, to maximize the effectiveness and validity of studies utilizing the GANNT model or similar multidisciplinary approaches, it is strongly recommended to foster collaboration among researchers from different departments. This collaborative effort ensures a more comprehensive analysis, promotes diverse expertise, and leads to more robust and reliable research outcomes.

7.3 Future Research Direction

This research sought to make significant advances in multivariate statistical process control by creating the GANNT chart. Not only was its development intended, but its practical implementation within industrial sectors as well. Furthermore, advanced statistical techniques should be utilized in order to enhance process control and quality management practices with an ultimate aim of improving operational efficiency and product quality.

Future research directions suggest further exploring and validating the assumptions underlying the GANNT model, particularly its assumptions involving assumptions regarding industries, process types, operational contexts etc. Even though initial assumptions were carefully taken into account when developing the model, it's still necessary to assess their robustness and applicability across various industries, process types etc. By conducting exhaustive studies that explore different scenarios, researchers can gain a more in-depth knowledge of its limitations and strengths thereby assuring its effectiveness in real world settings.

One future proposal involves applying the GANNT model across industries to assess its effects and assess performance. While current research primarily focused on developing and validating it in cement industry, testing its practicality and effectiveness across various industrial settings such as manufacturing, healthcare, transportation or telecommunications will allow researchers to test its practicality and assess industry-specific insights that reveal adaptability challenges or highlight any areas where further refinements or customization may be necessary. Such comprehensive analysis will reveal key insights into its versatility while pinpointing any areas requiring further modification or customization.

Implementation of the GANNT model across different industries enables researchers to gather empirical data and performance metrics, validating its effectiveness as well as creating an opportunity for constant refinement and enhancement. By analyzing collected data and assessing model's performance in real world applications, researchers gain invaluable insights into strengths, weaknesses and areas for improvement - an iterative feedback loop which ensures continuous enhancement and applicability to various industrial contexts.

Future research should also explore potential extensions or modifications of the GANNT model. As statistical process control is an ever-evolving field, new techniques and methodologies emerge regularly, so researchers should stay abreast of these advancements to incorporate any relevant innovations into the GANNT model; such as artificial intelligence (AI), machine learning algorithms or big data analytics which might improve predictive capabilities or decision making accuracy - thus staying abreast of technological innovations ensures it keeps pace with industrial challenges and needs.

In conclusion, expanding the application of the proven model beyond the cement factory to diverse industries provides valuable insights, facilitates cross-industry learning, and enables the development of a more generic and versatile model. By addressing the unique challenges and requirements of different sectors, the model can be refined to better serve the needs of a wide range of manufacturing industries, ultimately contributing to improved process control, optimization, and operational excellence. Furthermore, multivariate process analysis extends beyond the boundaries of the manufacturing industry and holds potential for application in service-oriented sectors as well.

REFERENCE

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *30*(2), 37–46. https://doi.org/10.1145/376284.375668

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record*, *22*(2), 207–216. https://doi.org/10.1145/170036.170072

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of Eleventh International Conference on Data Mining*. https://doi.org/10.1016/j.jbi.2007.05.004

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-Based Learning Algorithms Instance-Based Learning Algorithms. *Machine Learning*, *6*, 37–66.

Alpaydin, E. (2014). *Introduction to Machine Lerning*. The MIT Press Cambridge, Massachusetts London, England.

Alt, F. B. (1976). Corrections to Papers by Montgomery and Klatt. *Management Science*, *22*(10), 1167–1168. https://doi.org/10.1287/mnsc.22.10.1167

Arnold, J. C., & Reynolds, M. R. (2001). CUSUM control charts with variable sample sizes and sampling intervals. *Journal of Quality Technology*, *33*(1), 66–81. https://doi.org/10.1080/00224065.2001.11980048

Bai, D. S., & Lee, K. T. (1998). An economic design of variable sampling interval $\bar{X}$ control charts. *International Journal of Production Economics*, *54*(1), 57–64. https://doi.org/10.1016/S0925-5273(97)00125-4

Barnard, G. A. (1959). Control Charts and Stochastic Processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 239–257. https://doi.org/10.1111/j.2517-6161.1959.tb00336.x

Beattie, D. W. (1962). A Continuous Acceptance Sampling Procedure Based Upon a Cumulative Sum Chart for the Number of Defectives. *Applied Statistics*, *11*(3), 137. https://doi.org/10.2307/2985360

Bennett, K. P., & Demiriz, A. (1999). Semi-supervised support vector machines. In *In M. Kearns & S. Solla, D. C. (Ed.).Advances in neural information processing systems 11* (pp. 368–374). Cambridge, MA: MIT Press. https://doi.org/10.1109/ICIEA.2014.6931500

Bersimis, S., Panaretos, J., & Psarakis, S. (2005). Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry. *Proceedings of the 7th Hellenic European Conference on Computer Mathematics and Its Applications,*. http://arxiv.org/abs/0901.2880

Bersimis, S., Sgora, A., & Psarakis, S. (2017). Methods for interpreting the out-of-control signal of multivariate control charts: A comparison study. *Quality and Reliability Engineering International*, *33*(8), 2295–2326. https://doi.org/10.1002/qre.2191

Bersimis, S., Sgora, A., & Psarakis, S. (2021). A robust meta-method for interpreting the out-of-control signal of multivariate control charts using artificial neural networks. *Quality and Reliability Engineering International*, 1–34. https://doi.org/10.1002/qre.2955

Bissell, A. F. (1969). Cusum Techniques for Quality Control. *Applied Statistics*, *18*(1), 1. https://doi.org/10.2307/2346436

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the Annual ACM Conference on Computational Learning Theory*, 92–100. https://doi.org/10.1145/279943.279962

Boran, S., & Diren, D. D. (2017). Analysis of out of control signals in multivariate processes with multilayer neural network. *Acta Physica Polonica A*, *132*(3), 1054–1057. https://doi.org/10.12693/APhysPolA.132.1054

Braga, A. C., Barros, C., Delgado, P., Martins, C., Sousa, S., Velosa, J. C., Delgado, I., & Sampaio, P. (2018). Multivariate statistical process control based on principal component analysis: Implementation of framework in R. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10961 LNCS*, 366–381. https://doi.org/10.1007/978-3-319-95165-2_26

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L., Jerome Fridman, Charles J. Stone, & R. A. Olshen. (1984). Classification and Regression Trees. In *Taylor and Francis*. https://doi.org/10.4018/9781591405573.ch027

Brighton, H., & Mellish, C. (2002). Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery*, *6*(2), 153–172. https://doi.org/10.1023/A:1014043630878

Broomhead, I. S., & Lowe, D. (1988). Radial Basis Functions, mutli-variable functionalinterpolation and adaptive networks. In *Royal signal & radar establishment*.

Cai, D. (2006). Support Tensor Machines for Text Categorization Support Tensor Machines for Text Categorization ∗. *Science*, *April*.

Carot, V., Jabaloyes, J. M., & Carot, T. (2002). Combined double sampling and variable sampling interval $\bar{X}$ chart. *International Journal of Production Research*, *40*(9), 2175–2186. https://doi.org/10.1080/00207540210128260

Caruana, R. (1997). Multitask Learning. *Machine Learning*, *28*(1), 41–75. https://doi.org/10.1023/A:1007379606734

Castagliola, P., Zhang, Y., Costa, A., & Maravelakis, P. (2012). The variable sample size X^{-} chart with estimated parameters. *Quality and Reliability Engineering International*, *28*(7), 687–699. https://doi.org/10.1002/qre.1261

Castelli, V., & Cover, T. M. (1996). The relative values of labelled and unlabelled samples in pattern recognition with an unknown mixing parameters. *IEEE Transactions on Information Theory1*, *42*(6,November).

Chen, Y., Wang, K., & Zhong, P. (2016). One-Class Support Tensor Machine. *Knowledge-Based Systems*, *96*, 14–28. https://doi.org/10.1016/j.knosys.2016.01.007

Cheng, H.-P., & Cheng, C. (2007). A Support Vector Machine for Recognizing Control Chart Patterns in Multivariate Processes. *The 5th Asian Network for Quality CongressAt: Incheon, Korea*, *October*. https://www.researchgate.net/publication/228958247

Chiu, W. K. (1974). Economic Design of Cusum Charts for Controlling Normal Means. *Journal of Applied Statistics*, *23*(3), 420–433. https://doi.org/10.2307/2347134

Chou, C. H., Kuo, B. H., & Chang, F. (2006). The generalized condensed nearest neighbor rule as a data reduction method. *Proceedings - International Conference on Pattern Recognition*, *2*, 556–559. https://doi.org/10.1109/ICPR.2006.1119

Costa, A. F. B. (1994). $\bar{X}$ charts with variable sample size. *Journal of Quality Technology*, *26*(3), 155–163. https://doi.org/10.1080/00224065.1994.11979523

Costa, A. F. B. (1997). $\bar{X}$ chart with variable sample size and sampling intervals.

Journal of Quality Technology, *29*(2), 197–204. https://doi.org/10.1080/00224065.1997.11979750
Costa, A. F. B. (1999). X̄ Charts with Variable Parameters. *Journal of Quality Technology*, *31*(4), 408–416. https://doi.org/10.1080/00224065.1999.11979947
Costa, A. F. B., & Rahim, M. A. (2013). Economic design of and R charts under Weibull shock models. *Communications in Statistics - Theory and Methods*, *42*(21), 3902–3925. https://doi.org/10.1080/03610926.2012.748914
Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *1IEEE Transactions on Information Theory*, *16*(2), 143–156. https://doi.org/10.4018/978-1-60960-557-5.ch005
Crosier, R. B. (1986). A New Two-Sided Cumulative Sum Quality Control Scheme. *Technometrics*, *28*(3), 187. https://doi.org/10.2307/1269074
Crosier, R. B. (1988). Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics*, *30*(3), 291–303.
Das, T. K., & Jain, V. (1997). An economic design model for X charts with random sampling policies. *IIE Transactions (Institute of Industrial Engineers)*, *29*(6), 507–518. https://doi.org/10.1080/07408179708966357
Daudin, J. J. (1992). Double Sampling X̄ Charts . *Journal of Quality Technology*, *24*(2), 78–87. https://doi.org/10.1080/00224065.1992.12015231
De Magalhães, M. S., Costa, A. F. B., & Epprecht, E. K. (2002). Constrained optimization model for the design of an adaptive X̄ chart. *International Journal of Production Research*, *40*(13), 3199–3218. https://doi.org/10.1080/00207540210136504
De Magalhães, M. S., Epprecht, E. K., & Costa, A. F. B. (2001). Economic design of a Vp X̄ chart. *International Journal of Production Economics*, *74*(1–3), 191–200. https://doi.org/10.1016/S0925-5273(01)00126-8
Demircioğlu Diren, D., Boran, S., & Cil, I. (2020). Integration of machine learning techniques and control charts in multivariate processes. *Scientia Iranica*, *27*(6 E), 3233–3241. https://doi.org/10.24200/SCI.2019.50377.1667
Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com
Dhini, A. and I. S. (2016). Review on Some Multivariate Statistical Process Control Methods for Process monitoring. *Proceeding of the 2016 International Conference on the Industrial Engineering and Operations Management*, *iid*, 754–759.
Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060
Doganaksoy, N., Faltin, F. W., & Tucker, W. T. (1991). Identification of out of control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics - Theory and Methods*, *20*(9), 2775–2790. https://doi.org/10.1080/03610929108830667
Duncan, A. J. (1971). The economic design of (Equation presented)-charts when there is a multiplicity of assignable causes. *Journal of the American Statistical Association*, *66*(333), 107–121. https://doi.org/10.1080/01621459.1971.10482230
Ewan, A. D. (1963). When and How to Use Cu-Sum Charts. *American Society for Qualty*, *5*(1), 1–22.

Ewan, W. D., & Kemp, K. W. (1960). Sampling Inspection of Continuous Processes with No Autocorrelation Between Successive Results. *Biometrika*, *47*(3/4), 363. https://doi.org/10.2307/2333307

Ferrer, A. (2007). Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Quality Engineering*, *19*(4), 311–325. https://doi.org/10.1080/08982110701621304

Fisher, D., & Langley, P. (1980). Approaches to Conceptual Clustering. *Ninth International Joint Conference on Artificial Intellegence*.

Flaig, J. J. (1991). Adaptive control charts. In *Keats, J.B. and Montgomery, D.C. (Eds.), Statistical Process Control in Manufacturing, Marcel Dekker, New York* (pp. 111–122).

Frank Rosenblatt. (1960). PROCEEDINGS OF THE IRE Perceptron Simulation Experiments* FRANK ROSENBLATT. *Proceeding of IRE*, *48*(3), 301–309.

Furnkranz, J. F. (1999). Separate-and-Conquer Rule Learning. *Artificail Intellegence Review*, *13*, 3–54.

Gan, F. F. (1991). An Optimal Design of CUSUM Quality Control Charts. *Journal of Quality Technology*, *23*(4), 279–286. https://doi.org/10.1080/00224065.1991.11979343

Gani, W., Taleb, H., & Limam, M. (2011). An assessment of the kernel-distance-based multivariate control chart through an industrial application. *Quality and Reliability Engineering International*, *27*(4), 391–401. https://doi.org/10.1002/qre.1117

García-Díaz, J. C., & Aparisi, F. (2005). Economic design of EWMA control charts using regions of maximum and minimum ARL. *IIE Transactions (Institute of Industrial Engineers)*, *37*(11), 1011–1021. https://doi.org/10.1080/07408170500232214

Ghanbari, A., Vaghei, Y., & Noorani, S. M. R. S. (2014). Reinforcement learning with neural networks: A survey. *International Journal of Advanced Biological and Biomedical Research2*, *2*(5), 1398–1416. https://doi.org/10.1007/978-3-319-30933-0_47

Gibra, I. N. (1975). Recent Developments in Control Chart Techniques. *Journal of Quality Technology*, *7*(4), 183–192. https://doi.org/10.1080/00224065.1975.11980695

Goel, A. L., & Wu, S. M. (1973). Economically Optimum Design of Cusum Charts. *Management Science*, *19*(11), 1271–1282. https://doi.org/10.1287/mnsc.19.11.1271

Grossberg, S. (1987). Competitive learning: From interactive Activation to Adaprtive Reseonance. *Cognitive Science1*, *11*, 23–63.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, *37*, 1–47. https://doi.org/10.1016/j.neunet.2012.09.017

Guh, R. S., & Shiue, Y. R. (2008). An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts. *Computers and Industrial Engineering*. https://doi.org/10.1016/j.cie.2008.01.013

Gupta, B. (2011). FP-Tree Based Algorithms Analysis : FP-Growth, COFI-Tree and CT-PRO. *International Journal on Computer Science and Engineering (IJCSE)*, *3*(7), 2691–2699.

Habert, G., Miller, S. A., John, V. M., Provis, J. L., Favier, A., Horvath, A., & Scrivener, K. L. (2020). Environmental impacts and decarbonization strategies in the cement and concrete industries. *Nature Reviews Earth and Environment*, *1*(11), 559–573. https://doi.org/10.1038/s43017-020-0093-3

Han, J., Pei, J., & Yan, X. (2004). Sequential Pattern Mining by Pattern-Growt: Principles and Extensions. *Journal of Computer Science and Technology*, *19*(3), 257–279.

Hand, D. J. (1998). Data Mining : Statistics and More ? *The American Statisticians*, *52*(2), 112–118.

Hart, P. E. (1968). The Condensed Nearest Neighbor Rule. *IEEE Transactions on Information Theory*, *14*(3), 515–516. https://doi.org/10.1109/TIT.1968.1054155

Hawkins, D. M. (1981). A Cusum for a Scale Parameter. *Journal of Quality Technology*, *13*(4), 228–231. https://doi.org/10.1080/00224065.1981.11978762

Hawkins, D. M. (1991). Multivariate Control Based on Quality Variables. *Technometrics*, *33*(1), 61–75.

Hawkins, D. M. (1993). Cumulative sum control charting: An underutilized spc tool. *Quality Engineering*, *5*(3), 463–477. https://doi.org/10.1080/08982119308918986

He, J., Zhao, G., Zhang, H. L., Ramamohanarao, K., & Pang, C. (2015). An effective clustering algorithm for auto-detecting well-separated clusters. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2015-Janua*(January), 867–874. https://doi.org/10.1109/ICDMW.2014.78

He, S., & Zhang, C. (2011). Support vector data description based multivariate cumulative sum control chart. *Advanced Materials Research*, *314–316*, 2482–2485. https://doi.org/10.4028/www.scientific.net/AMR.314-316.2482

Healy, J. D. (1987). A note on multivariate CUSUM procedures. *Technometrics*, *29*(4), 409–412. https://doi.org/10.1080/00401706.1987.10488268

Heikes, R. G., Montgomery, D. C., & Yeung, J. Y. H. (1974). Alternative process models in the economic design of t control charts. *AIIE Transactions*, *6*(1), 55–61. https://doi.org/10.1080/05695557408974933

Ho, C., & Case, K. E. (1994). Economic design of control charts: a literature review for 1981-1991. *Journal of Quality Technology*, *26*(1), 39–53. https://doi.org/10.1080/00224065.1994.11979497

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodology. *Artificail Intellegence Review*, *22*(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Hotelling, H. (1931). the Generalization Student ' S Ratio *. *The Annals of Mathematical Science*, *2*(3), 360–378.

Hotelling, H. (1933). Analysis of a complex statistical variables into principal components 8. Determination of principal components for individuals. *Journal of Educational Psychology*, *24*, 498–520.

Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, *042*, 23–41. http://projecteuclid.org/euclid.bsmsp/1200500217

IEA. (2018). Technology Roadmap: Low-Carbon Transition in the Cement Industry. *International Energy Agency,* 66.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., & C.Dubes, R. (1988). Algorithms for Clustring Data. In *Prentice Hall Englewood Cliffs, New Jersey 07632*. https://doi.org/10.1126/science.311.5762.765

Jaromczyk, J. W., & Toussaint, G. T. (1992). Relative Neighborhood Graphs and Their Relatives. *Proceedings of the IEEE*, *80*(9), 1502–1517. https://doi.org/10.1109/5.163414

John, V. M., Quattrone, M., Abrão, P. C. R. A., & Cardoso, F. A. (2019). Rethinking cement standards: Opportunities for a better future. *Cement and Concrete Research*, *124*(March), 105832. https://doi.org/10.1016/j.cemconres.2019.105832

Jolliffe, I. T. (2002). Principal Component Analysis. In *Springer Series in statistics*. https://doi.org/10.4324/9781351206358-1

Jones, L. A. (2002). The statistical design of EWMA control charts with estimated parameters. *Journal of Quality Technology*, *34*(3), 277–288. https://doi.org/10.1080/00224065.2002.11980158

Kazmer, D. O., Westerdale, S., & Hazen, D. (2008). A Comparison of Statistical Process Control (SPC) and On-Line Multivariate Analyses (MVA) for Injection Molding. *International Polymer Processing*, *23*(5), 447–458. https://doi.org/10.3139/217.2192

Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished Manuscript*, *45*, 105.

Kearns, M., & Valiant, L. (1989). Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Journal of the ACM (JACM)*, *41*(1), 67–95. https://doi.org/10.1145/174644.174647

Keats, J. B., Annadi, H. P., Montgomery, D. C., & Runger, G. C. (1995). An adaptive sample size CUSUM control chart. *International Journal of Production Research*, *33*(6), 1605–1616. https://doi.org/10.1080/00207549508930231

Kemp, K. W. . (1962). The Use of Cumulative Sums for Sampling Inspection Schemes. *Society, Royal Statistical*, *11*(1), 16–31.

Kesavaraj, G., & Sukumaran, S. (2013). A Study On Classification Techniques in Data Mining. *Fourth International Conference on Computing, Communications and Networking Technologies*.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (2001). Self-organizing maps. In *Substance Use and MisuseSpringer Series In Information Science2*. https://doi.org/10.3109/10826089809115871

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Koshti, V. V. (2011). Cumulative sum control chart (CUSUM). *International Journal of Physics and Mathematical Science*, *1*(1), 28–32. https://doi.org/10.1016/b978-0-323-90152-9.00016-5

Kostiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica2*, *31*(1), 249–268. https://doi.org/10.1007/s10751-016-1232-6

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, *28*(4), 409–428. https://doi.org/10.1080/00224065.1996.11979699

Kresta, J. V., Macgregor, J. F., & Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, *69*(1), 35–47. https://doi.org/10.1002/cjce.5450690105

Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. In *Wiley interscience*. https://doi.org/10.1198/tech.2005.s320

Lee, M. H., & Khoo, M. B. C. (2006). Optimal statistical designs of a multivariate CUSUM chart based on ARL and MRL. *International Journal of Reliability, Quality and Safety Engineering*, *13*(5), 479–497. https://doi.org/10.1142/S0218539306002380

Lee, T. H., Hong, S. H., Kwon, H. M., & Lee, M. (2016). Economic statistical design of variable sampling interval X control chart based on surrogate variable using genetic algorithms. *Management and Production Engineering Review*, *7*(4), 54–64. https://doi.org/10.1515/mper-2016-0036

Li, B., Morris, J., & Martin, E. B. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *64*(1), 79–89. https://doi.org/10.1016/S0169-7439(02)00051-5

Linderman, K., & Love, T. E. (2000). Economic and economic statistical designs for MEWMA control charts. *Journal of Quality Technology*, *32*(4), 410–417. https://doi.org/10.1080/00224065.2000.11980026

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lorenzen, T. J., & Vance, L. C. (1986). The economic design of control charts: A unified approach. *Technometrics*, *28*(1), 3–10. https://doi.org/10.1080/00401706.1986.10488092

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics*, *34*(1), 46. https://doi.org/10.2307/1269551

Lucas, J. M. (1976). the Design and Use of V-Mask Control Schemes. *Journal of Quality Technology*, *8*(1), 1–12. https://doi.org/10.1080/00224065.1976.11980706

Maboudou-Tchao, E. M. (2019). High-dimensional data monitoring using support machines. *Communications in Statistics: Simulation and Computation*, *50*(7), 1927–1942. https://doi.org/10.1080/03610918.2019.1588312

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414. https://doi.org/10.1016/0967-0661(95)00014-L

Maravelakis, P. E., Bersimis, S., Panaretos, J., & Psarakis, S. (2002). Identifying the out of control variable in a multivariate control chart. *Communications in Statistics - Theory and Methods*, *31*(12), 2391–2408. https://doi.org/10.1081/STA-120017232

Mason, R. L., Tracy, N. D., & Young, J. C. (1997). A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology*, *29*(4), 396–406. https://doi.org/10.1080/00224065.1997.11979791

Mason, R. L., & Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. the American Statistical Association and the Society for Industrial and Applied Mathematics.

Matthew Carlyle, W., Montgomery, D. C., & Runger, G. C. (2000). Optimization problems and methods in quality control and improvement. *Journal of Quality Technology*, *32*(1), 1–17. https://doi.org/10.1080/00224065.2000.11979963

Mcculloch, W. S., & Pitts, W. (1943). A logical calculus nervous activity. *Bulletin of Mathematical Biology*, *5*(l), 115–133.

Michalski, R. S. (1980). Knowledge Acquisition Through Conceptual Clustering: a Theoretical Framework and an Algorithm for Partitioning Data Into Conjunctive

Concepts. In *International Journal of Policy Analysis and Information Systems* (Vol. 4, Issue 3, pp. 219–244).

Miljković, Z., Mitić, M., Lazarević, M., & Babić, B. (2013). Neural network Reinforcement Learning for visual control of robot manipulators. *Expert Systems with Applications*, *40*(5), 1721–1736. https://doi.org/10.1016/j.eswa.2012.09.010

Miller, S. A., Habert, G., Myers, R. J., & Harvey, J. T. (2021). Achieving net zero greenhouse gas emissions in the cement industry via value chain mitigation strategies. *One Earth*, *4*(10), 1398–1411. https://doi.org/10.1016/j.oneear.2021.09.011

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Molnau, W. E., Montgomery, D. C., & Runger, G. C. (2001). Statistically constrained economic design of the Multivariate Exponentially Weighted Moving Average control chart. *Quality and Reliability Engineering International*, *17*(1), 39–49. https://doi.org/10.1002/qre.379

Montgomery, D. (2009). Introduction to statistical quality control. In *John Wiley & Sons Inc.* https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Montgomery, D. C. (1980). The Economic Design of Control Charts: A Review and Literature Survey. *Journal of Quality Technology*, *12*(2), 75–87. https://doi.org/10.1080/00224065.1980.11980940

Montgomery, D. C., & Klatt, piliips J. (1972). The Economic Design of X Charts Used to Maintain Current Control of a Process. *Management Science1*, *19*(1), 76–89. https://doi.org/10.1080/0013791X.1958.10131811

Montgomery, D. C., Torng, J. C. C., Cochran, J. K., & Lawrence, F. P. (1995). Statistically constrained economic design of the EWMA control chart. *Journal of Quality Technology*, *27*(3), 250–256. https://doi.org/10.1080/00224065.1995.11979597

Murphy, B. J. (1987). Selecting Out of Control Variables With the T 2 Multivariate Quality Control Procedure. *The Statistician*, *36*(5), 571. https://doi.org/10.2307/2348668

Murthy, S. S. ., & Rambabu, Y. (1997). Design and application of economical process control charts. *Defence Science Journal*, *47*(1), 45–53.

Namikka, E., & Gibbon, G. J. (2003). Identification of Data Mining Techniques for industrial process analysis and control. *" 1st African Control Conference AFCON. Cape Town*.

Nantawong, C., Randhawa, S. U., & McDowell, E. D. (1989). A methodology for the economic comparison of x¯, cumulative sum, and geometric moving-average control charts. *International Journal of Production Research*, *27*(1), 133–151. https://doi.org/10.1080/00207548908942535

Ning, X., & Tsung, F. (2013). Improved design of kernel distance-based charts using support vector methods. *IIE Transactions (Institute of Industrial Engineers)*, *45*(4), 464–476. https://doi.org/10.1080/0740817X.2012.712237

Noskievicova, D., Smajdorova, T., & Tyleckova, E. (2020). Statistical Process Control in Big Data Environment. *Proceedings of the 2020 21st International Carpathian Control Conference, ICCC 2020*. https://doi.org/10.1109/ICCC49264.2020.9257251

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*.

Page, E. S. (1961). cumulative sum charts. *The Grants Register 2023*, *3*(1), 1–9. https://doi.org/10.1057/978-1-349-96053-8_16494

Park, C., Lee, J., & Kim, Y. (2004). Economic design of a variable sampling rate EWMA chart. *IIE Transactions (Institute of Industrial Engineers)*, *36*(5), 387–399. https://doi.org/10.1080/07408170490426116

Park, C., & Reynolds, M. R. (1994). Economic design of a variable sample size -. *Communications in Statistics - Simulation and Computation*, *23*(2), 467–483.

Park, C., & Reynolds, M. R. (1999). Economic Design of a Variable Sampling Rate $\bar{X}$ Chart. *Journal of Quality Technology*, *31*(4), 427–443. https://doi.org/10.1080/00224065.1999.11979949

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pérez-Suárez, A., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2019). A review of conceptual clustering algorithms. *Artificial Intelligence Review*, *52*(2), 1267–1296. https://doi.org/10.1007/s10462-018-9627-1

Perry, M. B., Spoerre, J. K., & Velasco, T. (2001). Control chart pattern recognition using back propagation artificial neural networks. *International Journal of Production Research*, *39*(15), 3399–3418. https://doi.org/10.1080/00207540110061616

Pignatiello, J. J., & Runger, G. C. (1990). Comparisons of Multivariate CUSUM Charts . *Journal of Quality Technology*, *22*(3), 173–186. https://doi.org/10.1080/00224065.1990.11979237

Poudyal, L., & Adhikari, K. (2021). Environmental sustainability in cement industry: An integrated approach for green and economical cement production. *Resources, Environment and Sustainability*, *4*(December 2020), 100024. https://doi.org/10.1016/j.resenv.2021.100024

Prabhu, S. S., Montgomery, D. C., & Runger, G. C. (1994). Combined adaptive sample size and sampling interval $\bar{X}$ control scheme. *Journal of Quality Technology*, *26*(3), 164–176. https://doi.org/10.1080/00224065.1994.11979524

Prabhu, S. S., Montgomery, D. C., & Runger, G. C. (1997). Economic-statistical design of an adaptive $\bar{X}$ chart. *International Journal of Production Economics*, *49*(1), 1–15. https://doi.org/10.1016/S0925-5273(96)00100-4

Prabhu, S. S., & Runger, G. C. (1997). Designing a multivariate EWMA control chart. *Journal of Quality Technology*, *29*(1), 8–15. https://doi.org/10.1080/00224065.1997.11979720

Qin, S. J., & Chiang, L. H. (2019). Advances and opportunities in machine learning for process data analytics. *Computers and Chemical Engineering*, *126*, 465–473. https://doi.org/10.1016/j.compchemeng.2019.04.003

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Ranjan, G. S. K., Kumar Verma, A., & Radhika, S. (2019). K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, *March*. https://doi.org/10.1109/I2CT45611.2019.9033691

Re, M., & Valentini, G. (2012). Ensemble Methods: A Review. In *Advances in Machine Learning and Data Mining for Astronomy* (Issue May). https://doi.org/10.1201/b11822-20

Reynolds, M. R. (1989). Optimal Variable Sampling Interval Control Charts. *Sequential Analysis*, *8*(4), 361–379. https://doi.org/10.1080/07474948908836187

Reynolds, M. R. (1995). Evaluating Properties of Variable Sampling Interval Control Charts. *Sequential Analysis*, *14*(1), 59–97.

https://doi.org/10.1080/07474949508836320
Reynolds, M. R. (1996). Shewhart and EWMA variable sampling interval control charts with sampling at fixed times. *Journal of Quality Technology*, *28*(2), 199–212. https://doi.org/10.1080/00224065.1996.11979660
Reynolds, M. R., Amin, R. W., & Arnold, J. C. (1990). Cusum charts with variable sampling intervals. *Technometrics*, *32*(4), 371–384. https://doi.org/10.1080/00401706.1990.10484721
Reynolds, M. R., Amin, R. W., Arnold, J. C., Nachlas, J. A., Technometrics, S., May, N., & Reynolds, M. R. (1988). X charts with Variable Sampling Intervals. *Technometrics*, *30*(2), 181–192.
Reynolds, M. R., & Arnold, J. C. (2001). Ewma control charts with variable sample sizes and variable sampling intervals. *IIE Transactions (Institute of Industrial Engineers)*, *33*(6), 511–530. https://doi.org/10.1080/07408170108936850
Ritter, G. L., Woodruff, H. B., Isenhour, T. L., & Lowry, S. R. (1975). An Algoritm for a selective Neighbor Decision Rule. *IEEE Transactions on Information Theory*, *May 1969*.
Roberts, A. S. W., Technometrics, S., & Aug, N. (1959). American Society for Quality Control Chart Tests Based on Geometric Moving Averages American Society for Quality Stable URL : http://www.jstor.org/stable/1266443 Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, *1*(3), 239–250.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, *323*(9), 533–536. https://doi.org/10.7551/mitpress/1888.003.0013
Runger, G. C., & Pignatiello, J. J. (1991). Adaptive Sampling for Process Control. *Journal of Quality Technology*, *23*(2), 135–155. https://doi.org/10.1080/00224065.1991.11979304
Saccucci, M. S., Amin, R. W., & Lucas, J. M. (1992). Exponentially weighted moving average control schemes with variable sampling intervals. *Communications in Statistics - Simulation and Computation*, *21*(3), 627–657. https://doi.org/10.1080/03610919208813040
Saniga, E. M. (1989). Economic statistical control-chart designs with an application to x and r charts. *Technometrics*, *31*(3), 313–320. https://doi.org/10.1080/00401706.1989.10488554
Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1023/A:1022648800760
Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965
Shao, Y. E., & Lin, S. C. (2019). Using a time delay neural network approach to diagnose the out-of-control signals for a multivariate normal process with variance shifts. *Mathematics*, *7*(10). https://doi.org/10.3390/math7100959
Singh, P., & Meshram, P. A. (2018). Survey of density based clustering algorithms and its variants. *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, *Icici*, 920–926. https://doi.org/10.1109/ICICI.2017.8365272
Singh, Y. K. (2006). *Fundamental of Reseach Methodology and Statistics*. New Age International Publishers.
Singla, S., & Malik, A. (2014). Survey on various improved Apriori Algorithms. *Ijarcce*, *3*(11), 8528–8531. https://doi.org/10.17148/ijarcce.2014.31139

Skellam, J. G. (1952). Studies in Statistical Ecology: I. Spatial Pattern. *Biometrika*, *39*(3/4), 346. https://doi.org/10.2307/2334030

Sukchotrat, T., Kim, S. B., & Tsung, F. (2010). One-class classification-based control charts for multivariate process monitoring. *IIE Transactions (Institute of Industrial Engineers)*, *42*(2), 107–120. https://doi.org/10.1080/07408170903019150

Sun, R., & Tsung, F. (2003). A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, *41*(13), 2975–2989. https://doi.org/10.1080/1352816031000075224

Sutton, R. S., & Barto, A. G. (2015). An introduction to reinforcement learning. In *A Bradford Book*. https://doi.org/10.4018/978-1-60960-165-2.ch004

Tan, I. N. G., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education , Inc.

Tax, D. M. J., & Duin, R. P. W. (2004). Support vector data description. *Machine Learning*, *54*, 45–66.

Taylor, H. M. (1968). The Economic Design of Cumulative Sum Control Charts. *Technometrics*, *10*(3), 479–488. https://doi.org/10.1080/00401706.1968.10490595

Tjortjis, C., & Keane, J. (2002). T3: A classification algorithm for data mining. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2412*, 50–55. https://doi.org/10.1007/3-540-45675-9_9

Toussaint, G. T. (1980). the Relative neighbourhood Graph of a finite planar set. *Pattern Recognition1*, *12*, 261–268.

Trout, E. A. R. (2019). The history of calcareous cements. *Lea's Chemistry of Cement and Concrete*, 1–29. https://doi.org/10.1016/B978-0-08-100773-0.00001-0

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Vazan, P., Janikova, D., Tanuska, P., Kebisek, M., & Cervenanska, Z. (2017). Using data mining methods for manufacturing process control. *IFAC-PapersOnLine*, *50*(1), 6178–6183. https://doi.org/10.1016/j.ifacol.2017.08.986

Werbos, P. J. (1988). Back Propagation : past and future. *IEEE International Conference on Neural Network1*.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics*, *2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning*, *38*(3), 257–286. https://doi.org/10.1023/A:1007626913721

Woodall, W. H., & Adams, B. M. (1993). The statistical design of cusum charts. *Quality Engineering*, *5*(4), 559–570. https://doi.org/10.1080/08982119308918998

Woodall, W. H., & Matoteng, Nc. M. (1985). Multivariate CUSUM Control Quality-Procedures. *Technometrics*, *27*(3), 285–292.

Woodall, W. H., & Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, *46*(1), 78–94. https://doi.org/10.1080/00224065.2014.11917955

Xia, B., Jian, Z., Liu, L., & Li, L. (2018). An effective multivariate control chart for detecting small mean shifts using support vector data description. *Advances in Mechanical Engineering*, *10*(11), 1–18.

https://doi.org/10.1177/1687814018810625
Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1995-June*, 189–196. https://doi.org/10.3115/981658.981684
Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications. *Studies in Computational Intelligence*, *115*, 715–762. https://doi.org/10.1007/978-3-540-78293-3_17
Yu, F. J., Rahim, M. A., & Chin, H. (2007). Economic design of VSI $\bar{x}$ control charts. *International Journal of Production Research*, *45*(23), 5639–5648. https://doi.org/10.1080/00207540701325512
Zaki, M. J., & Wong, L. (2001). Data mining techniques. *Machne Learning*, *42*(1/2), 31–60. https://doi.org/10.1017/S0962492901000058
Zegeye, A., Worku, A., Tefera, D., Getu, M., & Sileshi, Y. (2009). Introduction to research methods. In *Addis Ababa University Press*. https://doi.org/10.4324/9781315213033-1
Zha, Z. J., Mei, T., Wang, J., Wang, Z., & Hua, X. S. (2009). Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, *20*(2), 97–103. https://doi.org/10.1016/j.jvcir.2008.11.009
Zhang, Y., Castagliola, P., Wu, Z., & Khoo, M. B. C. (2012). The variable sampling interval $\bar{X}$ chart with estimated parameters. *Quality and Reliability Engineering International*, *28*(1), 19–34. https://doi.org/10.1002/qre.1209
Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, *5*(1), 30–43. https://doi.org/10.1093/nsr/nwx105
Zhou, Z., & Li, M. (2005). co-training : COREG. *IJCAI Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 908–913.
Zhu, X., & Lafferty, J. (2005). Harmonic mixtures: Combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, 1057–1064.
Zimmer, L. S., Montgomery, D. C., & Runger, G. C. (1998). Evaluation of a three-state adaptive sample size X control chart. *International Journal of Production Research*, *36*(3), 733–743. https://doi.org/10.1080/002075498193660

APPENDIX

Major Accomplishments during the PhD Study

1. Publications

    - Daniel A., Daniel K., & Eshetie B.(2022). Design multivariate statistical process control procedure in the case of Ethio cement, *International Journal of Quality & Reliability Management* ,DOI:10.1108/IJQRM-07-2021-0227
    - Daniel A., Daniel K., & Eshetie B.(2022). Advances in statistical quality control chart techniques and their limitations to cement industry, *Cogent Engineering9(1)*, https://doi.org/10.1080/23311916.2022.2088463

2. Co-Advising MSc students

    - More than 10 students are advised during study period

3. Course Offering

    - Industrial Management and Engineering Economy
    - Materials Handling Equipment
    - Quality Management

4. Attend international Conferences

    - Global Cement Conference on Cement Process, energy, and optimization
    - International Cement Conference

5. International Visit to research center

    - A three month visit to University Technische Hochschule Nürnberg NCT research center, Germany.

6. Social Services

    - Teaching High School Students in SOS Children Village