POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

ONLINE ANOMALY DETECTION OF INDUSTRIAL PROCESSES AND MACHINERY BASED ON LOGICAL ANALYSIS OF DATA AND STATISTICAL CONTROL CHARTS

RAMY MOHAMMED KHALIFA MOHAMMED
Département de mathématiques et de genie industriel
École Polytechnique de Montréal

Thèse présentée en vue de l'obtention du diplôme de *Philosophiae Doctor*
Génie industriel

Juin 2022

© Ramy Mohammed Khalifa Mohammed, 2022.

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

ONLINE ANOMALY DETECTION OF INDUSTRIAL PROCESSES AND MACHINERY BASED ON LOGICAL ANALYSIS OF DATA AND STATISTICAL CONTROL CHARTS

présentée par **Ramy Mohammed Khalifa MOHAMMED**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Sofiane ACHICHE, président
Soumaya YACOUT, membre et directrice de recherche
Samuel BASSETTO, membre et codirecteur de recherche
Ali AIDIBE, membre
Vahid EBRAHIMIPOUR, membre externe

DEDICATION

To the memory of my mother Wafaa El Hakim,
To my sister Nora, my wife Shireen, and
my lovely and my twin son Aser
For their endless love support,
and encouragement. . ....

ACKNOWLEDGEMENTS

First and foremost, I feel always indebted to Allah, the most gracious and the most merciful.

I would like to express my heartfelt gratitude and appreciation to my supervisor Soumaya Yacout. I am deeply thankful for her valuable guidance with her extensive knowledge and support throughout my PhD journey. Although she is busy, she always takes out from her precious time to review my work or even answer my questions. She always encourages me and pushes me to my full potential and beyond to bring out the best of me. Without her continuous support and dedication to her work with all credibility and honesty, my articles and thesis would not have come out in such a professional manner. I learned a lot and I am really proud to be one of her students.

I wish to deeply thank my co-supervisor Samuel Bassetto for his encouragement, sharing experience and knowledge, and useful suggestions. He always supports me morally and scientifically. He always inspires me with new insights. He spared no effort to advise and guide me to complete my PhD. I will always be grateful.

Furthermore, I would like to extend my thanks to the PhD committee to review my thesis and attend my PhD defense enthusiastically.

I wish to thank the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ) and Rolls Royce as well as the NSERC funds, who financially supported the SARA project. I would like to express special thanks to Rose-Ann Cusson (Rolls-Royce Canada) for assistance with valuable suggestions and guidance which help me to overcome the challenges.

I will forever be thankful to Yasser Shaban for providing this opportunity to be one of Soumaya Yacout's students. He is always in contact with me for support and advice. Many thanks to my wonderful and generous colleagues who always encourage and support me during my PhD.

I would like to express my deep appreciation to my mother, who passed away while I was studying here, who sacrificed a lot to ensure her son's accomplishment today with her painstaking and encouragement during this long time. I would also like to thank my sister who has always supported me and pushed towards the success. And last but not least, my sincere gratitude to my beloved wife, who encouraged me and stands the life pressures and the circumstances of being away from her throughout the PhD period. She never tires of lending a hand to me and being patient in absorbing these circumstances. Finally, my beloved son and my twin, whom I have always sought in this life for his happiness and comfort.

RÉSUMÉ

De nos jours, les processus de fabrication évoluent pour intégrer les technologies de l'industrie 4.0, à savoir celles nécessaires à la connectivité, à l'analyse, à l'évolutivité et à la collecte de données en temps réel. Les techniques classiques de surveillance de processus pour surveiller la qualité du processus et/ou la qualité du produit sont confrontées à plusieurs défis. Des quantités massives de données et d'informations sont échangées via de nombreux capteurs et contrôleurs au sein du système. Ils doivent être structurés, gérés et stockés dans une structure de base de données prédéfinie qui définit et caractérise les informations requises et les fournit à chaque partie prenante. L'équipe qualité est l'un de ces acteurs à qui il incombe de contrôler la qualité du processus/produit à l'aide de cartes de contrôle. Cependant, les cartes de contrôle classiques présentent des limitations qui augmentent proportionnellement à la complexité du processus de fabrication. La qualité 4.0 a été introduite en tant que nouveau paradigme qui intègre les technologies de l'industrie 4.0 pour améliorer la surveillance et le contrôle de la qualité, ainsi que la détection des anomalies à l'aide de techniques d'apprentissage automatique. Néanmoins, certaines techniques d'apprentissage automatique ne permettent toujours pas de remédier à l'augmentation des fausses alarmes et des phénomènes de détection manquée. De plus, les techniques d'apprentissage automatique sont utilisées soit pour la détection d'anomalies, soit pour l'identification d'anomalies. Ils ont besoin de suffisamment de données pour pouvoir identifier et détecter les anomalies.

Dans cette thèse, nous développons un modèle de données conceptuel et logique à l'aide d'un outil Entity-Relationship Modeling (*ERM*). Le modèle *ERM* définit toutes les informations requises par les parties prenantes. Il ingère, collecte, stocke, organise, nettoie, intègre, protège et maintient les données générées au sein de la fabrication dans une base de données structurée prédéfinie. Le modèle est facile à utiliser par les parties prenantes et garantit une qualité élevée des données. Nous avons utilisé l'*ERM* dans un cas réel pour gérer les données et informations pertinentes des processus d'inspection et de réparation dans le domaine de la maintenance aérospatiale. Les données stockées dans la base de données représentent les données historiques utilisées pour définir les indices de performance clés de la qualité d'un processus.

Le modèle de régression d'analyse logique des données (*LADR*) a été développé sur la base de modèles extraits à l'aide d'une approche *LAD* commune à appliquer aux problèmes de régression. Le modèle *LADR* est construit à l'aide de modèles extraits des données d'origine au lieu des variables indépendantes d'origine. Il a été intégré à la carte de contrôle pour

obtenir une nouvelle carte de contrôle basée sur un modèle. Il améliore la sensibilité de détection des anomalies. Contrairement à d'autres techniques d'apprentissage automatique, il exploite ses modèles pour effectuer une analyse des causes profondes de l'anomalie détectée.

La carte de contrôle basée sur *LADR* a été adoptée pour développer un nouveau mécanisme de surveillance et d'alerte en ligne. Ce mécanisme a été appliqué pour surveiller les conditions de fonctionnement du système d'entraînement par courroie. Des expériences ont été menées sur le système pour collecter ses signaux vibratoires uniquement en fonctionnement normal. Le mécanisme utilise les caractéristiques statistiques extraites des signaux collectés pour détecter et identifier toute anomalie rencontrée lors du fonctionnement du système.

La carte de contrôle basée sur *LADR* a produit de meilleures performances que d'autres cartes de contrôle bien connues basées sur l'apprentissage automatique. Il a réduit le taux de fausses alarmes et le taux de détections manquées par des pourcentages minimum de 95% et 50%, respectivement, par rapport aux approches actuelles.

ABSTRACT

Nowadays, manufacturing processes are changing to integrate industry 4.0 technologies, namely those needed for connectivity, analytics, scalability, and gathering real time data. The conventional process monitoring techniques for monitoring the process quality and/or product quality are facing several challenges. Massive amounts of data and information is exchanged through many sensors and controllers within the system. They are required to be structured, managed, and stored in a pre-defined database structure that defines and characterizes the required information and provides it to each stakeholder. The quality team is one of these stakeholders, whose responsibility it is to monitor the quality of the process/ product by using control charts. However, the conventional control charts have limitations that increase proportionally with the complexity of the manufacturing process. Quality 4.0 has been introduced as a new paradigm that integrates the industry 4.0 technologies to improve quality monitoring and control, and anomaly detection by using machine learning techniques. Nevertheless, some machine learning techniques still do not remedy the increase in false alarms and missed detection phenomena. Moreover, machine learning techniques are used either for anomaly detection or anomaly identification. They require sufficient data to be able to identify and detect anomalies.

In this thesis, we develop a conceptual and logical data model using an Entity-Relationship Modeling (*ERM*) tool. The *ERM* model defines all information required by stakeholders. It ingests, collects, stores, organizes, cleanses, integrates, protects, and maintains the generated data within manufacturing in a pre-defined structured database. The model is easy to use by stakeholders, and it ensures high quality of data. We used the *ERM* in a real case to manage the relevant data and information of the inspection and repair processes in the aerospace maintenance domain. The data stored in the database represents the historical data that is used to define the key performance indices of the quality of a process.

The Logical Analysis of Data regression (*LADR*) model was developed based on extracted patterns using a common *LAD* approach to be applied on regression problems. The *LADR* model is constructed using patterns extracted from the original data instead of the original independent variables. It has been integrated with the control chart to obtain a new model-based control chart. It improves the sensitivity of anomaly detection. Unlike other machine learning techniques, it exploits its patterns to perform a root cause analysis of the detected anomaly.

The *LADR*-based control chart was adopted to develop a new online condition monitoring

and warning mechanism. This mechanism was applied to monitor the operating conditions of the belt drive system. Experiments were carried out on the system to collect its vibration signals only during normal operation. The mechanism uses the extracted statistical features from the collected signals to detect and identify any anomaly that is experienced during the operation of the system.

The *LADR*-based control chart produced better performance than other well-known machine learning-based control charts. It reduced the rate of false alarms and the rate of missed detections by minimum percentages of 95% and 50%, respectively, regarding the current approaches.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

RÉSUMÉ v

ABSTRACT vii

TABLE OF CONTENTS ix

LIST OF TABLES xii

LIST OF FIGURES xiv

LIST OF SYMBOLS AND ACRONYMS xvii

LIST OF APPENDICES xix

CHAPTER 1 INTRODUCTION 1
- 1.1 Problem statement 3
- 1.2 General objective 4
- 1.3 Specific objectives 5
- 1.4 Research approach 6
- 1.5 Originality of Research 7

CHAPTER 2 THESIS ORGANIZATION 10
- 2.1 Deliverables 11

CHAPTER 3 LITERATURE REVIEW 13
- 3.1 Data management 13
  - 3.1.1 Data Types 14
  - 3.1.2 Data Modeling 14
  - 3.1.3 Entity Relationship Modeling (*ERM*) 18
- 3.2 Control charts 19
  - 3.2.1 Limitations of control charts 22
- 3.3 Machine Learning Based Control Charts 24

CHAPTER 4 ARTICLE 1: QUALITY 4.0: ENTITY RELATIONSHIP MODEL FOR INSPECTION AND REPAIR PROCESSES IN AEROSPACE DOMAIN 26
4.1 Abstract 27
4.2 Introduction 27
4.3 ER model Technique 29
4.4 Case study: Application on inspection and repair in aerospace domain 31
4.4.1 Description of the case study 31
4.4.2 Business requirement for SARA system 32
4.4.3 Conceptual ER model for SARA system 32
4.4.4 Logical ER model for SARA system 35
4.5 Discussion 40
4.6 Conclusion 41

CHAPTER 5 ARTICLE 2: DEVELOPING MACHINE-LEARNING REGRESSION MODEL WITH LOGICAL ANALYSIS OF DATA (LAD) 42
5.1 Abstract 43
5.2 Introduction 43
5.3 *LADR* regression 46
5.3.1 *LADR* methodology 47
5.4 Performance of the *LADR* 56
5.5 Validation of the *LADR* 71
5.6 Numerical application 74
5.7 Conclusion 76

CHAPTER 6 ARTICLE 3: ROOT CAUSE ANALYSIS OF AN OUT-OF-CONTROL PROCESS USING A LOGICAL ANALYSIS OF DATA REGRESSION MODEL AND EXPONENTIAL WEIGHTED MOVING AVERAGE 78
6.1 Abstract 79
6.2 Introduction 79
6.3 Literature review 80
6.4 Methodology 83
6.4.1 Overview of the *LADR* technique 84
6.4.2 *LADR* regression-based control chart 91
6.4.3 Root cause identification of the out-of-control process 94
6.5 Numerical example 96
6.5.1 Dataset description 96
6.5.2 Development of the *LADR* regression models 99

6.5.3 Construction of the control charts 100
6.5.4 Results of the $EWMA$ charts 100
6.5.5 Diagnosis of the root cause of the out-of-control signals 104
6.6 Conclusion 107

CHAPTER 7 ARTICLE 4: CONDITION MONITORING AND WARNING MECHANISM IN THE BELT DRIVE SYSTEM BASED ON *LADR* BASED RESIDUAL CONTROL CHART 108
7.1 Abstract 109
7.2 Introduction 109
7.3 Experimental Study 112
7.3.1 Experiment test-rig Description 112
7.3.2 Measurements and Data Description 114
7.4 Methodology 116
7.4.1 LADR regression technique 117
7.4.2 Residual Control Chart 121
7.4.3 Condition Monitoring and Warning Mechanism 122
7.5 Experimental case study to evaluate the mechanism 123
7.5.1 Multiple linear regression (MLR) 124
7.5.2 Support Vector Regression (SVR) 124
7.5.3 Random Forest Regression (RF) 125
7.6 Results and Discussions 126
7.7 Conclusion 131

CHAPTER 8 GENERAL DISCUSSION 133

CHAPTER 9 CONCLUSION AND RECOMMENDATIONS 135
9.1 Summary of Works 135
9.2 Future Research 137

REFERENCES 138

APPENDICES 158

LIST OF TABLES

Table 5.1 Addressing the research gaps 46
Table 5.2 The thresholds that are identified using *EW*, *KM*, 15%*STD*, and *QT* with step 0.2 method 50
Table 5.3 Three tables defining the positive and negative classes that are obtained by the first three thresholds of the dataset from table 5.2, when using the 15%*STD* method. (Class=0 :Positive observations , Class=1: Negative observations) 51
Table 5.4 The generated patterns at the first three thresholds of the 15%*STD* method for the illustrative dataset and the binary values of the patterns' independent variables X_{P_j},j=1,...,10 53
Table 5.5 Characteristics of the four datasets 56
Table 5.6 The performance of the regression models for *Boston Housing* 60
Table 5.7 The performance of the regression models for *Computer Hardware* 71
Table 5.8 The performance of the regression models for *Auto-mpg* 71
Table 5.9 The performance of the regression models for *Servo* 71
Table 5.10 The performance of the regression models for *Airfoil Self-Noise* 72
Table 5.11 The performance of the regression models for *Concrete Strength* 72
Table 5.12 Friedman test for the best *LADR* model using the four discretization methods for all datasets 74
Table 5.13 Friedman test for the best *LADR* model and other regression models for all datasets 74
Table 5.14 A comparison between the performance of *LADR* and CR 75
Table 5.15 A comparison between the performance of *LADR* and LR 76
Table 6.1 Illustrative example for the steps in the methodology 85
Table 6.2 The classes and thresholds using the *KM* method for the illustrative example 87
Table 6.3 Defining the positive and negative classes for the first threshold (τ_1) using the *KM* method. 87
Table 6.4 The generated patterns at the first two thresholds of the *KM* method for the illustrative dataset and the binary values of the patterns' independent variables X_{P_j}, j=1,...,5 89
Table 6.5 The pattern's covered zones and classes 91
Table 6.6 Generation of special cause 93

Table 6.7 The performance of the $LADR$ models for the Concrete manufacturing process 99
Table 6.8 The $FAR\%$ and $MDR\%$ for the regression based EWMA methods ... 104
Table 6.9 The pattern's covered zone and class for the 32nd measurement sample 105
Table 6.10 Concrete manufacturing: The classes, zones, and thresholds using the EW method 106
Table 7.1 Samples of the data used 118
Table 7.2 The performance of regression models 126
Table 7.3 The Patterns of the *LADR-KM* model 127
Table A.1 Boston housing dataset 158
Table A.2 Computer Hardware dataset 159
Table A.3 Auto-MPG dataset 159
Table A.4 Servo dataset 159
Table A.5 Airfoil Self-Noise dataset 160
Table A.6 Concrete Compressive Strength dataset 160
Table C.1 Files and figures description 171

LIST OF FIGURES

Figure 3.1 Quality 4.0 framework 13
Figure 3.2 The stages of Data modeling 15
Figure 3.3 Data modeling approaches 17
Figure 3.4 The notations of the *ERM* 19
Figure 3.5 The bridge from data modeling to quality monitoring 19
Figure 3.6 The characteristics of the Control chart 20
Figure 3.7 Type *I* & *II* errors 22
Figure 4.1 Three axes of Database management 28
Figure 4.2 Data modeling stages 29
Figure 4.3 The notations for ER model 30
Figure 4.4 Inspection and Repair stages 31
Figure 4.5 ER Modeling design steps 32
Figure 4.6 Conceptual ER Modeling for SARA system 34
Figure 4.7 Logical ER Modeling for SARA system 38
Figure 4.8 EQUIPMENT entity in the SARA Model 39
Figure 4.9 ORDER entity in the SARA Model 40
Figure 5.1 Steps of a *LAD* approach 46
Figure 5.2 Difference between *LAD* and linear regression form 47
Figure 5.3 A diagram of the *LADR* methodology 48
Figure 5.4 Data processing steps 53
Figure 5.5 *Boston Housing* : the *LADR* measures of performance using the *KM* method 58
Figure 5.6 *Boston Housing* : the *LADR* measures of performance using the *EW* method 58
Figure 5.7 *Boston Housing* : the *LADR* measures of performance using the *%STD* method 59
Figure 5.8 *Boston Housing* : the *LADR* measures of performance using the *QT* method 59
Figure 5.9 *Computer Hardware*: the *LADR* measures of performance using the *KM* method 61
Figure 5.10 *Computer Hardware*: the *LADR* measures of performance using the *EW* method 61

Figure 5.11 *Computer Hardware*: the $LADR$ measures of performance using the $\%STD$ method 62
Figure 5.12 *Computer Hardware*: the $LADR$ measures of performance using the QT method 62
Figure 5.13 *Auto-mpg*: the $LADR$ measures of performance using the KM method 63
Figure 5.14 *Auto-mpg*: the $LADR$ measures of performance using the EW method 63
Figure 5.15 *Auto-mpg*: the $LADR$ measures of performance using $\%STD$ method . 64
Figure 5.16 *Auto-mpg*: the $LADR$ measures of performance using the QT method 64
Figure 5.17 *Servo*: the $LADR$ measures of performance using the KM method . . . 65
Figure 5.18 *Servo*: the $LADR$ measures of performance using the EW method . . . 65
Figure 5.19 *Servo*: the $LADR$ measures of performance using the $\%STD$ method . 66
Figure 5.20 *Servo*: the $LADR$ measures of performance using the QT method 66
Figure 5.21 *Airfoil self-noise*: the $LADR$ measures of performance using the KM method 67
Figure 5.22 *Airfoil self-noise*: the $LADR$ measures of performance using the EW method 67
Figure 5.23 *Airfoil self-noise*: the $LADR$ measures of performance using the $\%STD$ method 68
Figure 5.24 *Airfoil self-noise*: the $LADR$ measures of performance using the QT method 68
Figure 5.25 *Concrete strength*: the $LADR$ measures of performance using the KM method 69
Figure 5.26 *Concrete strength*: the $LADR$ measures of performance using the EW method 69
Figure 5.27 *Concrete strength*: the $LADR$ measures of performance using the $\%STD$ method 70
Figure 5.28 *Concrete strength*: the $LADR$ measures of performance using the QT method 70
Figure 6.1 The $LADR$ methodology flow chart 86
Figure 6.2 Anomaly detection using $LADR - EWMA$ chart 93
Figure 6.3 The prevalence for each pattern per each class 95
Figure 6.4 Concrete manufacturing: *LADR-EW* based $EWMA$ chart for Phase I 101
Figure 6.5 Concrete manufacturing: *LADR-EW* based $EWMA$ chart for Phase II 102
Figure 6.6 Concrete manufacturing: LR based $EWMA$ chart 102
Figure 6.7 Concrete manufacturing: SVR based $EWMA$ chart 103
Figure 6.8 Concrete manufacturing: PLS based $EWMA$ chart 103

Figure 6.9 Concrete manufacturing: $MARS$ based EWMA chart 104
Figure 6.10 Concrete manufacturing: The prevalence for each pattern in class C4 106
Figure 7.1 Test-rig Description 113
Figure 7.2 The pretension gauge 114
Figure 7.3 Time and Frequency domains for the normal operation of belt drive system at N=1000 RPM and T=70 N 117
Figure 7.4 Schematic of proposed condition monitoring and warning Mechanism 123
Figure 7.5 Random Forest Regression model 125
Figure 7.6 The pattern's prevalence for each class 128
Figure 7.7 The *LADR-KM-RCC* for belt drive system 129
Figure 7.9 The *MLR-RCC* for belt drive system 129
Figure 7.8 The *SVR-RCC* for belt drive system 130
Figure 7.10 The *RF-RCC* for belt drive system 130
Figure C.1 G.U.N.T machinery diagnostic system (PT 500.14) Description 172
Figure C.2 The pretension gauge 172
Figure C.3 G.U.N.T (PT 500.14) - the healthy and faulty belts 173
Figure C.4 G.U.N.T (PT 500.14) - presence of unbalanced weights 173

LIST OF SYMBOLS AND ACRONYMS

ANN	Artificial Neural Networks
ARL	Average Run Length
ASD	Adaptive Step-Down approach
CL	Centerline
CMT	Condition Monitoring Techniques
CR	Combinatorial Regression
CSLV	Cycle Since Last Visit
CSN	Cycle Since New
CSO	Cycle Since Overhaul
DBMS	Database Management System
DTR	Decision Tree Regression
EN	Elastic net
ERM	Entity relationship model
EW	Equal Width
EWMA	Exponential Weighted Moving Average
ID	Indentifier
KM	Kmeans
KNN	K-Nearest Neighborhood
LAD	Logical Analysis of Data
LADR	Logical Analysis of Data Regression
LAR	Least Absolute Residual
LASSO	Least Absolute Shrinkage and variable Selection
LCL	Lower Control Limit
LP/N	Last Part Number
LR	Linear Regression
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
MSE	Mean Square Error
MYT	Mason, Young, and Tracy decomposition
P.C.	Point Cloud
P/N	Part Number
PBR	Pseudo Boolean Regression model
PCA	Principal Component Analysis

PLS	Partial Least Square
PolyR	Polynomial Regression
QT	Quantile method
R^2	Coefficient of Determination
RCC	Residual Control Chart
RF	Random Forest
RMS	Root Mean Square
S/N	Serial Number
SARA	"Système d'Analyse et de Réparation Automatisée" (Automated inspection, Analysis and Repair System)
SPC	Statistical Process Control
SVM	Support Vector Machine
SVR	Support Vector Regression
TSLV	Time Since Last Visit
TSN	Time Since New
TSO	Time Since Overhaul
UCL	Upper Control Limit
V/N	Visit Number
VIF	Variance Inflation Factor
VMT	Vibration-based Monitoring Techniques
WPD	Wavelet packet decomposition

LIST OF APPENDICES

Appendix A *UCI* Datasets 158

Appendix B *LADR* models 161

Appendix C ARTICLE 5: EXPERIMENTAL VIBRATION DATA COLLECTED FOR A BELT DRIVE SYSTEM UNDER DIFFERENT OPERATING CONDITIONS 166

CHAPTER 1 INTRODUCTION

A few years ago, a study was carried out comparing automobile transmissions that were manufactured in the USA and Japan. This comparative study showed there was a big difference between the two countries based on an analysis of performance and repair cost. Random transmission samples were selected from the two sources and their quality characteristics were measured as well. At that time, Japan outperformed the USA because they had higher quality transmissions at a lower cost [1].

The quality of products is the main concern in industrial applications to achieve consumers' requirements. It is one of the most significant factors that creates fierce competition between products [2]. Accordingly, when the quality is improved, the performance of the business will be enhanced. In addition, it increases not only the process productivity but also the safety process and reliability of the system. Garvin [3] described the evaluation of the quality of any product as follows:

1. An evaluation of the product's performance according to required specifications.
2. Identification of the product's reliability, durability and serviceability.
3. Determination of the product's aesthetics and additive features.
4. The reputation of the company and the perceived quality of the product.
5. Conformance to regulations and standards.

Industries seek a level of quality, in which their processes or products can achieve a required target level. However, some variations can appear during this process. High quality means a reduction in variations so that they are within an acceptable range, closer to the target. In other words, the higher quality the product, the greater the reduction in variability in both the processes and products. This refers to the previously mentioned comparative study on automobile transmissions, in which the products' statistical quality distribution of the two countries had the same target mean, but the width of the distribution variance (i.e. normal distribution of a high variance) in the USA is greater than that in Japan. The results in the USA had greater variability, which corresponds to lower quality. Accordingly, non-conforming products were accepted by American companies with defective characteristics. Furthermore, the transmissions from American automobiles needed too many repairs, more rework and greater effort as well as wasted time and cost.

There are two types of process variations: common cause variations and assignable cause variations. Common cause variations are the natural variability in any process that has no significant effect on the stability of the process. These variations are due to unavoidable causes that can be eliminated only by better product or process design; nevertheless, the process is statistically controlled. On the other hand, the assignable cause variations result in abnormal variability, which affects the process performance. These can be caused by the presence of machine faults, non-conforming raw material, or operator errors. Thus, the process is statistically out of control [4, 5].

Thus, quality enhancement is achieved by monitoring, analyzing, and controlling process variability. Process monitoring and control are constructed in three stages: (1) fault detection of any abnormal patterns; (2) fault diagnosis to understand a pattern; (3) applying corrective actions to bring the process back to normal conditions.

Statistical process control (SPC) is one of the methods that monitors the process to reduce variability and improve quality [6]. Once SPC detects any assignable causes associated with the process, it provides a notification of the presence of abnormal variations. Hence, a corrective decision can be taken to avoid loss of quality. SPC contains seven major tools, which are often called the "magnificent seven" [7]. The control chart is an SPC tool that is used in many industrial applications [8, 9]. They are used in a statistical hypothesis to monitor the variability of quality characteristics. Nevertheless, the control chart faces several limitations due to an increase in the complexity of manufacturing processes.

Recently, the Quality 4.0 paradigm has been introduced under the title of Industry 4.0 which digitalizes quality management using artificial learning techniques [10]. Several studies have implemented machine learning techniques with control charts to overcome their drawbacks, such as the sensitivity of the results to the chosen parameters and the need to determine accurate control limits, autocorrelation in the dataset, and the difficulty of handling higher-dimensional data. These drawbacks have led to false alarms, which are called false positives, and/or missed fault detection, which are called false negatives. Hence, these misleading results affect process monitoring, and therefore the product or process quality declines. Various machine learning techniques have been used for feature extraction and selection to maintain the important variables that are required to be monitored by the chart. Consequently, the performance of the control charts in the detection of assignable causes has shown some improvement compared to before. Furthermore, machine-learning techniques have helped control charts identify the type of anomaly that causes abnormal behavior. Moreover, they have identified the variables that have contributed to that fault.

Machine learning enriched the control charts and increased their performance in fault detec-

tion. However, the latest research shows that there is still some need to improve false alarms and/or missed detection rates. This also applies to the accuracy of the model variables and determines the parameters that are calculated from the data. In other words, a more accurate model results in a better description of the process data.

Data management is considered an essential aspect of Quality 4.0 which is carried out before the implementation of machine learning algorithm-based control charts. Data management is used to establish, communicate, demonstrate, and invest in a unified data vision [11]. It identifies the data type generated from the manufacturing processes, then applies cleaning for invaluable resources to ensure the quality of the given data. Therefore, a data model is developed to provide all of the necessary data to easily monitor the quality characteristics of a manufacturing process.

1.1 Problem statement

Quality is considered to be a crucial aspect of both processes and products and is a competitive advantage for various industrial organizations and companies in the global market. Most industries have now realized that maintaining the quality of a process and/or product is not an option. They strive towards improving quality by reducing the variability in both the processes and products to meet customer requirements and to conform to standards, as well as increase productivity performance. Control charts are used for process monitoring and quality control to detect the anomalies experienced in the process and, accordingly, improve the quality of the process and product. An increase in the volume and variety of data that is automatically exchanged through the sensors and controllers in a system, increases the complexity of the manufacturing processes. The conventional control chart has several challenges that affect decision-making.

Recent advances in modern technologies in various industrial applications have led to continuous and large data streams that are collected via the data acquisition systems [12, 13]. The collected data is stored in complicated and varying structures with different formats, which is difficult to be implemented directly to control charts or even use in other analyses. Data preparation and management is a critical operation, which enables the efficient and effective use of data in order to become easy to identify, understand and manage the storage of historical data and current data, and to identify its location for traceability, accessibility, and reusability. Therefore, a data model is compulsorily required to be designed to integrate and define the necessary process variables in a unified data structure. Consequently, these variables will be easily monitored by the control chart [14].

Despite improvements in the performance of control charts, several limitations have become more apparent as the manufacturing processes have become more complex. False alarms and missed detections are two critical problems that are presented within control charts. The control charts' performance deteriorates with the presence of auto-correlated process data, which affects the accuracy of the chart parameters [15]. Also, the charts' performance declines when they are subject to high-dimensional data [16]. However, machine learning techniques contribute to solving the drawbacks of control charts. Different models are generated to reduce the dimensionality of the dataset, which has a large number of variables. Thus, a few dependent variables are monitored via the control charts. Also, regression models manage the auto-correlated data, so the process can be accurately monitored. There are existing techniques that still produce false alarms and/or missed faults. It is essential to obtain an accurate technique structure that will lead to a better description of the process data. Accordingly, this is reflected in the selection of the chart parameters that lead to high performance of fault detection.

The identification of an anomaly experiences in the manufacturing process is important for decision-making to bring the process back to its normal operation. The control charts do not determine the root causes when an anomaly is detected. Several fault isolation techniques are used to identify the faulty variables that contribute to an anomalous process. Although these techniques overcome the drawbacks of existing methods, they still suffer from a lack of handling higher dimensional process data that contain high correlations [17, 18]. Some techniques work well when there is one faulty variable and others assume that a few variables are responsible for that shift in the process [19]. Generally, these techniques acquire enough historical fault data or randomly generate training data that describes the different types of anomalies [20–23].

1.2 General objective

This thesis aims to provide quality leaders with versatile tools to deploy a Quality 4.0 transformation in terms of data management and analytics to improve the conventional process quality monitoring and control tools. This set of tools are are oriented towards two main parts, data and information flow within the manufacturing system, and improving of the quality of the process and product. A data model is proposed and built to support the digital transformation strategy of data management in manufacturing processes. It is used to establish a structured plan that favours information exchange to improve traceability and to allow the data flow to be mapped within the process. Subsequently, the thesis proposes a new machine learning technique that is integrated with control chart to monitor process variabil-

ity. It overcomes the drawbacks of existing approaches. The development of a compatible technique increases the control chart's performance in anomaly detection and performing root cause analysis. The proposed tools ensure well structured and descriptive database that easily define the key performance of indices for all disciplines. Moreover, the tools increases the quality of the process and product that conform standard and specification, and improves the productivity.

1.3 Specific objectives

In order to achieve the general objective, several tools were developed to overcome the diverse and heterogeneous of the captured data from the manufacturing processes, and improve process quality monitoring in terms of anomaly detection and identification. The objectives are as follows:

Objective 1: Design an approach for data modeling

Design and build a data model that organizes and manages the diverse and heterogeneous data exchange through the sensors and controllers within a process. The proposed data model captures and stores the data in a pre-defined structure and ensures a high quality of data. It favours accessibility, traceability, and reproducibility that will help for better communication between stakeholders, specifically for the quality department and to track relevant information. This structure of the data model can result in the conservation of human expertise for further data analysis, exploration, and exploitation. More importantly, the data model will provide the key performance of index for a process that represents the stage of data preparation for the next objective.

Objective 2: Develop an accurate machine learning technique for monitoring process quality

Develop a machine learning technique that provides an accurate model structure with high performance. This technique exploits the historical data captured from the manufacturing and obtains an inferential model for the monitored process. The structure of the technique is considered the extracted patterns that describe the original data. The independent variables are the patterns with the same dependent variable(s). Therefore, the key feature is to obtain better independent variables that are more interpretable to describe the key performance of a quality index and understand process conditions. This technique will be integrated with the control charts in the next objective. The integration is used for anomaly detection and identification in a process to ensure a high quality of that process.

Objective 3: Reduce false alarms and missed detections

Develop a new model-based control chart that monitors the variability within a process, re-

duces the false alarm and missed detection rates, and overcomes the limitations of the control charts. High false alarms and missed detections are crucial problems that are inherent within the control charts. Integrating a machine learning technique increases the sensitivity of the control charts. They are implemented to obtain a regression model describing the relationship between the independent variables and the dependent variable. The proposed approach combines the proposed machine learning technique in objective 2 with the conventional control chart to improve the performance of detecting anomalies that can be presented in a process. The proposed approach reduces false alarm and missed detection rates by minimum percentages of 95% and 50%, respectively, regarding the current approaches. This reduces the downtime and improves the productivity of the manufacturing.

Objective 4: Identify the root causes of the anomaly with pattern recognition

Use the same machine learning technique developed to identify the root cause of the anomaly once it is detected. This technique can determine the anomaly's reason via the patterns, which are extracted from the process data.

1.4 Research approach

We established a research approach to meet the research objectives previously mentioned. This approach is as follows:

1. **Data Modeling:** Data preparation and management is a critical operation, which enables the efficient and effective use of data in order to become easy to identify and understand, and to manage the storage of historical data and current data. We proposed a data modeling approach using Entity-Relationship Modeling (*ERM*) technique to collect, store, organize, clean, integrate, and protect the diverse and heterogeneous data exchange through sensors and controllers within a process. The *ERM* provides well-structured database and improves the data quality. It provides the stakeholders' information easily and rapidly, and standardizes the communications between them. Furthermore,Provide the key performance indices for all stakeholders, especially quality team.

2. **Predictive Modeling:** Machine learning technique is used to obtain the relationship between the dependent and independent variables. It uses the historical data in the database to create a regression model which is used to predict accurately the online real-time data. Thus, we developed a Logical Analysis of Data Regression technique (*LADR*) as a new regression technique to obtain an accurate model that describes

the process data. The *LADR* is based on a standard *LAD* methodology [24]. It is constructed based on extracted hidden patterns in the original process data. The *LADR* handles the curse of dimensionality and auto-correlation phenomenon.

3. **Anomaly Detection:** Since the control charts have several limitations, we integrate the proposed *LADR* with the control chart to monitor quality characteristic and increase the sensitivity of the anomaly detection when presented in the process. Consequently, providing accurate model reduces the false alarm rate (FAR) and/or missed detection rates (MDR).

4. **Anomaly Diagnosis:** The control charts are not designed to identify the root cause of the detected anomaly. Thus, we proposed the same *LADR* to determine the reason for the detected anomaly in the process by using the interpretable patterns that construct the *LADR* model. Accordingly, the variables that contribute to the anomaly are identified without resorting to collecting or generating sufficient data for different anomalies conditions.

1.5 Originality of Research

The originality and novelty of this research is as follows:

1. A data modeling technique is used to develop a conceptual and logical models in order to characterize and manage the diverse and heterogeneous data exchange through a manufacturing system. They are considered essential steps in the Quality 4.0 paradigm. They achieve high levels automation including high quality of stored data and avoids redundant, incomplete, inconsistent Data compared to the current data models that were used in the manufacturing. Moreover, it remedies the absence of standardization in terminology of communication between stakeholders and provides the required key performance of indices for all stakeholder. Furthermore, It has been implemented for the first time to be used for inspection and repair processes in Aerospace domain.

2. A new machine learning technique *LADR* is developed to obtain a regression model from the historical data in manufacturing database. The *LADR* is based on the standard *LAD* methodology where the independent variables represent the extracted patterns from the historical data. The *LADR* addresses the research gaps and limitations of the recent researches of *LAD*-based regression approaches. Moreover, it provides significant results compared to the other well-known technique.

3. New three discretization methods are introduced to *LADR* to improve not only the concept of the *LAD* based regression approaches but also, improve its performance. These methods are responsible for converting the problem from classification to regression to obtain strong patterns that describe the data. The results demonstrate improvement in the performance of the *LADR* compared to recent researches of *LAD*-based regression approaches as well as the other well-known machine learning techniques.

4. A clear methodology is presented for implementation of the *LADR* technique to build a regression model based on strong patterns that are extracted from the original data using *cbmLAD*. Unlike the recent researches of *LAD*-based regression approaches, *LADR* has no limitation on the degree of the generated patterns, which affect the accuracy of the regression model.

5. To the best of our knowledge, *LAD* has never been integrated with the control chart. *LADR* has been adopted as a regression adjustment with the control chart has been introduced to detect any anomaly present in the process. This integration shows a reduction in the false alarms and missed detection rates compared to other approaches.

6. A new methodology to perform a root cause analysis to identify the reason for the anomaly experienced in the process using the same regression model that was obtained with the *LADR* technique. Unlike other machine learning- based control chart, the *LADR* does not need to collect or generate sufficient data for different anomalous conditions or to acquire any additional classifier.

7. A new condition monitoring and warning mechanism is introduced based on Logical Analysis of Data Regression (*LADR*) and Residual control chart (*RCC*). This mechanism exploits the strength of the *LADR* to detect any faults in an industrial system and identify the root cause of the detected fault to take the appropriate corrective action.

The research approach leans towards both developing the cutting edge of research and industrial applications. The research objectives are not only working on improving the performance of the used methodologies but also performing concept improvement. In this work, we present the *LADR* technique, a novel regression model by introducing three discretization methods, which haven't been used in such types of regression problems to the best of our knowledge. Thus, our proposed regression model serves as a conceptual extension of the standard *LAD* methodology to suit regression problems, thus expanding the field of using *LAD* methodology in regression problems. We show in this thesis that our proposed method compares favorably to the current state of the art of *LAD*-based regression approaches, as

well as other well-known machine learning techniques. Moreover, we provide a clear and detailed description of our method while addressing the gaps in the literature. That said, we keep versatility and applicability in the industry in mind while developing our proposed approach. We develop a novel *LADR*-based control chart methodology that facilitates the root cause of detected anomalies presented in industrial processes. We show that our approach outperforms state-of-the-art methods in this regard. This boosts the ability to take adequate corrective actions to eliminate such anomalies in industrial processes and sustain the process in-control operation. Our approach can be used in many industrial applications, such as production lines, machining processes, and quality improvement.

CHAPTER 2 THESIS ORGANIZATION

The thesis is presented in eight chapters. The current chapter is Chapter 1, which provides a brief introduction about the process quality monitoring techniques and the new trend of digitalization of quality management. It also addresses the problem statement and challenges, in addition to demonstrating the general objectives and originality of this research. Chapter 2 is divided into two parts. The first part reviews the data management architecture to control and characterize the data, its traceability, and its adaptability for use by the stakeholders. It provides the concept for designing a data model that organizes and manages the relevant process data in a well-structured database to ensure high data quality. On the other hand, the second part provides a background about the control charts and their limitations in process monitoring. Then, it introduces the importance of integrating machine learning techniques with conventional control charts to increase the sensitivity of anomaly detection and diagnosis.

Chapter 3 presents the concept of data modeling to store, organize, and manage the data generated from the sensors and controllers of a process. It describes the design of conceptual and logical models using a technique called "Entity Relationship, *ER*". The *ER* has been applied in a real case study on inspection and repair processes in the Aerospace domain. It is considered an essential step in data preparation for Quality 4.0.

Chapter 4 introduces a new regression technique, *LADR*, based on the standard *LAD* methodology. It explains the methodology for implementating the *LADR*, showing how it strengthens the developed model. The performance of the new technique has been evaluated using different datasets. Moreover, it has been compared with the performance of well-known regression techniques.

Chapter 5 provides a *LADR*-based control chart as a new model-based control chart to improve the performance of anomaly detection during process quality monitoring. Unlike other integrations, a *LADR*-based control chart is not only used for anomaly detection but also for performing root cause analysis to identify the reason for that anomaly. Therefore, the methodology of the proposed integration is described in the terms of fault detection and identifying the root cause of the anomalous process. To evaluate the performance, the results of the proposed technique are compared with those of the other techniques.

Chapters 6 and 7 focus on monitoring the operation of the belt drive system under different conditions. In Chapter 6, extensive experiments are carried out to collect vibration signals during the operation of the system. It describes in detail the description and importance of

collected signals and how the experiments are performed. This is considered an introduction into proposing a new condition monitoring and warning mechanism based on the Logical Analysis of Data Regression ($LADR$) and Residual control chart (RCC) in Chapter 7. The implementation of the proposed mechanism is elaborated upon for monitoring and fault detection and diagnosis during the operation of the belt drive system.

Chapter 8 presents a summary of the contributions of this thesis, areas for further research, and some concluding remarks.

2.1 Deliverables

The following is a list of the outcomes of this thesis:

1. Khalifa, R.M., Yacout, S. & Bassetto, S. (2021). Developing machine-learning regression model with Logical Analysis of Data (LAD). Computers and Industrial Engineering, 151, 16 pages. Retrieved from https://doi.org/10.1016/j.cie.2020.106947

2. Khalifa, R.M., Yacout, S. & Bassetto, S. (2021). Quality 4.0 : entity relationship model for inspection and repair processes in aerospace domain. Paper presented at the 6th North American Conference on Industrial Engineering & Operations Management (IEOM 2021), Monterrey, Mexico (11 pages).

3. "Root Cause Analysis of an Out-of-Control Process Using a Logical Analysis of Data Regression Model and Exponential Weighted Moving Average"

   - Authors: Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto
   - Submitted to Journal of Intelligent Manufacturing (JIMS) on 4th of March 2022. The paper is under review.

4. "Experimental vibration data collected for a belt drive system under different operating conditions"

   - Authors: Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto, Yasser Shaban
   - Submitted to Data in Brief, ElSEVIER on 16th of May 2022.

5. "Condition monitoring and warning mechanism in the belt drive system based on Logical Analysis of Data regression based residual control chart"

   - Authors: Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto, Yasser Shaban

- Submitted to Mechanical Systems and Signal Processing, ElSEVIER on 16th of May 2022.

CHAPTER 3 LITERATURE REVIEW

Quality 4.0 has been introduced as a new concept in the era of Industry 4.0 [25]. The American Society of Quality (ASQ) states that Quality 4.0 is the organizational excellence and the future of quality management [26]. Quality 4.0 represents the impact of the digital transformation of quality management in terms of quality tools and technology, people, and processes [10]. Quality 4.0 closely aligns quality management with the era of Industry 4.0 to enhance the efficiency, performance, innovation, and business models of an organization [27]. The Quality 4.0 framework has 11 axes as depicted in Figure 3.1. This thesis addresses the data management and process quality monitoring which are related to the two axes "Data" and "Analytics", respectively, in the framework.

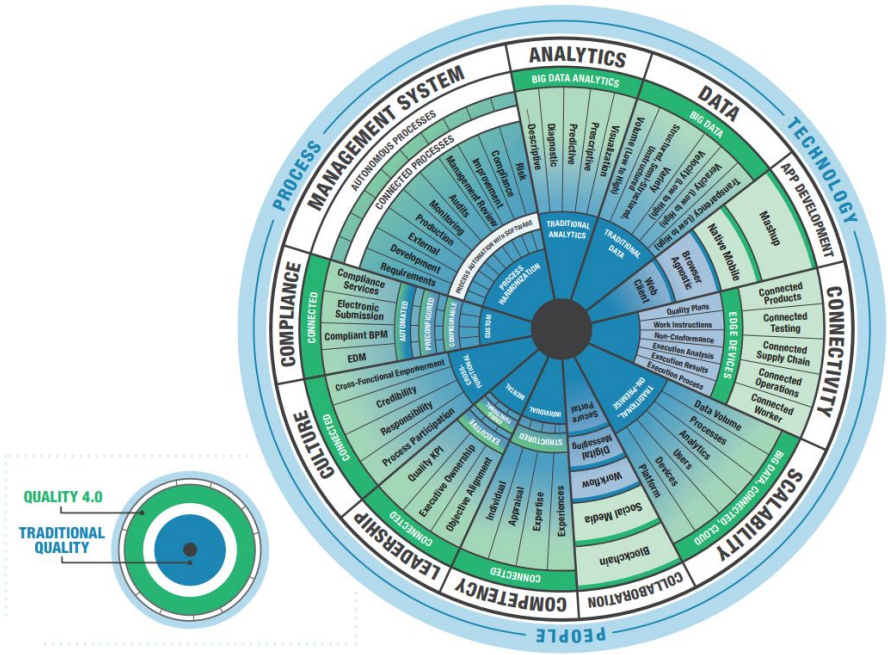


Figure 3.1 Quality 4.0 framework [10]

3.1 Data management

Data-driven decisions have been considered the heart of quality management for decades [10]. With the rapid progress of, and complexity in manufacturing, there is a large amount of diverse and heterogeneous data exchange through sensors and controllers within a manufacturing system. Since data is a valuable aspect of process quality monitoring, it must be handled

and managed using a data management system. The data management system is used to ingest, collect, store, organize, cleanse, govern, integrate, protect, and maintain the required data for purposes of further analysis. In other words, it is an architecture that is used to achieve the requirements of data for the life cycle in a manufacturing system [28, 29]. Consequently, it characterizes, identifies, and controls the data or information generated by the system. It establishes and standardizes communications between operations' stakeholders, and in this case, quality leaders and engineers.

3.1.1 Data Types

Data is usually characterized in terms of 5 V's. These include 1) Volume, which refers to the size and amount of the collected data, 2) Value, which indicates the importance and use of the data, 3) Velocity, which is the rate of the data stream, 4) Veracity, which means the accuracy of data , 5) Variety, which refers to the diversity of different types of data that can be summarized into three types: structured, semi-structured, and unstructured data [10, 30].

Structured data conforms a data model in which the collected data has been well organized and stored in a pre-defined structure with a certain hierarchical format [31]. Generally, it follows a tabular format with various rows and columns where their relationship is well defined. It usually resides in a relational database management system (RDBMS) which manages the relationships in the database. This type is eminently searchable for algorithms or even for human-generated queries. It is easily used and understandable by users in an organization. Conversely, unstructured data does not conform to a data model and the collected data is heterogeneous, irregular, and has no pre-defined structure such as videos, text, XML, ..etc. [32]. It constitutes 80-95% of the available data [30]. Unstructured data can be human-generated (such as emails, text files, and digital photo) or machine-generated (such as sensor data, weather data, and seismic imagery). It usually resides in a non-relational database management system. The cons of the unstructured data include requiring a data science expert to be able to understand and analyze the generated data so that it can be used later, and specialized tools that are necessary to manipulate this type of data. On the other hand, semi-structured data lies between the structured and unstructured data types [31]. Although it is semi-organized and does not conform to a data model, it is easier to be analyzed compared to unstructured data type.

3.1.2 Data Modeling

With the growth of data and information that is gathered from different sources and in different formats, there is a need to formulate and store data in a structured format [33]. The

data modeling process is used to design a robust data model that characterizes the required data of the database in an entire organization and describes the relationships between these data. It organizes and defines the structure of the data stored in the database. The data model ensures that the stored data meets business requirements by being a high quality of data. High data quality means 1) accuracy, 2) validity, in that the stored data follows pre-defined standards in the database in terms of type, size, and format, 3) integrity, in that the validity of the relationships across the data stored can be traced, 4) completeness, 5) consistency, in that the data is in sync across the database, and 6) timeliness, meaning that the data is available when it is needed [34].

The data model has four stages in designing a database, as depicted in Figure 3.2 [35, 36]. The Business requirement is the first and the most important stage. It aims to understand the manufacturing process and the interactions between the stakeholders and the life cycle of the process. In this stage, the required data is identified and characterized based on the process and the stakeholders' requirements, whether for further analysis or to prepare reports [37]. Then, the collected data and information is used to design a conceptual data model. The conceptual data model represents a high level of abstraction which defines what the manufacturing process contains in terms of concepts and rules [38]. This model is refined and converted into a logical data model, which provides more detail about the structure of the data. It maps all of the elements (attributes) of the data in the structure and defines the relationships between them. Finally, a physical data model is developed, which is the actual implementation of the database using Database Management System (DBMS) software [37].

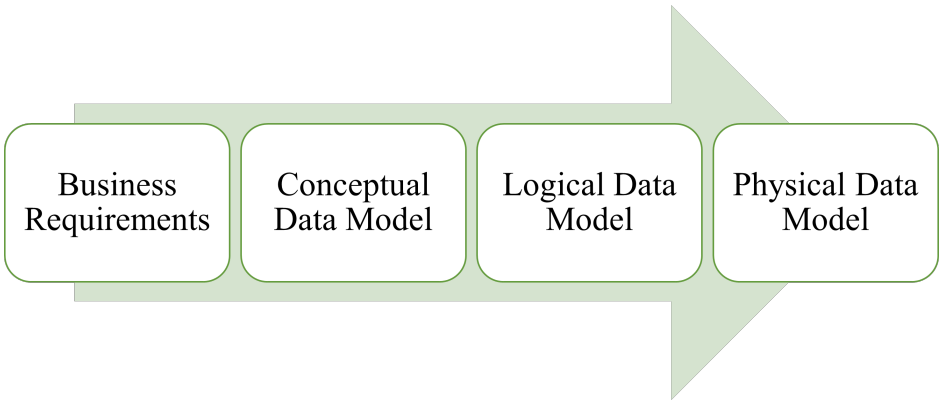


Figure 3.2 The stages of Data modeling

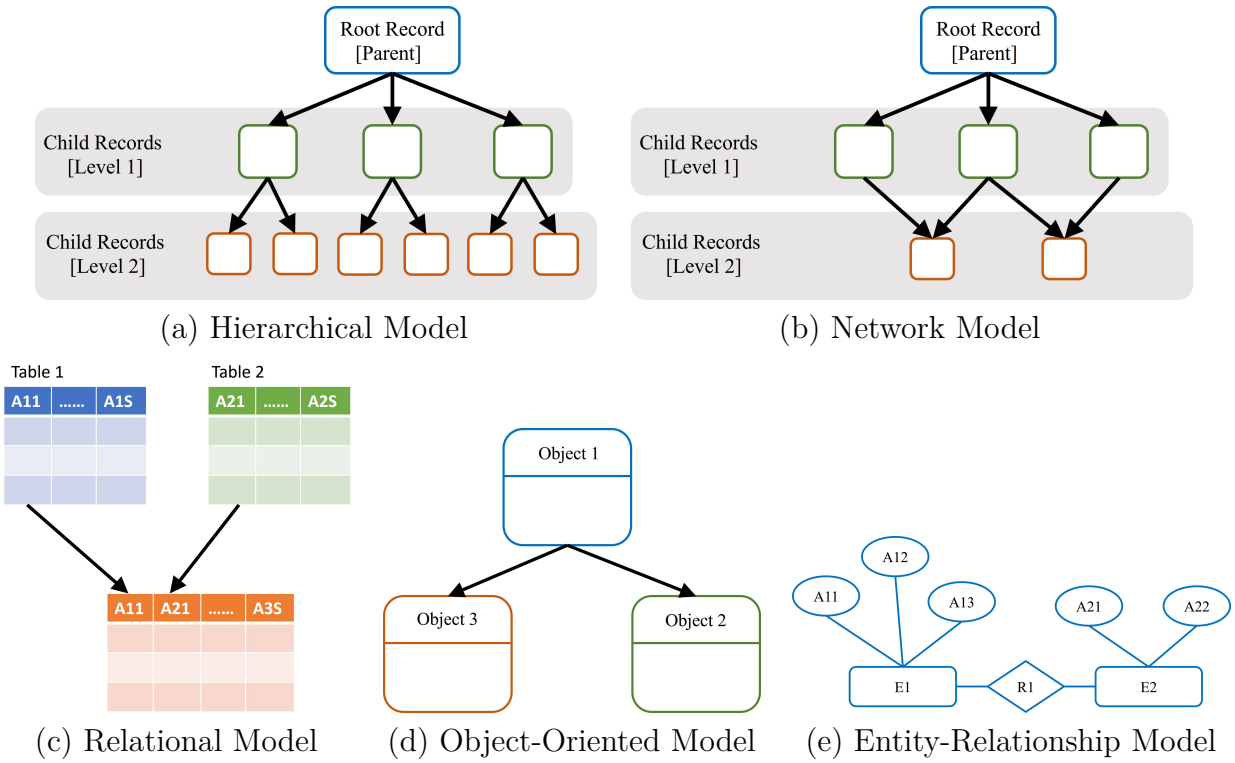
Several types of data modeling approaches have evolved along with the growth of data storage required by business organizations. There are five main types: 1) Hierarchical Data Modeling (*HDM*) [39], 2) Network data modeling (*NDM*) [40], 3) Relational Data Modeling (*RDM*)

[41], 4) Object-Oriented Data Modeling (*OODM*) [42], 5) Entity-Relationship Data Modeling (*ERM*) [43], as depicted in figure 3.3. Hierarchical Modeling is considered the oldest data model. It organizes the data in the form of a tree-like structure, which is represented by the parent-child relationship. The root record is the peak of the structure, which has a set of parents' records. Each parent's record has one or more children. This model manages a large amount of data and improves data sharing. However, it is rarely used due to its complexity of implementation. Network data modeling looks like *HDM*. Nevertheless, it is easier than *HDM* in representing complex relationships, which include multiple parent records. *NDM* organizes the data in a graph where the child can have one or more parent records. However, the structure of the database becomes more complex than *HDM*. Moreover, it is difficult to apply structural changes to the existed database. Relational data modeling represents the collection of data and information, which is organized as related-based multiple tables. *RDM* facilitates the design of the database and promotes the independence of the structure. It supports multi-level relationships between the tables with large amounts of data. Object-Oriented Data Modeling represents the data and information in the form of objects. The objects with similar functionalities are gathered and linked to different other objects. Although it can handle different types of data, it is still limited and mostly a theoretical approach. In addition, the obtained model can be complicated to design or understand. Entity-Relationship modeling is a graphical approach that represents the data and information in the form of entities. Each entity has one or more attributes that describe that entity. The *ERM* provides a better description of the stored data in the database and reduces the data redundancy and inconsistency to ensure high data quality in the database.

Various data modeling languages are widely used in computer science and software engineering to implement different data modeling approaches. They aim to present the information and data structures of reality as designed by the approach [44]. The following are the modeling languages:

1. Unified Modeling Language (*UML*)
   UML is a general-purpose modeling language, which is mainly used to design software. It is a graphical modeling language that constructs, standardizes, and documents the visualization of the designed software system. There are several *UML* methods which can be divided into two types: structural and behavioral *UML*. It facilitates the design of complex software and provides clear and real communication of the designed software. It consumes much time to design and maintain the *UML* code [45].

2. Extensible Markup Language (*XML*)
   XML is a modeling language that is used to store and share the data in self-descriptive



(a) Hierarchical Model (b) Network Model

(c) Relational Model (d) Object-Oriented Model (e) Entity-Relationship Model

Figure 3.3 Data modeling approaches

structures. The relationship among the data is following the concept of the hierarchical modeling approach. The *XML* describes these structures using tags (<text>) through text formatting. These tags are determined by Document Type Definition (*DTD*) that defines the data structure, including all attributes. The *DTD* is used to verify whether the *XML* is valid or not [46]. Generally, the *XML* is characterized as it simplifies the data sharing, transporting, and availability. Nevertheless, it uses excessive syntax in terms of a large number of entities and tags. Thus, JavaScript Object Notation (*JSON* is considered one of its alternatives.

3. Systems Modeling Language (*SysML*)
   SysML is a modeling language that is developed from the Object Management Group (OMG) [47]. It is considered an extension of *UML* to support the analysis, specification, design, and validation of systems engineering applications. The *SysML* is more flexible and expressive compared to *UML*. It allows modeling a wide variety of complex systems from different views in terms of behavior, structure, or requirement [44]. This language is easy to learn and implement because its notions are not ambiguous.

4. Entity-Relationship Diagram (*ERD*)

ERD is an essential tool that is used for the data management system. It is based on Entity-Relationship Modeling. It describes the data structure as entities and defines the relationship between each entity. Each entity contains several attributes that define that entity [48].

3.1.3 Entity Relationship Modeling (*ERM*)

ERM is one of the most common data modeling techniques. It was first developed by Chen [49]. It represents the structure of the database of a specific domain in a set of entities and defines their relationships [50, 51]. The database structure can have several entities. Each entity can correspond to persons, objects, spaces, or concepts in the real world, and it can include several attributes. An attribute is used to define the characteristics of an entity. The attribute can be a simple attribute in which is an atomic attribute, and cannot be further subdivided, and a composite attribute, which can be represented with a set of simple attributes. Moreover, it is classified according to its value in a multi-valued attribute, as it can have two or more values, and as a derived attribute, which derives its value from other attribute(s) in the database [52]. Each entity can have two types of keys, a primary (unique) key, and a foreign key. A primary key attribute is a unique identifier that identifies the entity and cannot be repeated. On the other hand, the foreign key in an entity is considered a primary key of another entity and it sets up a relationship with that entity. The relationship between two entities in *ERM* is defined using connecting lines and cardinality (Kashmira 2018). Cardinality is determined by the maximum number of times that an instance within an entity can relate to many instances within another entity. Consequently, cardinality can describe the relationship in one-to-one (1:1), one to many (1:M), and many to many (M: M). The notations of the *ERM* are summarized in Figure 3.4.

The *ERM* provides a visual representation of the designed database in terms of defining entities, attributes, and relationships. Therefore, it allows the business users to easily observe the data flow, along with the database, and understand the data stored in [36]. It improves the data quality and ensures that there is no missing or redundant data. The conceptual model is simple and easily converted into the logical model and physical model that needs to be implemented using DBMS software.

Once the data model is designed and implemented in the manufacturing process, it will easily collect the process data and define key performance indices for quality, traceability, and reproducibility control. The stored data in the database will be considered historical data. Therefore, we use the historical data to determine the parameters of the quality tool. Subsequently, the quality tool will be able to monitor the process online to detect any anomaly

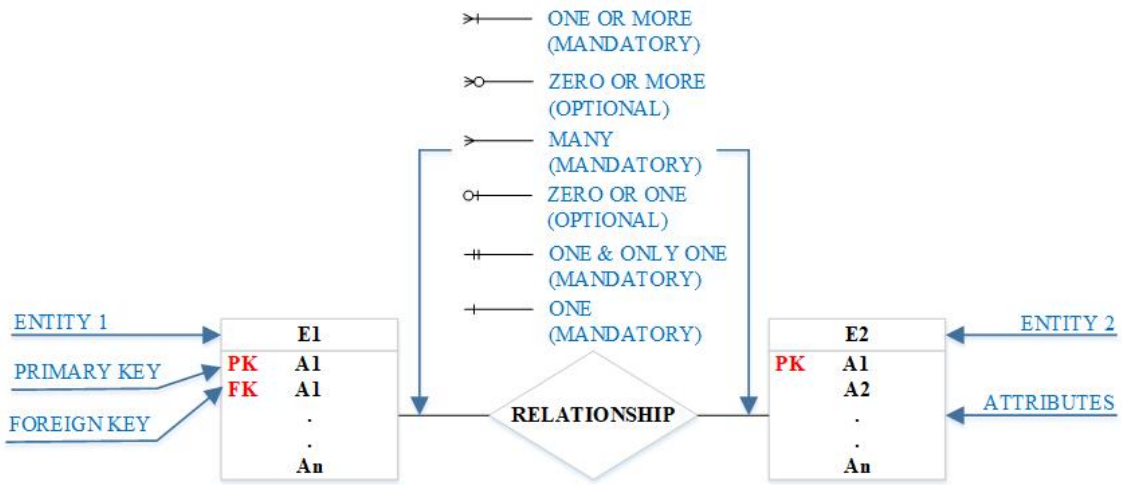


Figure 3.4 The notations of the *ERM*

experienced in the process, and hence it identifies its root cause to take appropriate corrective actions, as shown in Figure 3.5.

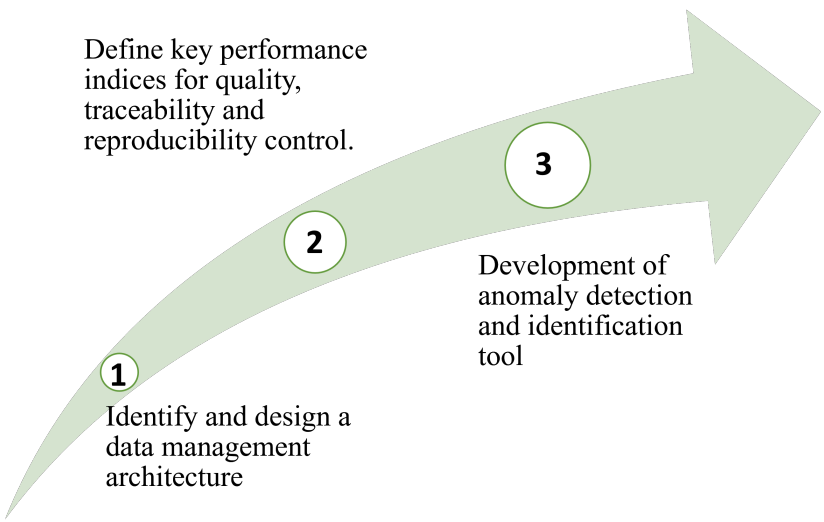


Figure 3.5 The bridge from data modeling to quality monitoring

3.2 Control charts

Control charts are statistical tools for monitoring process variations over time. They graphically represent variations in the quality characteristics, so these charts can determine the condition of the process and whether it is in-control or out-of-control. In 1924, Dr. Wal-

ter Shewhart developed the first control chart and used it to monitor and detect potential variations in the process [7, 53].

Control charts were developed for industrial applications, and they were also used in different fields. They were used in the prediction of failures in business works [54], quality management of education [55], medical applications [56], monitoring ecological systems [57], and improvement of sports applications [58].

The control chart is introduced in Figure 3.6. It is used to monitor quality characteristics following a normal distribution, $N(\mu_0,\sigma^2)$ where μ_0 is the target mean and σ is the standard deviation. It contains three horizontal lines to describe the quality characteristic of the product versus the order/time of the sample. A centerline is the target value, which represents the mean value of the monitored quality characteristic. The other two horizontal lines represent the control limits of the control chart, the upper control limit (UCL) and lower control limits (LCL). When the observation at time, t, falls within the control limits, the process is considered in-control. An out-of-control process occurs when the observation falls outside these limits or shows abnormal variations. Although the process is within the control limits, it may be out-of-control. Process observations can form a systematic or non-random pattern, which indicates abnormal behaviors.

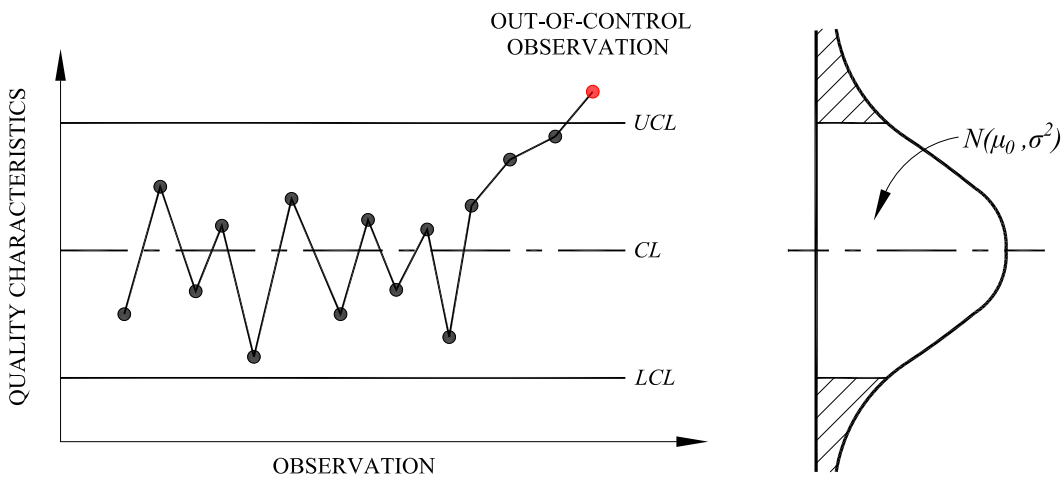


Figure 3.6 The characteristics of the Control chart

Therefore, a statistical hypothesis demonstrates the condition of the sample mean (μ_0) with respect to the target value of the process (μ_0). The null hypothesis (H_0) is when the process state is in-control in contrast to the alternative hypothesis (H_a) that states an out-of-control sample as given in equation (1).

$$\begin{aligned} H_0 &: \mu_1 = \mu_0 \\ H_a &: \mu_1 \neq \mu_0 \end{aligned} \tag{3.1}$$

The hypothesis test evaluates the performance of the control chart based on two types of errors, type I error (α) and type II error (β). Type I error is when the process is at an in-control state, but the control chart rejects H_0 and declares an out-of-control process. It is an important statistical parameter to indicate the control chart limits, which are represented by the acceptance region. This region is defined by the standardized z_0-statistic such -$z_{\alpha/2}$ $\leq Z_0 \leq z_{\alpha/2}$. The designers usually determine the suitable α–value (common values 0.01 to 0.1) for control charts to avoid detecting the normal data as anomalous ones. In a type II error, the process is originally out-of-control and the control chart accepts H_0 showing an in-control process. This β–value is calculated using the following equation:

$$\beta = \phi(z_{\alpha/2} - \frac{\delta}{\sigma}\sqrt{m} - \phi(-z_{\alpha/2} - \frac{\delta}{\sigma}\sqrt{m}) \tag{3.2}$$

Where $z_{\alpha/2}$ is z-statistic at $\alpha/2$ (two tailed test), δ is the difference between μ_0 and μ_1, m is the sample size and $\phi(.)$ denotes the standard normal cumulative distribution function.

The parameters must be adequately adjusted to reduce these errors and be effective in monitoring the process. Figure 3.7 illustrates two types of given data, normal and anomalous data, and represents the type I error and type II error in grey and green colors, respectively. Average run length (ARL) is a measure of performance of the control chart describing the two previous error types. The ARL is the average number of samples that must be plotted in the control chart after an assignable cause has happened and before a sample falls outside the control limits, thus declaring the process to be out-of-control [1]. When the process is in-control, large in-control ARL (ARL_0) contributes to a reduction in false alarms. Conversely, small out-of-control ARL (ARL_1) is needed for out-of-control processes in order to rapidly detect the change [59]. Hence, the ARL_1 is much smaller than ARL_0.

The control charts are implemented in two phases, phase I and phase II [60]. Phase I is a retrospective analysis that uses the available process data to determine the control limits. Thus, these data are data on the normal conditions and this is for two reasons. The first reason is to construct reliable limits for the control charts and accordingly, to easily monitor the process variations in the future. The second reason is to reduce the source of the error of type I and II. On the other hand, phase II is applied for online monitoring of the process variability.

Shewhart control charts [61] are used to detect a large mean shift in the process. They

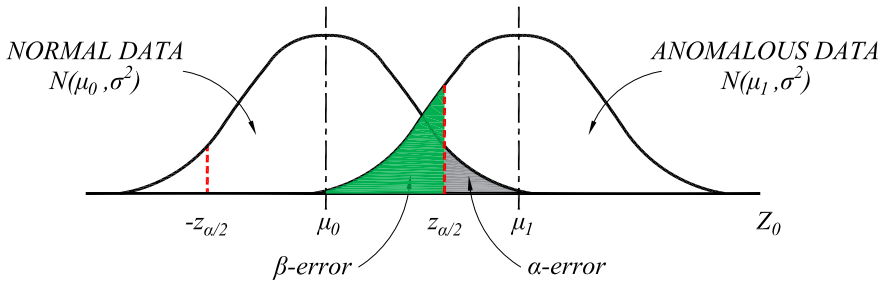


Figure 3.7 Type *I* & *II* errors

represent a memoryless control chart, as they only use the last (current) sample information and ignore the rest. Thus, these charts are less sensitive to small or moderate mean shifts. Consequently, alternative control charts were developed to detect the small deviations in the process with reliable parameters, mean and variance. They took all of the previous information into consideration when monitoring the process variations.

The control charts are classified into univariate and multivariate types. The univariate control chart is used to monitor a single quality characteristic in the process. The most popular control chart for detecting small mean shifts are exponential weighted moving average (*EWMA*) [62] and Cumulative Sum (*CUSUM*) [63]. Most industrial applications have more than one variable that needs to be monitored in a process. One of the univariate control charts can be applied for each variable separately, which is considered an incorrect choice for many reasons. The univariate control chart may fail to detect the small deviations in the process when several variables are correlated. Accordingly, the false alarms increase during an operational process. Moreover, if there is a large number of variables, this will reflect directly on the charts, in addition to certain challenges in monitoring the process. In order to overcome these drawbacks, they have obtained Multivariate control charts [64–66].

3.2.1 Limitations of control charts

Although there have been improvements in control charts to monitor the quality of the process, some limitations have been explored that affect their performance.

(**A**) **Autocorrelation**

The control charts were constructed based on the assumption of normal and independent data distribution, $N(\mu_0,\sigma^2)$. However, many applications provided streams of data that had high dependency; this phenomenon is called autocorrelation [67]. A sample autocorrelation function is estimated to determine the correlation between the observations that are L-time periods apart as the following:

$$r_l = \frac{\sum_{t=1}^{m-l}(x_t - \bar{x})(x_{t-l} - \bar{x})}{\sum_{t=1}^{m}(x_t - \bar{x})^2}, l = 0, 1, ..., L \tag{3.3}$$

Where r_l is the value of autocorrelation function at l, x_t is the observation at time t, x_{t-l} is the observation that lags with l-time period where $l \leq m/4$ [1]. Thus r_l is plotted at different values of l with limits that were identified by the two standard deviations of r_l. So, if the value exceeds the limits, the plot detects non-zero autocorrelations, which are enough to disturb the performance of the control chart.

Accordingly, conventional control charts could not be used to avoid misleading consequences [68]. This has led to signaling a large number of false alarms and time delay to detect the mean shift if the data of the process has a positive autocorrelation [15].

(**B**) **The Curse of Dimensionality**

Modern technological progress in various industrial applications has led to a continuous increase in the number of interest variables (n). This makes monitoring the performance of process more cumbersome [16]. Increasing n increases the ARL_1 performance at the same parameters [69]. In other words, the conventional control charts cannot handle a large number of variables because this increases the performance of ARL_1. There is a time delay for detecting the mean shift in the process [70].

(**C**) **False alarm rate**

High false alarm signals are crucial problems that are inherent within the control charts, resulting from the determination of the control limits [71]. A false alarm is the detection of incorrect out-of-control points in the process. Many factors affect the control chart limits that cause these alarms. However, narrower control limits improve the detection of small mean shifts, but they increase the false alarm rate as a result of ARL [1, 72]. Moreover, when the non-normality assumption is not valid, this leads to a distortion in process monitoring, causing false alarms. Furthermore, estimated parameters may maintain inaccurate control limits, which increase the false alarm rate in addition to misleading results [60, 73, 74]. Refer to autocorrelation data, which is one of the main reasons for false alarms.

(**D**) **Anomaly identification**

Most industrial processes experience different anomalies during the operation of a system. Identification of an anomaly plays an important role in taking corrective action to return the system to a normal state. Although the control charts can detect an

anomaly presented in the process, they are not designed to identify the root cause of that anomaly [70]. Additional techniques are required to determine the anomaly online when the process is out-of-control.

3.3 Machine Learning Based Control Charts

The main goals of the Quality 4.0 paradigm in manufacturing are: (1) to monitor the process and/or product accurately and overcome the drawbacks of the conventional control charts to ensure anomaly free process, (2) to perform real-time root cause analysis if an anomaly is detected, (3) to speed up decision-making to take corrective action [26, 75]. Companies are striving towards the implementation of quality and integrating machine learning techniques with conventional control charts.

Machine learning is a form of artificial intelligence that applies a variety of algorithms to describe the system data [76]. Machine learning has a set of techniques that have the ability to detect patterns in a given data and exploit these patterns for predictions and/or performing decision making [77]. Machine learning techniques were implemented to extract unknown knowledge and describe the relationships within the dataset. They have the ability to have high-dimensional data and obtain relationships for complex and dynamic datasets, even chaotic behaviors [76, 78]. They have been applied in different fields, such as fault diagnosis in industrial processes, climatic science, biology and genetics, business and finance, etc.

Machine learning techniques play an important role in increasing the sensitivity of the control charts. They are used for anomaly detection when any abnormal behavior is experienced in a process. Several Machine learning methods were combined with control charts, not only to describe the process variability but also to recognize the occurrence of any anomalies. They are implemented to describe the relationship between the independent variables and the dependent variable [79]. A univariate control chart is then used to monitor the variations of the dependent variable instead of using a multivariate control chart. The control chart monitors variations in the residuals obtained from the regression model. The residuals are the difference between the true values and the predicted values of the model. The residual has no evidence of the autocorrelation phenomenon. If the residual value at any time falls outside the control chart limits, an anomaly is detected. Moreover, machine learning techniques are implemented for anomaly diagnosis as well as anomaly detection in the process.

Some techniques are integrated with control charts to handle the curse of dimensionality with highly correlated process variables such as Partial Least Square *PLS*-based *EWMA* [80],

Principal Component Analysis *PCA* with an *EWMA* [81–86]. They obtain a number of uncorrelated components or variables from the original correlated variables. The number of these components or variables is less than the original. A Poisson Principal Component Regression (*PPCR*)-based control chart was proposed for monitoring count data and Poisson processes. *PPCR* combines the Poisson regression and *PCA* [87]. This strategy considered the impact of collinear independent variables on the dependent variable in contrast to integrating Poisson regression with a control chart [88]. Thus, it preserves the relevant information in the multicollinearity data and reduces the false alarms rate. Many research studies extended the use of *PCA* to develop two new schemes to monitor mixed (continuous and categorical) quality characteristics simultaneously [89–93]. A support vector regression (*SVR*) was combined with *EWMA* in which *SVR* handles non-linear relationships compared to multiple linear regression (*MLR*) [72]. Hybrid approaches were proposed that combined two or more machine learning techniques such as *PCA*-support vector machine (*SVM*) [94]. Furthermore, other approaches have been developed for time series forecasting such as Autoregressive moving average (*ARMA*) [95, 96] and Autoregressive integrated moving average (*AMIRA*) [97]. The quality characteristics of some industrial applications, such as tool wear, are drifting linearly over time which the conventional approaches can not handle. Therefore, a regression spline control chart was introduced to monitor the quality characteristics that exhibit with nonlinear profile over time [98].

After detecting an anomaly, it is essential to identify the root cause of that anomaly. Several machine learning techniques were used to diagnose an anomaly. The least absolute shrinkage and variable selection (*LASSO*) is used as a variable selection method with a penalized term [17]. It shrinks the coefficients of its regression model until they tend to a value of zero. Each time, by solving *LASSO*'s objective function, it provides transition points. These transition points are active variables that do not change with the value of the penalty term. The order of the variables entered the active set was equivalent to the relevant variables that contributed to the anomalous process fault. However, this has a major disadvantage when the process data has highly correlated variables. Therefore, a hybrid technique was proposed that combines LASSO and ridge regression, called Elastic Net (*EN*) [18]. Furthermore, several pattern recognition techniques are implemented to diagnose the anomalous process such as Decision Tree (*DT*) [99, 100], Support Vector Machine (*SVM*) [8, 101, 102], Random Forest (*RF*) [103], K-Nearest Neighborhood (*KNN*) [104, 105], and Artificial Neural Network (*ANN*) [20, 106, 107]. Hybrid techniques were integrated with control charts, such as *SVM-ANN* [108–110], *PCA-SVM* [94], and *PCA-LASSO* [17]. Practical speaking, it is essential to have sufficient data or generate random data that describes various anomalies for training these techniques.

CHAPTER 4 ARTICLE 1: QUALITY 4.0: ENTITY RELATIONSHIP MODEL FOR INSPECTION AND REPAIR PROCESSES IN AEROSPACE DOMAIN

Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto

Published in:

Proceedings of the International Conference on Industrial Engineering and Operations Management, Monterrey, Mexico, November 3-5, 2021

4.1 Abstract

For decades, data-driven decisions have become the core of quality improvements. A new paradigm has been introduced for the digitalization of quality management with "Quality 4.0". Automatic data and information exchange are the essential steps of the Quality 4.0 paradigm to achieve automation of the manufacturing systems. The goal of this paper is to design and manage the database of the automated system as the first stage of data preparation for quality management. For this stage, we use the Entity Relationship (ER) modeling technique to develop conceptual and logical models. This technique is used in a real case study to organize and manage the database of inspection and repair processes in the aerospace manufacturing. The real merit of the developed models is to create a well-structured database that describes the system's data flow with high data quality.

Keyword: Quality 4.0, Database management, Data modeling, Entity Relationship (ER) Modeling, automated inspection

4.2 Introduction

Quality 4.0 paradigm is a new concept introduced under the general approach of Industry 4.0. It represents the digitalization of quality management where it monitors and controls the quality of the process and/or product [10, 111]. Data and information exchange are the cornerstone of the Quality 4.0 paradigm to achieve high reliability in the automation of the manufacturing systems [10]. In such systems, a large amount of information flow between sensors and controllers independently of continuous interaction with humans, which poses some challenges for existing traditional systems regarding their ability to handle and manage the sheer amount of data and information due to the lack of the necessary tools and infrastructures [112]. A clear example of such a challenge can be found in the aerospace domain in which visual inspection is the main method for inspecting parts [113, 114]. Since this process is carried out by talented and experienced human inspectors, it is still prone to error which can range from 20% to 30% [115]. Moreover, there is a great possibility of missing data and the presence of data redundancies and inconsistency. The stakeholders take much time to prepare their necessary information to do the report and/or analysis. Further, the data are entered manually, so there can be an absence of standardization of communication between stakeholders. As such, numerous researches are trending to automate the inspection and repair using the available data in the process [116, 117]. This current trend is to replace the human element with more accurate automated inspection and repair machines. In turn, suitable tools for handling and managing the generated data are required.

Information modeling should be applied on the events engine parts; inspection, maintenance, repair, and overhaul; to capture the relevant data attributes and describe the current and historical event of the part. This model provides the accurate information required by the stakeholders such as the operators, engineering, and scheduling team to prepare their reports, or analysis or even to make decisions. Consequently, we need to define the necessary data to be managed in the inspection and repair processes in terms of three axes as shown in figure 4.1. Therefore, the importance of efficient Database management tool grows.

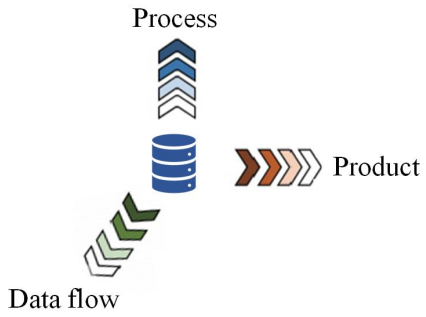


Figure 4.1 Three axes of Database management

Database management plays an important role in organizing and storing data in an appropriate structure to maximize their value. Implementation of database management does not only provide data structure with necessary information required by the process but also, governs the data entry in terms of the six-dimensional data quality; accuracy, validity, integrity, completeness, consistency, and timeliness [34].

Data modeling is commonly used in database design. It creates an abstract model as a visual representation in order to describe the structure of the stored data in addition to additional consistent constraints [33]. The real merit of data modeling is not only to ensure that all required data match the business needs but also, to avoid the presence of any redundant or missing data by respecting its constraints. A data model was proposed as modeling of the maintenance task knowledge of Boeing B737 aircraft [118]. The model is used to capture all information related to the process and history of the aircraft. It supports the stakeholders with the necessary information for maintenance execution to ensure compliance with airworthiness. Okoh proposed a data model that represented the accurate visualization of maintenance tasks applied on the aircraft engines [119]. It supplies the relevant engine information over time which provide new insights and prediction about the lifespan of the engine through-life engineering services. Rodger provided a preliminary data model that was built based on the primary data flow of the aircraft maintenance, repair, and overhaul (MRO) [120]. It focuses on the captured information of the repaired parts of Boeing 777 and

Sikorsky's UH-60 helicopter.

The data model follows four steps to design a database as depicted in figure 4.2. The first stage is Business requirements. It is the most important stage which is dependent on the identification and characterization of the data required in the system. Moreover, understanding the process is the main step to define the necessary information to be collected. The collected information in the first stage is translated into a formal and independent model which is the "Conceptual model". It is a wide coverage picture that identifies the business concepts and their rules. Then, a "Logical model" converts the previous model into a data structure including more details. It defines the data elements in the structure and the relationships between them. The final stage is the "Physical model" that provides the schema which is implemented physically using Database Management System (DBMS) software [37].

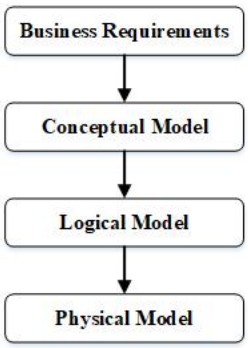


Figure 4.2 Data modeling stages

Entity relationship (ER) model [49] is one of the known data modeling for a database design. It is a graphical tool that is used to provide a conceptual form of the database structure [121]. ER preserves the information of the system and diminishes the data redundancy in storage [44]. In this paper, we use ER modeling technique to design database for inspection and repair in the aerospace domain. We propose a conceptual and logical ER model of a real case as the first step of data preparation for Quality 4.0. The paper is organized in sections as follows. Section II describes the methodology of ER modeling in the design of a database. Section III shows a description of the case study and discusses the first three stages in the data model regarding figure 4.2. Section IV presents the conclusion and future works.

4.3 ER model Technique

ER is a data modeling technique that represents the database of the system in a set of entities that are stored in a database and their relationships [50]. The entity can represent

persons (such as inspectors), spaces (such as shop floor), objects (such as Order and inspection machine), or concepts (such as inspection and repair processes) about storing data in the database. The real application consists of several entities. Each entity contains a set of attributes that define the characteristics of that entity. The attribute can be simple, composite, derived, or multi-valued. The simple attribute is an atomic attribute; e.g. Serial number (S/N) of inspected or repaired part. While the composite one is composed of a set of simple attributes; e.g. Defect dimension attribute is represented by depth (D), width (W), and Length (L). The attribute can have more than one value, which is called a multi-valued attribute. This means that each S/N can have more than one defective location. The values of the derived attribute are derived from another attribute(s) that exists in the database; for example, when determining the number of defects per inspected part. Each entity must have a primary (unique) key(s) that uniquely identifies instances of that entity which can not be repeated; each inspected part has its own S/N. The foreign key in an entity represents a primary key of another entity, which establishes a relationship with that entity. ER modeling defines the relationships between the entities using different types of connecting lines [122]. Each relationship indicates the number of instances of a certain entity relates to one instance of another entity, which is called cardinality. Cardinality can be one to one (1:1), one to many (1: M), and many to many (M: M). The relationship can be mandatory when every instance of one entity must have a relationship with the other entity. The optional relationship does not require the necessary participation of one entity in a relationship with the other entity. Figure 4.3 summarizes the necessary notations for ER diagram.

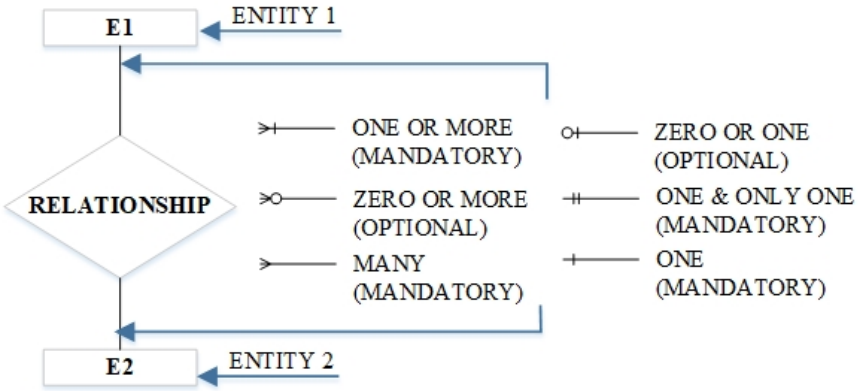


Figure 4.3 The notations for ER model

ER diagram facilitates defining the terms entities, attributes, and relationships. In addition, it provides a preview for connection between each table, which allows constructing databases quickly. Finally, it provides a better understanding of the information stored in the database [36].

4.4 Case study: Application on inspection and repair in aerospace domain

4.4.1 Description of the case study

Rolls Royce is considered one of the leading companies in the aerospace domain. Rolls Royce maintenance center is aspiring in cooperation with AV&R Vision & Robotics, Polytechnique Montreal, Conseil national de recherches Canada (CNRC), and Laval university to develop automated system called SARA which is "Système d'Analyse et de Réparation Automatisée" (Automated inspection, Analysis and Repair System). In the current system, the inspectors perform Visual inspection on the mechanical parts to find the surface defects based on work instructions and the necessary measurement tools for inspection. They assess the defective locations in the inspected parts approximately and compare them with the limits in the engine manual. The new system aims to develop automated analysis, inspection, and repair systems for mechanical parts of engines using the machine vision instead of human vision as depicted in figure 4.4. In the inspection stage, the mechanical part is inspected by the inspection machine to check for any surface defects such as nick, scratches, pitting, etc. If any defect is detected, it is assessed and classified in the defect classification step based on the measuring the defect geometry (depth, width, length, and type) in the engine manual. Then, the decision in the sentencing step declares whether this part can be repaired or if it is scrapped. In case of being repairable, the defect locations are defined, and the repair service is carried out by the repair machine to meet standard specifications.

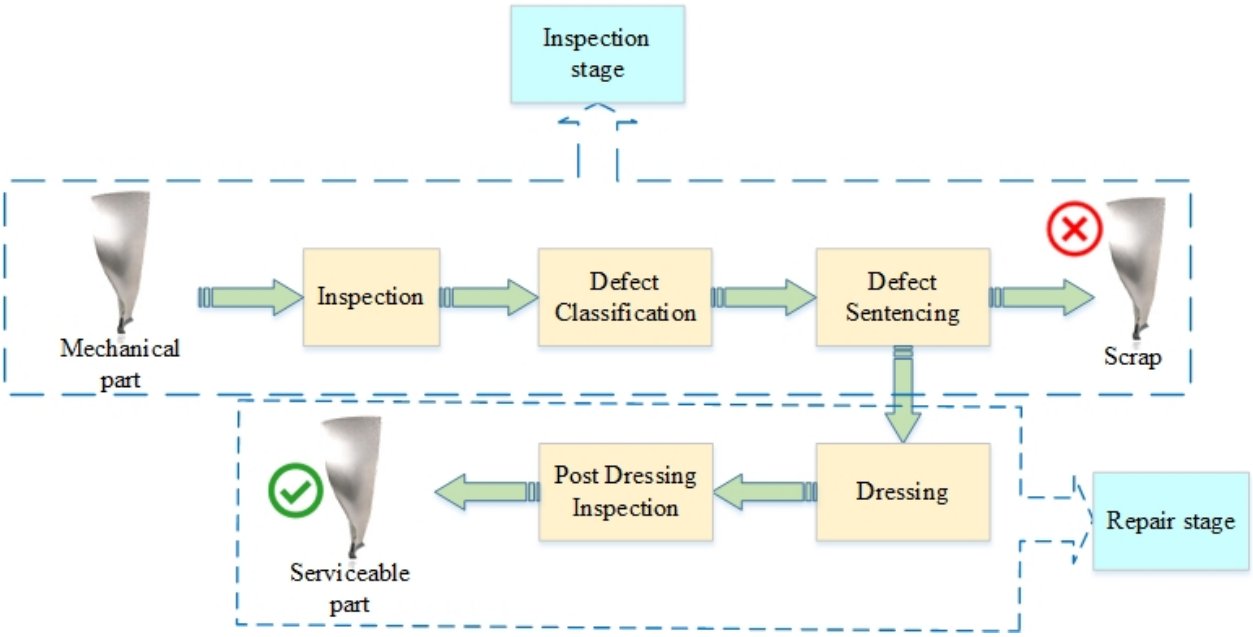


Figure 4.4 Inspection and Repair stages

Database management (DBM) provides capabilities to access, integrate, cleanse, govern, store, and prepare data for further analysis. Furthermore, it characterizes and controls the information generated by the SARA system and proposes a data management structure for operations. It establishes and standardizes communications between operations' stakeholders such as inspectors, engineers, Material Review Board (MRB), etc. In this paper, we propose an ER model for database design and management. To create a model for SARA system, we need to follow the steps of data modeling in figure 4.2 and those of ER model in figure 4.5.

Figure 4.5 ER Modeling design steps

4.4.2 Business requirement for SARA system

A strategy has been established to determine Rolls Royce's requirements and the objectives. First, we understand the operational processes in the current system; the inspection and repair; and their characteristics and constraints. Three mechanical parts are recommended as candidates for investigations: Fan blade, Curvic teeth of fan disc, and High pressure turbine (HPT) shaft. Several meetings are organized with the stakeholders, such as inspectors, operators, engineers, Material Review Board (MRB),etc.; to identify and characterize the necessary required data for the SARA system. We conclude that since SARA will be an automated system, the identification of the defect dimensions accurately in the model will shorten the inspection. Some mechanical parts are sent to the laboratory to confirm that the defects are beyond the limits. On the other hand, the data model will provide the necessary information for each stakeholder to prepare their reports and analysis. One of these stakeholders is the Quality engineer who is responsible for process monitoring and maintaining the inspection and repair process within the quality specifications based on Quality 4.0.

4.4.3 Conceptual ER model for SARA system

The conceptual ER model for the SARA system defines the system's concept by identifying the entities in addition to the relationships between them. The model has 4 main

stages: Sensor and measurement, inspection, sentencing, and repair. SARA model consists of nine entities as depicted in figure 4.6: (1) SENSOR, (2) P.C. DATA, (3) ORDER, (4) EQUIPMENT, (5) INSPECTION, (6) DEFECTS, (7) SENTENCING, (8) REPAIR, and (9) RE-INSPECT. The entities are described as the following:

1. SENSOR: This entity defines the sensors or measurement tools, which are responsible for providing the measurements and scanning of the inspected parts.

2. P.C. DATA: It refers to the point clouds (P.C.) for the surfaces of the inspected parts.

3. ORDER: It includes the essential information related to the engine and define the mechanical part(s) that is (are) required to be inspected.

4. EQUIPMENT: It means the mechanical parts that are required to be inspected. Therefore, this entity characterizes the details on the inspected mechanical part.

5. INSPECTION: This entity encapsulates the system objectives. It defines information related to the inspection of the part in terms of notifications and time duration taken for inspection.

6. DEFECT: It defines all information related to the characterizations of the detected defects if any.

7. SENTENCING: It provides information about the condition of the inspected mechanical part by the automated inspection machine whether being serviceable, repairable, or scrap.

8. REPAIR: If the decision for the inspection machine is that the inspected part is repairable, so the repair scheme and procedures will be defined automatically by the repair machine, then applied to the part.

9. RE-INSPECT: After repairing the inspected mechanical part, it is essential to check the repair procedures. Therefore, the inspected part will be sent to the inspection machine to be reinspected. Consequently, the RE-INSPECT entity defines the status of repair operation of the repaired part whether it is confirmed or not.

Furthermore, the SARA model also includes 13 relationships which represent the interactions between the entities (1) COMPOSE OF, (2) ASSIGN TO, (3) SEND TO, (4) CONTAIN, (5) DETECT, (6) PROVIDE, (7) Has, (8) Has, (9) SENTENCED BY, (10) ORDER TO, (11) REPAIRED BY, (12) RE-INSPECT, and (13) CONFIRMED BY to as depicted in figure

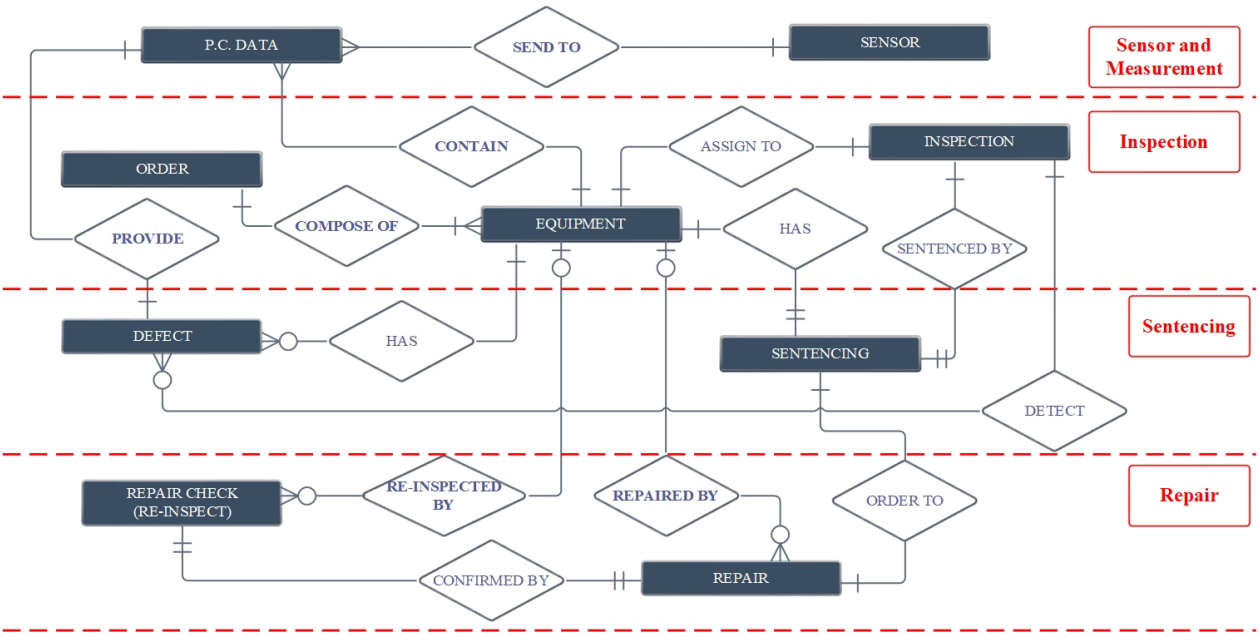


Figure 4.6 Conceptual ER Modeling for SARA system

4.6. The relationships and cardinality are obtained based on the constraints (business) rules required by the inspection and repair process. These constraints are as the following:

1. ORDER and EQUIPMENT: The inspection machine receives the orders; each order must compose of information of the inspected part(s) and the required task. One or many parts can be included in one order. Engine information is provided in the case of an internal order, where parts from an engine that is located on the shop floor are sent for inspection.

2. EQUIPMENT and INSPECTION: Each part is assigned to the inspection process. However, one or many parts may be included in one inspection order, each part is defined by its own inspection identifier. Therefore, each inspection identifier defines the inspection information of one part even the parts are included in the same order.

3. SENSOR and P.C. DATA: Each sensor in the inspection machine sends many measurements in terms of point clouds.

4. EQUIPMENT and P.C. DATA: Each inspected part contains many numbers of point clouds that describe its surfaces.

5. INSPECTION and DEFECT: Each inspection order may detect one or many numbers of defects per each inspected part.

6. DEFECT and P.C. DATA: P.C. DATA provides one point cloud for each defect detected on the inspected part.

7. EQUIPMENT and DEFECT: During the inspection, each inspected part may have one or more defects. Thus, the information of these defects are determined for each inspected part.

8. EQUIPMENT and SENTENCING: Each inspected part must have one and only one sentencing notification to identify its condition. Therefore, each notification must be assigned to one and only one inspected part.

9. INSPECTION and SENTENCING: Each inspection order has one and only one sentencing notification. In other words, each sentencing notification includes one inspection order. When many parts are included in one inspection order, each inspected part; defined by inspection identifier; must have one sentencing notification.

10. SENTENCING and REPAIR: When the sentencing notifies that the inspected part is repairable, one sentencing notification may be ordered to one repair order. The repair entity may include one sentencing notification, and none is in case of being scrapped part.

11. EQUIPMENT and REPAIR: One inspected part may be repaired by the repair machine in case of being repairable based on the sentencing notification in the sentencing entity. So, one repair order may include one and only one inspected part for repair. However, one or many parts may be included in one repair order, each part is defined by its own repair identifier.

12. EQUIPMENT and RE-INSPECT: One part may be reinspected by the inspection machine. One re-inspect notification may include one part. One or many parts may be included in one re-inspect order, each part is defined by its own re-inspect identifier.

13. REPAIR and RE-INSPECT: After repairing the repairable parts, they are sent to the inspection machine to confirm their repairs. Each re-inspect notification confirms each repair order.

4.4.4 Logical ER model for SARA system

The logical ER model is converting the conceptual one to data structures. Consequently, the attributes for each entity are identified to describe that entity based on the SARA's

requirements in term identification of the primary key(s) per each entity as shown in figure 4.7. The description of attributes per each entity are as follows:

1. SENSOR: A sensor ID acts as a primary key to identify the sensor name and its description. The sensor is used to scan and measure the points on the surfaces of the inspected part.

2. P.C. DATA: It includes an identifier the describes the zone that is represented by point clouds and the stored location where can be referred to these point clouds for understanding the characterizations of different types of defects if any.

3. ORDER: An order is carried out with an order number and at a defined date. The order number is considered the primary key of this entity because it is unique. The type of sales order may be internal or external. If it is an internal order, so the functional location, engine description, number of inspected parts, the owner and owner description must be defined. On the other hand, the external order concerns the information of the number of parts required to be inspected, and the required task whether, repair or overhaul.

4. EQUIPMENT: Each inspected part is defined by a unique identifier. Each inspected part has its own serial number(S/N). The reason for using an identifier instead of S/N as a primary key is that one part can be assigned to inspection, repair, and re-inspect several times, so the identifier differentiates between each time. Part and material numbers are specified based on the type of the inspected part such as fan blade, HPT shaft, or Curvic teeth of fan disc. The Part number can be modified after several repairs, so it is necessary to mention the last part number (LP/N) if any. Generally, these numbers are found in the engine manual. In addition to the inspected part information related to total Time Since New (TSN), Cycle Since New (CSN), Time Since Last Visit (TSLV), Cycle Since Last Visit (CSLV), and Time Since Overhaul (TSO).

5. INSPECTION: An inspection ID acts as a primary key to describe the notification number of the inspected part(s). In other words, if the inspection order includes many parts, so each inspected part has its own inspection ID. Moreover, it determines the time duration taken by the inspection machine. It includes the forecasting required time to finish the inspection.

6. DEFECT: The entity is defined with the Equipment ID of the inspected part which is considered a primary and foreign key at the same time. It is a foreign key because it

belongs originally to the EQUIPMENT entity as a primary key. On the other hand, it is a primary key, in addition to defect ID because defect ID can be repeated for different inspected parts. for example, defect ID#1 can be found in two or more inspected parts having different Equipment IDs. Thus, the Equipment ID and defect index describe the entity because they are unique together.

7. SENTENCING: Each inspected part has its sentencing notification that determines its condition. Plant and location attributes are used to mention that the process is carried out using the inspection machine in the SARA system. The decision type can be serviceable or repairable or scrap.

8. REPAIR: The entity is defined by repair ID because more than one inspected part may be subjected to the same repair procedures when the original task is an overhaul; for example, the 24 fan blades of the compressor. Therefore, these parts have the same repair order. Moreover, it identifies the repair scheme and procedures in addition to the duration time taken for repair and forecasting time.

9. RE-INSPECT: A re-inspect ID defines the entity. It provides a confirmation number for the repaired part(s) to demonstrate that the repair procedures are carried out as engine manual regulations. This is carried out by measuring the depth after repair. Plant and location attributes are used to mention that the process is carried out using the inspection machine in the SARA system. The duration times taken for re-inspect and forecasting time are determined.

Therefore, all the required information and business requirements related to the SARA system are captured by the proposed logical model. The model documents the inspection and repair process through the previously mentioned entities. It defines the attributes and primary key for each entity and clearly describes the relationships between these entities. To validate the SARA model, the relationships of the model should satisfy the business rules of the application. Figure 8 shows a transaction sample of the EQUIPMENT entity. This transaction presents the necessary information of the EQUIPMENT entity as in figure 4.7. Additionally, it shows its relationship with the other entities. The inspected part identified by "Equipment_ID =1" is ordered by the order number "Ordert_Order_No" = 14000" which is linked to the primary key "Order_No" of ORDER entity as shown in figure 4.9. The essential information of the engine and mechanical part are defined when exploring the "Order_No = 14000", similarly for the INSPECTION, SENTENCING, REPAIR, REINSPECT entities as depicted in figure 4.8.

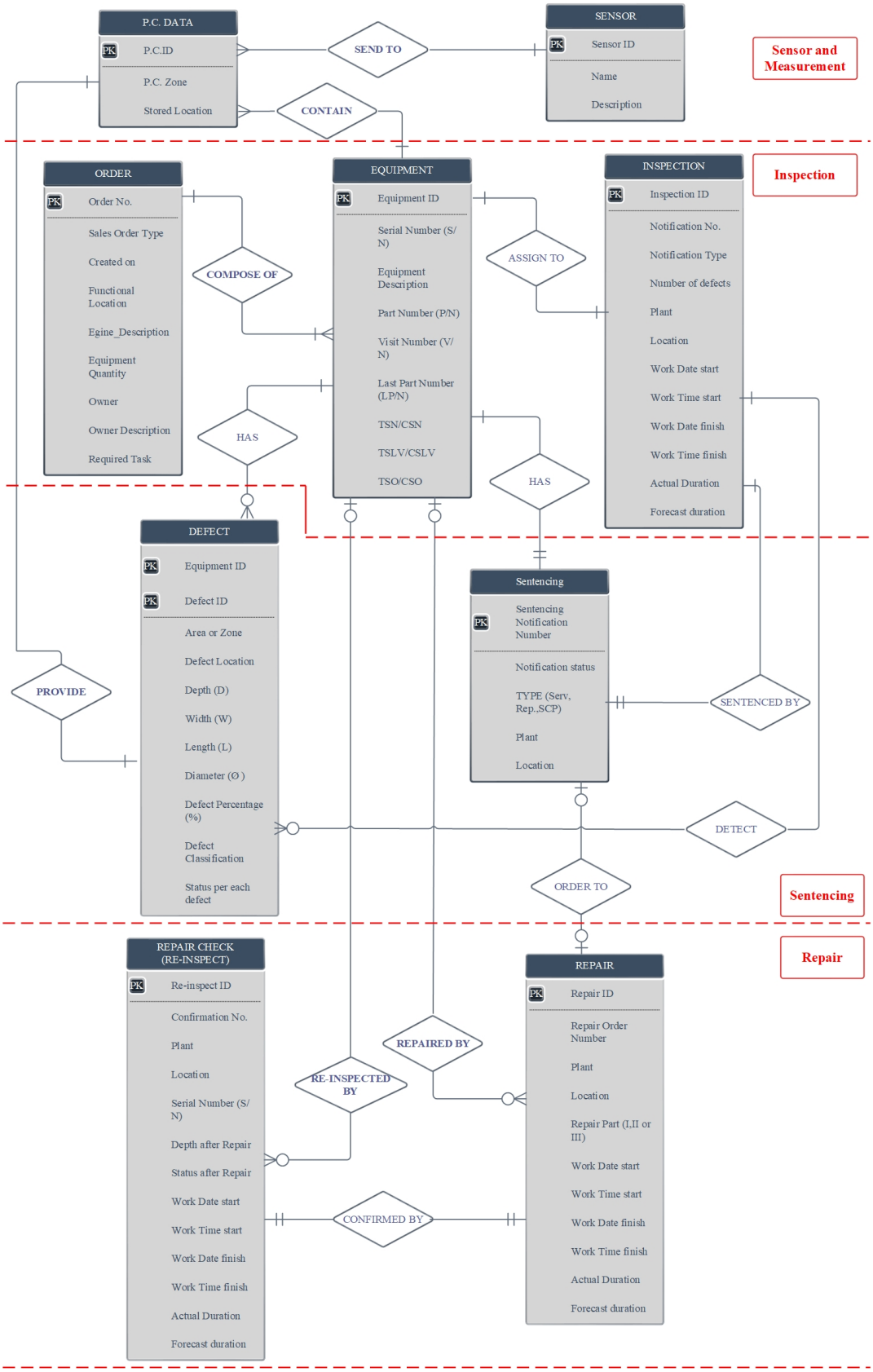


Figure 4.7 Logical ER Modeling for SARA system

phpMyAdmin

Current server: MySQL

Recent Favorites

New
information_schema
mysql
performance_schema
rr
rr
rr_old
Tables
New
defect
equipment
inspection
minor_record_sheet_r
ordert
reinspect
repairt
sentencing
Views
sys

Server: MySQL:3306 » Database: rr_old » Table: equipment

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges | Operations | Triggers

Showing rows 0 - 0 (1 total, Query took 0.0003 seconds.)

```
SELECT * FROM `equipment`
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

+ Options

	Equipment_ID Serial Numberof the inspected equipment	S_N Serial Numberof the inspected equipment	Equipment_Description Description of the inspected equipment	P_N Part number of the inspect equipment	V_N Current visit number of the inspect equipment	LP_N Last Serial number of the inspect equipment if it ...	TSV Time Since New	CSN Cycle Since New	TSLV Time Since Last Visit	CSLV Cycle Since Last Visit	TSO Time Since Overhaul
Edit Copy Delete	1	23995660	FAN BLADE	FW33513	1	0	4124.12	1207.00	0.00	0.00	0.00

Check all With selected: Edit Copy Delete Export

Show all | Number of rows: 25 Filter rows: Search this table

Query results operations

Print Copy to clipboard Export Display chart Create view

phpMyAdmin

Current server: MySQL

Recent Favorites

New
information_schema
mysql
performance_schema
rr
rr
rr_old
Tables
New
defect
equipment
inspection
minor_record_sheet_r
ordert
reinspect
repairt
sentencing
Views
sys

Server: MySQL:3306 » Database: rr_old » Table: equipment

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges | Operations | Triggers

CSO Cycle Since Overhaul	ordert_Order_No	inspection_Inspection_ID	sentencing_Sentencing_Notification_No	repairt_Repair_ID	reinspect_Reinspect_ID
0.00	14000	2	431083172	1	1

Figure 4.8 EQUIPMENT entity in the SARA Model

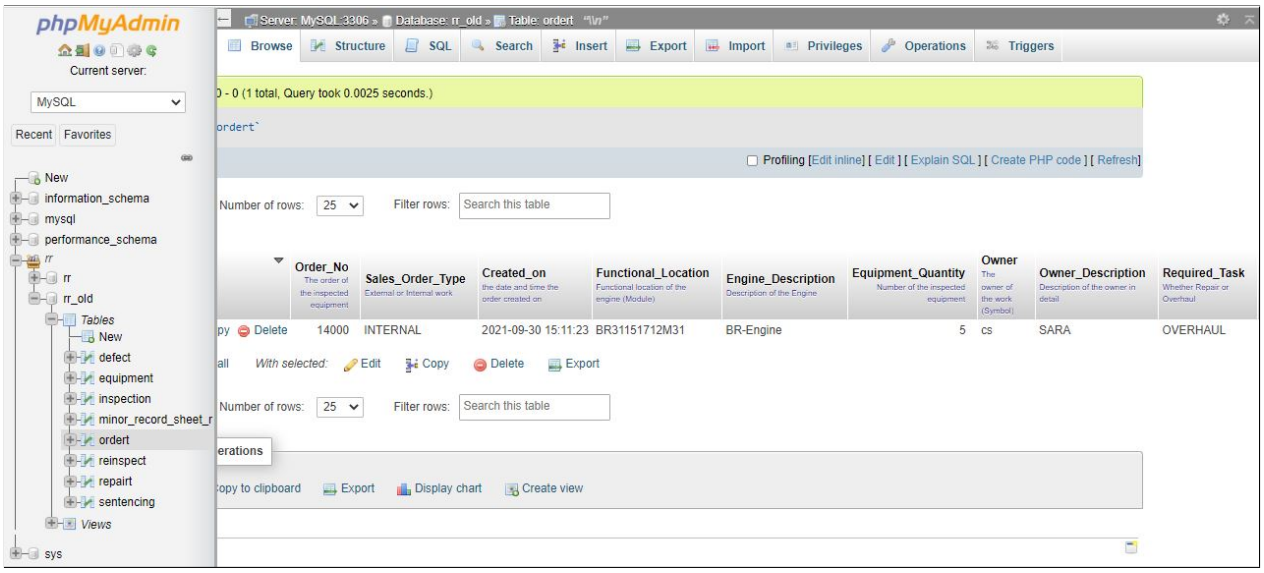


Figure 4.9 ORDER entity in the SARA Model

4.5 Discussion

The SARA model captures all the information relevant to the aspects of the SARA system based on the three axes. The model defines the criteria of the three candidate parts; fan blade, HPT shaft, and Curvic teeth of fan disc; and traces their inspection and repair processes. Moreover, it represents the data flow through the process. The ER modeling technique assists in mapping out and managing the information flow in database management of the SARA system by the representation of the conceptual and logical models. The relevant entities are defined in addition to the identification of the relationships between them. Each entity contains a set of attributes that clearly describe the characteristic of that entity. Consequently, it provides a representation of a structured database with high data quality in terms of completeness, consistency, accuracy, absence of data redundancy, and integrity. The model meets the requisite information for the stakeholders, which were checked with each stakeholder before. The constraints and business rules are implemented in the model to ensure the completeness and accuracy of fulfilling the mandatory data and check even the optional data (i.e., external order). Therefore, the model reduces data redundancy. Further, the constraints ensure the data entry type, format, and size. The data are consistent as the information is the same and synchronized across the model. The model ensures data integrity where the relationships connect and trace all data in the database. Finally, the model supports the automated system to store, organize and provide the necessary information required by stakeholders.

4.6 Conclusion

Quality 4.0 anticipates the digital transformation of quality management will improve the process and/or product quality and will increase productivity. Data automation supports the vision by monitoring the processes and collecting the necessary information. When the data management process is improved, it will be possible to identify the process variables that are required to be monitored by quality tools, for example, the control chart. ER modeling technique is widely used to design a data model that represents the data of the system in entity sets and describes their relationships. Moreover, it provides the key performance indices for quality to detect any anomalies during system operation. In this paper, we developed the conceptual and logical ER models for the SARA system which is an automated inspection and repair in the aerospace maintenance and repair domain. These models aim to organize the data in the system by identification of nine entities that represent the process, and their relationships with each other and their attributes. They characterize and control the information generated by the SARA system. Consequently, the data flow along the SARA model is visually represented and easily understood. The models favor accessibility, traceability, and reproducibility for better communication between inspectors and design specialists and tracking relevant information.

For future works, we will convert the logical *ER* model for the SARA system to a physical model that is implemented in the system. The physical data model represents the actual database design according to what is carried out in the logical one. The physical model defines the data types; numeric, string, date, or time, etc.; the foreign keys that link the entities, and the necessary constraints. It is used to support the implementation of the database using *DBMS* software. It provides a Data Definition Language (*DDL*) file, which is generated by MySQL Workbench software that will be implemented in the SARA system. Consequently, the automation of inspection and repair in the SARA system allows to enter the data that are defined in the model automatically, and provides the information that each stakeholder is interested in. After data preparation, we are targeting the implementation of the Quality 4.0 aspects. By providing the key performance indices for quality of inspection and repair actions, the information's variability of inspection and repair processes will be monitored by quality monitoring control tools based on machine learning and statistical control chart. This tool assesses the quality of the repair process based on the confirmation rate of the repaired parts to be within pre-specified limits. When the repair quality gets beyond these limits, the tool does not only detect abnormal behavior in the process but also, it determines the root cause of that out-of-control process. Then, corrective actions are taken to avoid several reworks of repair and maintain the repair process within the quality limits.

CHAPTER 5 ARTICLE 2: DEVELOPING MACHINE-LEARNING REGRESSION MODEL WITH LOGICAL ANALYSIS OF DATA (LAD)

Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto

Published in:
Computers & Industrial Engineering, 2021
(*DOI: 10.1016/j.cie.2020.106947*)

5.1 Abstract

This paper proposes a regression model based on Logical Analysis of Data (*LAD*). *LAD* is known as a combinatorial Boolean supervised data mining technique for pattern generation. It is used mainly for classification problems, and has demonstrated high accuracy compared to other classification techniques. In this paper, we extend the use of *LAD* to deal with supervised data with continuous responses. We derive a *LAD* regression model (*LADR*). Three discretization methods that transform the values of the response into a set of thresholds are tested. At each threshold, *LAD* analyzes the data as a two-class classification problem and extracts the prescriptive patterns for each class. *LADR* regression uses the generated patterns from the original data by using *cbmLAD software* to fit a numerical continuous dependent response. Therefore, a normalized regression model with only binary independent variables is obtained. *LADR* has been applied for six datasets and obtains better results compared with the linear regression (LR), support vector regression (SVR), Decision Tree Regression (DTR), Random Forest (RF), and Polynomial Regression (PolyR). The performance is evaluated by the Mean Square Error (MSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE) based on a 10-fold cross validation.

Keyword: Regression techniques, Logical Analysis of Data (*LAD*), *LADR* regression, Discretization methods, Combinatorial Regression (CR)

5.2 Introduction

Recently, companies have started to examine and enable Industry 4.0 technologies to monitor and control the manufacturing process [111, 123]. A new paradigm has been introduced under the title of Industry 4.0 called "Quality 4.0". Quality 4.0 serves manufacturers with maintaining and improving quality management by using machine learning techniques, which use online sensor data to monitor process performance [10]. Data analytic is one common feature for both Industry 4.0 and Quality 4.0. Thus, companies are striving towards building their own analytic strategy to analyze the available data and to extract useful knowledge.

As the volume of data increases, it becomes difficult to analyze it with traditional statistical tools, such as ANOVA, control charts and statistical regression modeling. Data mining becomes useful in this situation. This is the process of extracting useful information in the form of patterns from the data [124, 125]. These patterns have physical meanings that describe and interpret the hidden events and phenomena that are taking place. Hence, they are called prescriptive patterns, as they identify interpretable knowledge [126]. Machine learning provides the technical tools for data mining. It is a form of artificial intelligence

that applies a variety of algorithms to analyze the data [76]. It has a set of techniques that are implemented to describe the relationships within the dataset and to exploit these relationships for diagnosis and prognosis.

In this paper, we construct a regression model to predict process performance based on patterns that are extracted by the data mining of sensors' readings. These patterns describe the hidden phenomenon of natural and abnormal variation in process performance. The patterns are generated by using Logical Analysis of Data (*LAD*), which is a non-statistical supervised data mining technique for pattern generation [127]. The concept of *LAD* is based on Boolean logic and combinatorial optimization, which is adapted to classification problems [128]. *LAD* is used to extract patterns from the training set of the original data. These patterns are hidden rules that differentiate between classes. These rules are also used for the prediction of certain events such as failure, and for anomalies' detection [129]. The generated patterns are characterized by the following [130]: (1) degree (d) is the number of variables that define the pattern, (2) prevalence of the pattern is the proportion of the positive (negative) observations that are covered by the pattern to the total positive (negative) observations, (3) homogeneity of a pattern is the proportion of the positive (negative) observations to total observations (positive and negative) that are covered by the pattern. It has been proven that the most explanatory pattern has a low degree, high prevalence and high homogeneity. A strong pattern is one that has the highest prevalence amongst all patterns in its class. This characteristic is needed to provide a robust pattern, which is capable of explaining a phenomenon, particularly in the case of noisy data [131]. *LAD* was employed for various classification problems in two classes [127] and a multi-class [132]. It was implemented in condition based maintenance (CBM) applications [24], in financial applications [133], in industrial chemical processes [134], in the airline industry [135], and in medical applications [136]. *LAD* has the advantage of (1) generating prescriptive patterns that have physical meaning, and (2) it is not based on any statistical assumptions, as is usually the case with statistical techniques.

The *LAD* classification technique has demonstrated high accuracy compared to other classification techniques. Two approaches were proposed in the literature to create *LAD* regression models. The first approach is a Pseudo Boolean Regression model (PBR) [137], which uses the generated patterns to build the regression model based on a minimization of the Least Absolute Residual (LAR). LAR is the sum of the absolute residual value between the actual values and the model responses. In this approach, the number of the generated patterns is very large even for a small size of datasets. In order to reduce computational efforts, a Column generation algorithm is used to select the optimal set of generated patterns. These patterns are used in the regression algorithm. This algorithm represents an optimization problem that

is solved by using linear programming based on LAR criterion. The second approach is a Combinatorial Regression (CR) that was proposed as a new extension of the standard *LAD* classification technique [138]. The original classification problem is transformed into several classification sub-problems based on a continuous numerical response. In other words, CR chooses the number of thresholds on that response to define these sub-problems. For each sub-problem, *LAD* is used to generate the patterns. Consequently, all of the generated patterns are gathered. CR finds the coefficients of the independent variables of the regression model by minimizing the sum of the square residual (MSE) criterion.

The two approaches have performance that is comparable to other regression techniques, PBR [137] and CR [138]. In fact, CR and PBR are closely related, as they are constructed based on data transformation by mapping non-binary data into Q-dimensional $\{0,1\}^Q$. Although both techniques are similar in process and have the same form of results, the CR differs in the techniques of pattern generation and the calculation of the coefficients or weights of the regression model, which are called the loss function. Both methods have limitations and ambiguities. The PBR is limited to the degree of the generated patterns of, at most 3, and the CR relies on selecting the number of thresholds used for obtaining the classification sub-problems. Nevertheless the rule for selecting these thresholds is not provided in [138].

This paper presents a *LAD* Regression (*LADR*) technique for building regression models. *LADR* overcomes the limitations of the previous two techniques as shown in table 5.1 by providing its four novelties: (1) the discretization process that represents the cornerstone of the technique. It is implemented to maintain a pre-specified number of thresholds to obtain several classification sub-problems based on the response. This paper introduces three discretization methods; Equal width *(EW)*, K-means *(KM), and Percentage of standard deviation* (%STD)*;* that present the selection rule of the thresholds. (2) The generation of strong patterns that cover all observations in the dataset, without restriction on the degree of these patterns by using *cbmLAD software* [24]. (3) The data preparation and pre-processing of the generated patterns. (4) The development of the regression model and the evaluation of its performance. The *LADR* approach determines the best discretization method and the appropriate number of intervals (or thresholds) that provide a high accuracy of the regression model.

This paper is organized as follows. Section 2 presents the *LADR* technique and discusses how it strengthens the regression model. Section 3 evaluates the performance of the *LADR* by using different datasets. Section 4 provides a comparison with the results of other regression techniques, in addition to the combinatorial regression method. Finally, section 5 concludes the paper.

Table 5.1 Addressing the research gaps

Research gaps	Addressing the gaps
Limitation on the degree of the generated patterns that provide less accurate regression model.	Generation of strong patterns with different degrees that describe the data.
The rule for selecting the thresholds for obtaining the classification sub-problems is not mentioned.	Three discretization methods *EW*, *KM*, and *%STD* are proposed to obtain these thresholds

5.3 *LADR* regression

LAD is one of the knowledge discovery approaches that extracts the hidden patterns in the dataset and constructs a theory formulation for the prediction of future events. *LAD* is considered a supervised learning technique because it relies on historical data with labeled classes [139]. It is mainly used for classification problems. The three main steps of *LAD* are the binarization of the dataset's variables, generation or extraction of patterns, and theory formulation by defining a discriminant function as shown in figure 5.1. Further details are found in [140], in addition to illustrative example in [141].

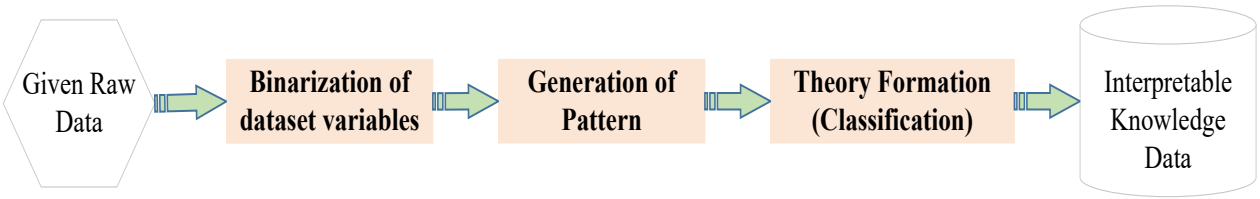


Figure 5.1 Steps of a *LAD* approach

The objective of this paper is to transform the standard *LAD* to a regression modeling technique to solve various regression problems. It aims to exploit the strength of *LAD* to extract patterns that cover all of the observations in the given dataset. Accordingly, the independent variables of the original data are transformed into patterns P_j , j=1,. . . ,J , and the dependent response ,Y, is regressed on the patterns that cover the observations instead of the observations themselves. It has been shown in PBR [137] and CR [138] that the regression model, which is based on patterns, is more accurate than those that are based on the observations themselves. While observations may be noisy or inaccurate, patterns, and in particular, strong patterns, are more robust because they characterize a group of observations. The X_{P_j} is a binary variable that indicates whether pattern j covers an observation or not. These patterns are used as the independent variables of the regression model that minimize the MSE, as shown in figure 5.2.

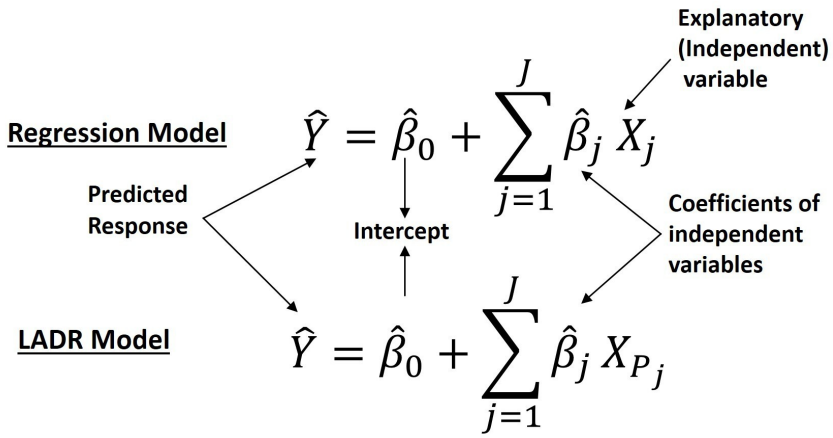


Figure 5.2 Difference between *LAD* and linear regression form

Although the dependent response is in a numerical and continuous form, discretization methods are used to allow the generation of patterns that characterize the variability of the response. Discretization methods are used to select several thresholds based on the response and transform the dataset Ω to several sub-problems. For each threshold, the dataset is divided into two classes. The observations (ω) that have response values below that threshold constitute the first class, while the others are in the second class. Therefore, the independent variables are binarized at each threshold, and consequently, *LAD* methodology is applied to generate patterns as two-class classification. The key point is to obtain the appropriate number of thresholds that provides the most accurate model. Since the selection of threshold criterion is not defined in the CR method, in this paper, we propose *EW*, *KM*, and $\%STD$ based on the response. Therefore, we propose a *LADR* regression technique that selects the best discretization method and provides better results compared to the other regression techniques, including CR.

5.3.1 *LADR* methodology

The *LADR* technique consists of four main steps: (1) threshold selection, by dividing the data into N-intervals to maintain the thresholds. (2) Pattern generation, by using the *LAD* methodology and *"cbmLAD software"* [24] to generate patterns at each threshold. (3) Data preparation by gathering the patterns in a single patterns' set to obtain the independent variables X_{P_j}, j=1,...,J , instead of X_j, then the application of data processing on this data. (4) Modeling and validation, by obtaining the regression model and by evaluating its performance and accuracy

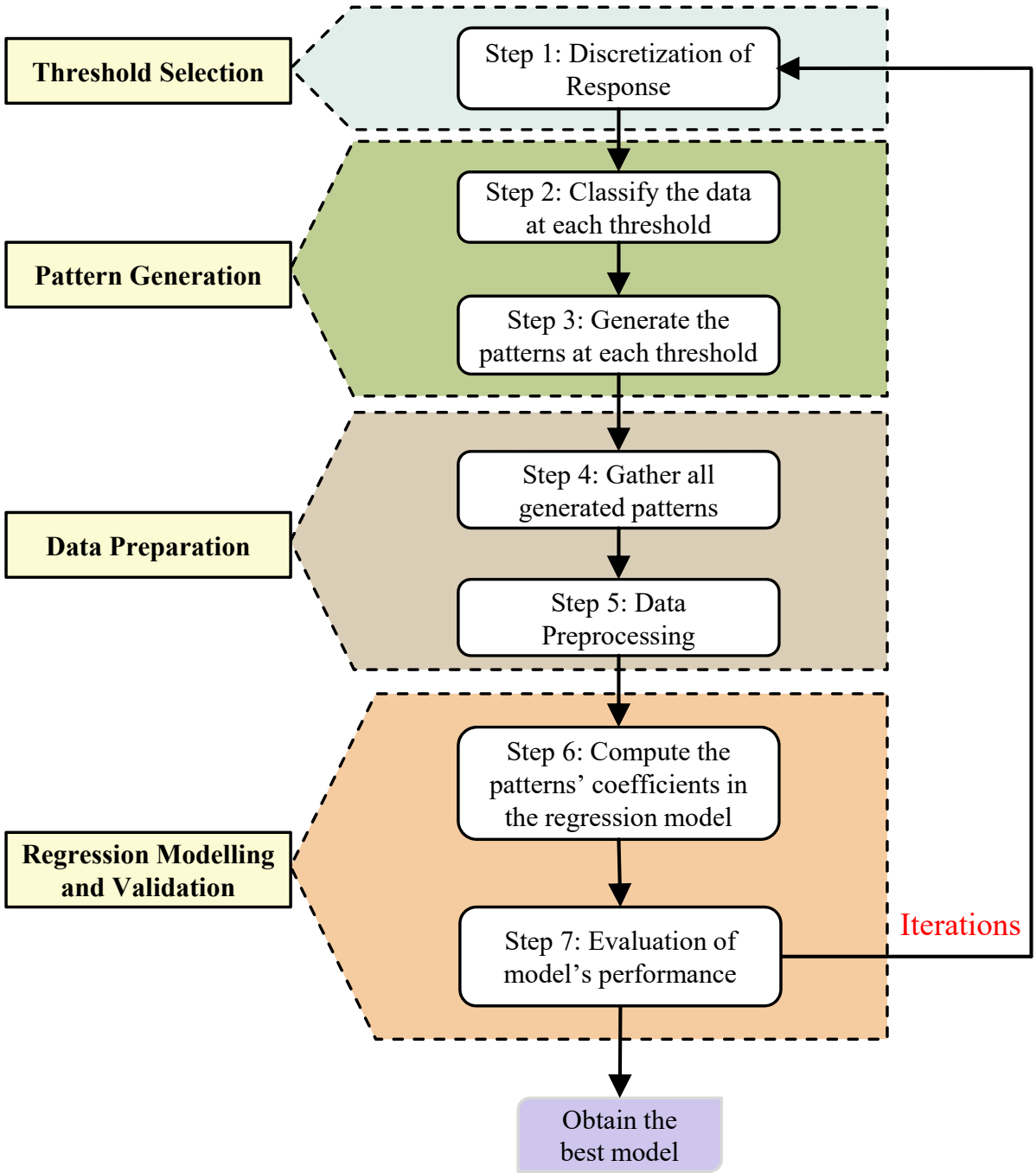


Figure 5.3 A diagram of the *LADR* methodology

Step 1: Discretization of the response

The discretization process partitions the values of the response Y by using thresholds. These thresholds gather the values of the response into a set of intervals. We consider a dataset $\Omega$of n independent variables with m observations. We sort the data in ascending order based

on the values of the response Y. We search for I-thresholds τ_1, τ_2,..., τ_I on the response. The following process was performed to maintain the appropriate number of thresholds that could provide a regression model that has a good fit [142, 143]:

1. Equal width (EW):

   Creating the same interval width by dividing the range of original ordered response's values into N-intervals. The first threshold is determined by the first value of the second interval, and so on. The value of N is obtained by iteration.

2. K-means Clustering (KM):

   Creating N-intervals based on the construction of N-clusters that minimize the sum of the distances of each response's value to the gravity center of its cluster [144, 145]. The kth threshold is the first value of the $(k+1)^{th}$ cluster.

3. Percentage of standard deviation of the response *(%STD)*:

   Selecting a pre-specified percentage of the standard deviation (STD) of the response, by sorting the values of the response variable. For each consecutive values u and v, where u <v if the v-u >τ, a new threshold, which is the average of the consecutive values, is introduced [137].

In this paper, we compare the proposed discretization methods with a natural one, which is called Quantile method (QT). It considers the data distribution and creates N-intervals based on quantiles. The range of the quantiles is]0,1[. The QT is used to divide the values of response using different steps: 0.01, 0.02, 0.05, 0.1, 0.2, 0.25, and 0.5, which obtain number of thresholds: 99, 49, 19, 9, 4, 3, and 1 respectively.

Table 5.2 presents a dataset as an example to illustrate the three methods. The dataset contains only one independent variable X and one dependent response Y. The values of the response are sorted ascendingly. Let us assume splitting the response values into five intervals ($N = 5$) for both EW and KM. In this case, four thresholds are obtained for both methods. For the third method, we assume 15% standard deviation of the response. The thresholds are created when the condition that u-v is greater than that percentage is satisfied. The 15%STD criterion creates 12 thresholds. Refer to QT, assume the step is 0.2 that constructs four thresholds.

Each discretization method creates a set of thresholds on the response of the dataset. It prepares the data for *cbmLAD software* [24] to generate the patterns in the second step. The selection of the best discretization method and the appropriate number of thresholds for the dataset will be mentioned later.

Table 5.2 The thresholds that are identified using *EW*, *KM*, 15%*STD*, and *QT* with step 0.2 method

X	Y	*EW*		*KM*		*%STD*		*QT*	
		N	Threshold	N	Threshold	u-v	Threshold	N	Threshold
0.95	87.33	1		1				1	
0.87	87.59	1		1		0.26		1	
1.02	89.05	1		1		1.46	$\tau_1 \geqslant 88.32$	1	
1.01	89.54	1		1		0.49	$\tau_2 \geqslant 89.30$	1	
1.11	89.85	2	$\tau_1 \geqslant 89.75$	1		0.31	$\tau_3 \geqslant 89.70$	2	$\tau_1 \geqslant 89.79$
0.99	90.01	2		1		0.16		2	
1.2	90.39	2		1		0.38	$\tau_4 \geqslant 90.20$	2	
0.98	90.56	2		1		0.17		2	
1.15	91.43	2		2	$\tau_1 > 90.56$	0.87	$\tau_5 \geqslant 91.00$	3	$\tau_2 \geqslant 91.08$
1.23	91.77	2		2		0.34	$\tau_6 \geqslant 91.60$	3	
1.15	92.52	3	$\tau_2 \geqslant 92.17$	2		0.75	$\tau_7 \geqslant 92.15$	3	
1.26	93.25	3		3	$\tau_2 > 92.52$	0.73	$\tau_8 \geqslant 92.89$	3	
1.32	93.41	3		3		0.16		4	$\tau_3 \geqslant 93.31$
1.19	93.54	3		3		0.13		4	
1.4	93.65	3		3		0.11		4	
1.29	93.74	3		3		0.09		4	
1.36	94.45	3		3		0.71	$\tau_9 \geqslant 94.10$	5	$\tau_4 \geqslant 93.88$
1.43	94.98	4	$\tau_3 \geqslant 94.59$	3		0.53	$\tau_{10} \geqslant 94.72$	5	
1.46	96.73	4		4	$\tau_{13} > 94.98$	1.75	$\tau_{11} \geqslant 95.86$	5	
1.55	99.42	5	$\tau_4 \geqslant 97.01$	5	$\tau_4 \geqslant 96.73$	2.69	$\tau_{12} \geqslant 98.08$	5	

Step 2: Classify the data at each threshold

After the selection of the discretization method, the data is classified for each threshold, τ_i, into two sets: Set Ω_i^+ of positive observations, which have responses that are greater or equal to the value of that threshold, and set Ω_i^- of negative observations. $Y(\omega)$ is the response at the observation ω, and i=1,...,I are the thresholds.

$$\begin{aligned} \Omega_i^+ &= \{\omega \in \Omega | Y(\omega) \geqslant \tau_i\}, i = 1, ..., I \\ \Omega_i^- &= \{\omega \in \Omega | Y(\omega) < \tau_i\}, i = 1, ..., I \end{aligned} \tag{5.1}$$

In table 5.2, the *(%STD)* method obtains 12 thresholds at 15%. For the first threshold (τ_1), the response is classified in Ω_1^+ if its value is greater or equal to 88.32, otherwise it is classified in Ω_1^-. Thus, the positive class starts from Y= 89.05 to the last observation 99.42, while the negative class consists of the first two observations (Y=87.33 and 87.59). This procedure is repeated for all thresholds. Therefore, a table is created for each threshold to identify the positive and negative observations. Twelve tables, containing positive and negative observations, are created. For instance, table 5.3 shows only three of the tables.

Step 3: Generate patterns at each threshold

Based on the previous step, and for the three datasets that are shown in table 5.3, *"cbm-LAD software"* [24] acts as a two-class classifier. The patterns are generated by solving the

following optimization problem:

$$\text{maximize} \sum_{\phi \in \Omega^+} \prod_{\substack{j=1 \\ \phi_j \neq \psi_j}}^{n} (1 - y_j) \tag{5.2}$$

$$\text{subject to} \sum_{\substack{j=1 \\ \gamma_j \neq \psi_j}} y_j \geqslant 1 \quad , \forall \gamma \in \Omega^- \tag{5.3}$$

$$y_j \in 0, 1 \quad , \forall j = 1, ..., n \tag{5.4}$$

Where n: number of attributes,
$y_j = 1$ if attribute j is included in the pattern with a value equals ψ_j, 0 otherwise,
ψ_j : the value of the jth attribute in the ψ observation, where j = 1,...,n ,

Table 5.3 Three tables defining the positive and negative classes that are obtained by the first three thresholds of the dataset from table 5.2, when using the 15%STD method. (Class=0 :Positive observations , Class=1: Negative observations)

Threshold 1 (τ_1)		
Class	X	Y
1	0.95	87.33
1	0.87	87.59
0	1.02	89.05
0	1.01	89.54
0	1.11	89.85
0	0.99	90.01
0	1.2	90.39
0	0.98	90.56
0	1.15	91.43
0	1.23	91.77
0	1.15	92.52
0	1.26	93.25
0	1.32	93.41
0	1.19	93.54
0	1.4	93.65
0	1.29	93.74
0	1.36	94.45
0	1.43	94.98
0	1.46	96.73
0	1.55	99.42

Threshold 2 (τ_2)		
Class	X	Y
1	0.95	87.33
1	0.87	87.59
1	1.02	89.05
0	1.01	89.54
0	1.11	89.85
0	0.99	90.01
0	1.2	90.39
0	0.98	90.56
0	1.15	91.43
0	1.23	91.77
0	1.15	92.52
0	1.26	93.25
0	1.32	93.41
0	1.19	93.54
0	1.4	93.65
0	1.29	93.74
0	1.36	94.45
0	1.43	94.98
0	1.46	96.73
0	1.55	99.42

Threshold 3 (τ_3)		
Class	X	Y
1	0.95	87.33
1	0.87	87.59
1	1.02	89.05
1	1.01	89.54
0	1.11	89.85
0	0.99	90.01
0	1.2	90.39
0	0.98	90.56
0	1.15	91.43
0	1.23	91.77
0	1.15	92.52
0	1.26	93.25
0	1.32	93.41
0	1.19	93.54
0	1.4	93.65
0	1.29	93.74
0	1.36	94.45
0	1.43	94.98
0	1.46	96.73
0	1.55	99.42

ϕ_j : the value of the jth attribute in the ϕ observation, where $\phi \in \Omega^+$ and $j = 1,...,n$,
γ_j : the value of the jth attribute in the γ observation, where $\gamma \in \Omega^-$ and $j = 1,...,n$.

It generates patterns that characterize the observations in the dataset, and distinguishes between two classes. Positive and negative patterns for each dataset are generated at the threshold *"i"*. A Positive (negative) pattern is a conjunction of binary attributes which covers at least one positive (negative) observation, while it is not valid for all negative (positive) observations in the dataset. Thus, a pattern set (Pi) is formed of positive and negative patterns that are generated at each threshold "i", where i=1,...,I, as given in equation (5.5).

$$Pi = \{Pi^+\} \cup \{Pi^-\}, i = 1, ..., I \tag{5.5}$$

Refer to our example at the first threshold τ_1 in table 2, the *cbmLAD* obtains a pattern set $P1$=$\{ Y \geq 88.32\} \cup \{ Y < 88.32\}$.

Step 4: Gather all generated patterns

In this step, all of the generated patterns at all thresholds are gathered in a single dataset $\mathcal{P}$=$\cup Pi$, i=1,...,I by dropping the notion of positivity and negativity of these patterns. Subsequently, each pattern in $\mathcal{P}$ is represented by an independent variable X_{P_j} where j=1,...,J is the pattern index.

Table 5.4 presents the generated patterns for the first three thresholds of the 15%STD that is applied on the illustrative dataset. When an observation is covered by a pattern j ,X_{P_j} takes the value of 1, otherwise its value is 0. As in the example, the first observation is covered by patterns 2, 5, and 9, while it is not covered by patterns 1, 3, 4, 6, 7, and 8.

Step 5: Data pre-processing

Data pre-processing is implemented on the independent variables of the *LADR* regression model, which correspond to the generated patterns. Data pre-processing is carried in five stages to detect the duplicated patterns, the correlated patterns, the dependencies between patterns, the multi-collinearity phenomena, and insignificant patterns in the regression model as shown in figure 5.4.

(A) Removal of the duplicated patterns

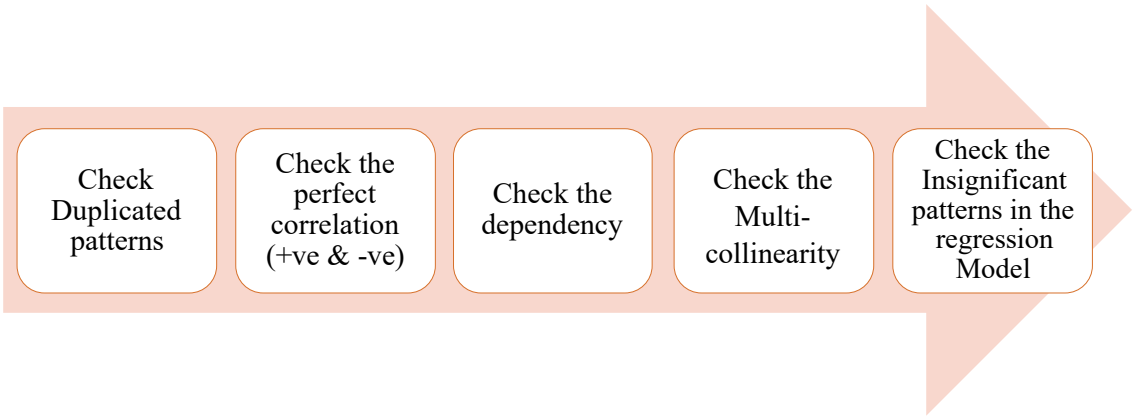


Figure 5.4 Data processing steps

Table 5.4 The generated patterns at the first three thresholds of the 15%STD method for the illustrative dataset and the binary values of the patterns' independent variables X_{P_j},j=1,. . . ,10

Threshold 1 (τ_1)		Threshold 2 (τ_2)				Threshold 3 (τ_3)			
X>0.965	X<0.965	X>0.965 X<1.015	X>1.065	X<0.965	X>1.015 X<1.065	X>1.065	X>0.965 X<1	X<0.965	X>1 X<1.065
X_{P_1}	X_{P_2}	X_{P_3}	X_{P_4}	X_{P_5}	X_{P_6}	X_{P_7}	X_{P_8}	X_{P_9}	$X_{P_{10}}$
0	1	0	0	1	0	0	0	1	0
0	1	0	0	1	0	0	0	1	0
1	0	0	0	0	1	0	0	0	1
1	0	1	0	0	0	0	0	0	1
1	0	0	1	0	0	1	0	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	1	0	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0

The same pattern can be generated in more than one dataset. For example, in table 5.4, pattern 2 (X_{P_2}) is the same as pattern 5 and 9 ($X_{P_{5\&9}}$), and patterns 4 and 7 ($X_{P_{4\&7}}$) are duplicated. In this case, one of them is kept in the pattern set $\mathcal{P}$, and the other patterns are removed.

(B) Correlated patterns

Gathering a large number of patterns in the previous steps allows for the presence

of correlated patterns. The correlation coefficient (R) is a statistical measure that determines the strength of the linear relationship between the variables [146]. The R ranges between -1 to 1 and is formulated as follows:

$$\mathrm{R}_{X_{P_j}X_{P_h}} = \frac{m\sum_{i=0}^{m} X_{P_j}X_{P_j} - \sum_{i=0}^{m} X_{P_j}\sum_{i=0}^{m} X_{P_h}}{\sqrt{(m\sum_{i=0}^{m} X_{P_j}^2 - (\sum_{i=0}^{m} X_{P_j})^2)(m\sum_{i=0}^{m} X_{P_h}^2 - (\sum_{i=0}^{m} X_{P_h})^2)}}$$

$$\forall j\&h = 1, \cdots, J \text{ where } j \neq h$$

$$\mathrm{R}_{X_{P_j}X_{P_h}} = \begin{cases} 1 & \text{Prefect positive correlation} \\ 0 & \text{No correlation} \\ -1 & \text{Prefect negative correlation} \end{cases} \tag{5.6}$$

Where X_{P_j} and X_{P_h} are any pair of patterns and $\mathrm{R}_{X_{P_j}X_{P_h}}$ is the measure of correlation between patterns X_{P_j} and X_{P_h}.

When the patterns are correlated, the accuracy of the coefficients in the regression model decreases, causing a large discrepancy. In this stage, a correlation matrix of the patterns is constructed to trace the correlated patterns that have R-values of 1 and -1. R equals 1 means two patterns cover the same observations. On the other hand, if R is equal to -1, this means that one pattern is the complement of the other one. In both cases, one of the two patterns is removed. In table 5.4, patterns 1 and 2 ($X_{P_{1\&2}}$) are perfectly correlated patterns, which may affect the model. Since pattern 2 (X_{P_2}) is the complement of pattern 1 (X_{P_1}), X_{P_2} can be removed.

(C) Dependent patterns

Dependencies may occur among patterns where the values of a pattern may depend on another pattern(s). We use equation (5.7) to calculate the coefficient of determination (R^2), which is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables. In practice, a pattern is considered dependent on another if its R^2 is greater to or equal to 90%. These procedures are repeated for all patterns.

$$\mathrm{R}^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y}_i)^2} \tag{5.7}$$

Where Y_i and $\hat{Y}_i$ is the actual value and predicted value of the response at the ith observation and $\bar{Y}_i$ is the mean value of the responses' values.

(D) Multi-collinearity

A correlation coefficient R that is higher than 90% may cause a multi-collinearity problem [147]. Multi-collinearity exists when two or more patterns have strong linear correlations to one another [148]. A variance inflation factor (VIF) [147,149] is used to verify the multi-collinearity phenomenon. It is used to measure how the coefficients of a regression model are inflated as a result of one or more collinear independent variables.

$$\mathrm{VIF}_j = \frac{1}{1 - \mathrm{R}_j^2} \tag{5.8}$$

Where R_j^2 is the coefficient of determination obtained by regressing the jth pattern on the other patterns, while ignoring the response Y. If the pattern has VIF >10 [150] which is equivalent to R_j^2=0.9, this implies a multi-collinearity, and this pattern is removed. This procedure is performed sequentially, and after each pattern's removal, a new regression model is obtained.

(E) Insignificant patterns

By calculating the p-value using a F-test, patterns with insignificant contribution to the regression model are those with p-value>0.05.

Step 6: Compute the patterns' coefficients in the regression model

In this step, the patterns that remain after the pre-processing step form the regression model. The independent variables of that model are the binary variables, (X_{P_j}), j=1,...,J of the remaining patterns, while the original response Y, is the dependent variable. The weights or coefficients of these patterns are calculated by minimizing the value of mean square error (MSE) as depicted in figure 5.2. The MSE is the average of square difference between the true and the model responses. In this paper, we use the linear regression algorithm to build the model. The MSE, R^2 and MAE are calculated for the obtained model, where MAE is calculated using equation (5.11).

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} (\mathrm{MSE}) \tag{5.9}$$

$$\mathrm{MSE} = \frac{\sum_i^m (Y_i - \hat{Y}_i)^2}{m} \tag{5.10}$$

$$\text{MAE} = \frac{\mid \sum_i^m (Y_i - \hat{Y}_i) \mid}{m} \tag{5.11}$$

Where $\hat{\boldsymbol{\beta}}$ is the vector of coefficients of the regression model.

Step 7: Evaluation of the model's performance

k-fold cross-validation is used to estimate the coefficient of determination (R^2), the mean square test error (MSE), and mean absolute error (MAE). The k-fold cross-validation splits the given data into the required k-folds, where each fold is a good representation of the whole data. In practice, 5 to 10 folds are the typical values that are performed in a k-folds cross validation. In a k- fold cross validation, the model is constructed by using (k-1) fold, and it is tested by using the remaining fold. This procedure is repeated k times and the average of R^2, MSE, and MAE calculated for assessment. After consecutive iterations using different number of thresholds with each discretization method, we obtain the best model that has the lowest MSE. Therefore, we determine the appropriate number of thresholds that explain the given dataset.

5.4 Performance of the *LADR*

To validate and evaluate the performance of the *LADR*, we implement it to six different known datasets. These datasets are *Boston housing*, *Computer Hardware*, *Auto-mpg* , *Servo*, *Airfoil Self-Noise*, and *Concrete Strength*, which are found in the UCI Machine Learning Repository [151]. Researchers use these datasets to compare their techniques and algorithms. All of the datasets contain one dependent variable. The characteristics of these datasets are presented in table 5.5. In addition, all the features for each dataset are identified in appendix A.

Table 5.5 Characteristics of the four datasets

Dataset	No. of observations	No. of independent variables	Range of response
Boston housing	506	13	[5 – 50]
Computer Hardware	209	6	[6 – 1150]
Auto-mpg	398	6	[9 - 46.5]
Servo	167	4	[0.13 – 7.1]
Airfoil Self-Noise	1503	5	[103.38 – 140.98]
Concrete Strength	1030	8	[2.33 – 82.60]

A pre-processing step is applied for these datasets. The aim of this step is to detect any missing values, outliers, duplicated instances or noise, and to apply any necessary data transformation, whether for response or independent variables, or for both [152, 153].

We seek a performance assessment of each of the three discretization methods, and its effect on the accuracy of a regression model. The *LADR* technique is an integration of the R-script using *R-studio software* [154] and *cbmLAD software* [24]. The best results are based on finding an appropriate number of thresholds and the discretization method that captures the variability in the given data.

The performance results of all models are represented in terms of MSE, R^2, and MAE. These are calculated based on the average of 10-fold cross validation. In cross validation, MSE is used as an appropriate metric of the true test error. It evaluates the predictive performance of the model when new observations are assessed. Consequently, it detects the presence of overfitting in that model, if any. Therefore, the lower MSE, the higher prediction of the model. In the following figures, we describe these metrics of the *LADR* model per each threshold for both the *KM* and *EW* methods. The results are explored for 1 to maximum 14 thresholds, which are equivalent to 15-intervals and clusters for *KM* and *EW* methods, respectively. For the STD method, we evaluate the model performance per each percentage of the standard deviation of the data's response. The maximum percentage of the standard deviation is 15%. On the other hand, the *QT* is carried out based on the step values as previously mentioned in step 1 in section 2. The selection of the best model for each discretization method is based on the MSE, which has a minimum value.

Figures 5.5, 5.6, 5.7 and 5.8 present the results of *LADR* models based on *KM*, *EW*, %*STD*, and *QT*, respectively for *Boston Housing* dataset. These figures demonstrate a variation in results using different methods at different thresholds. According to the statistical results of LADR-*KM* models in figure 5.5, the K-means with 5 clusters (*KM*=5) provides the most accurate predictive model. The *KM*=5 clusters of the data are based on its response in 4 thresholds. Figure 5.6 depicts that the best LADR-*EW* occurs when dividing the data into 11 equal width intervals (*EW*=11), which is equivalent to 10 thresholds. In figure 5.7, clustering the data based on 2% of the standard deviation of the response leads to the appropriate number of thresholds. At step 0.1, the best LADR-*QT* is obtained as shown in figure 5.8.

Table 5.6 shows the values of the performance measures, MSE, MAE and R^2 when using the best LADR-*KM*, LADR-*EW* and LADR-*STD*. The three proposed methods have better performance than the best LADR-*QT* model using step 0.1 as shown in figure 5.8. For further validation of the results, we implement five other different regression techniques. These

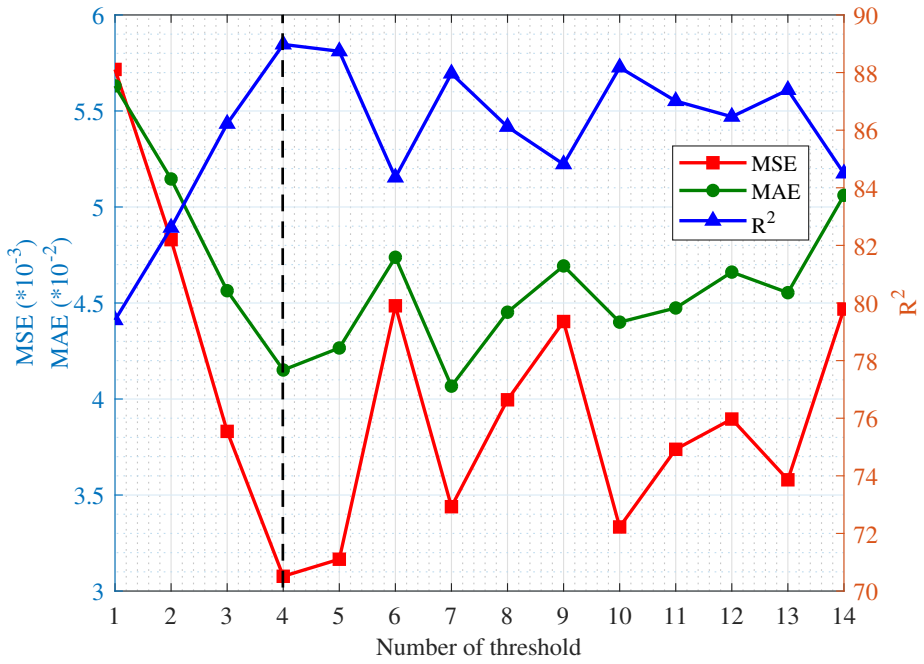


Figure 5.5 *Boston Housing* : the *LADR* measures of performance using the *KM* method

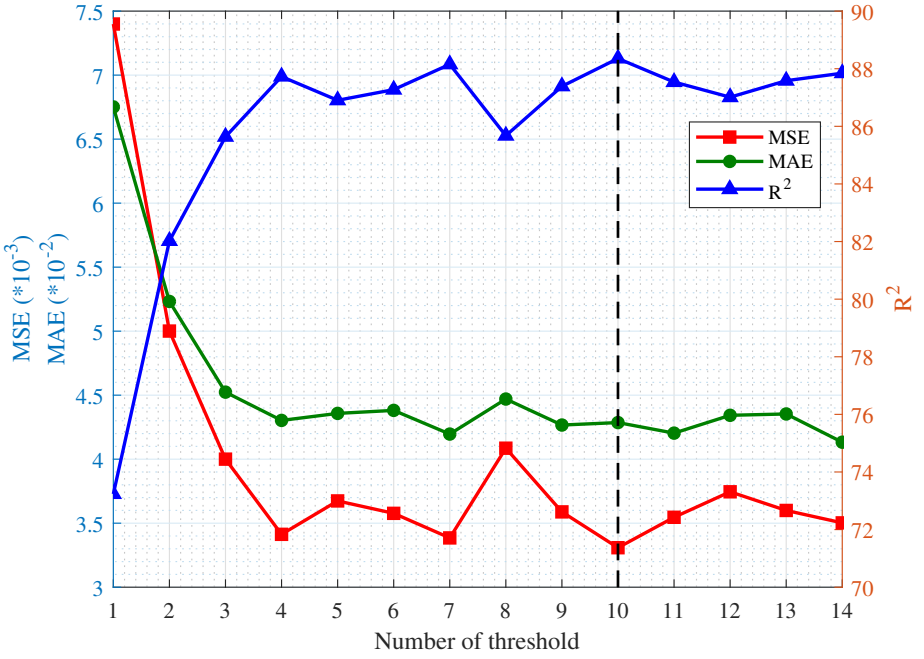


Figure 5.6 *Boston Housing* : the *LADR* measures of performance using the *EW* method

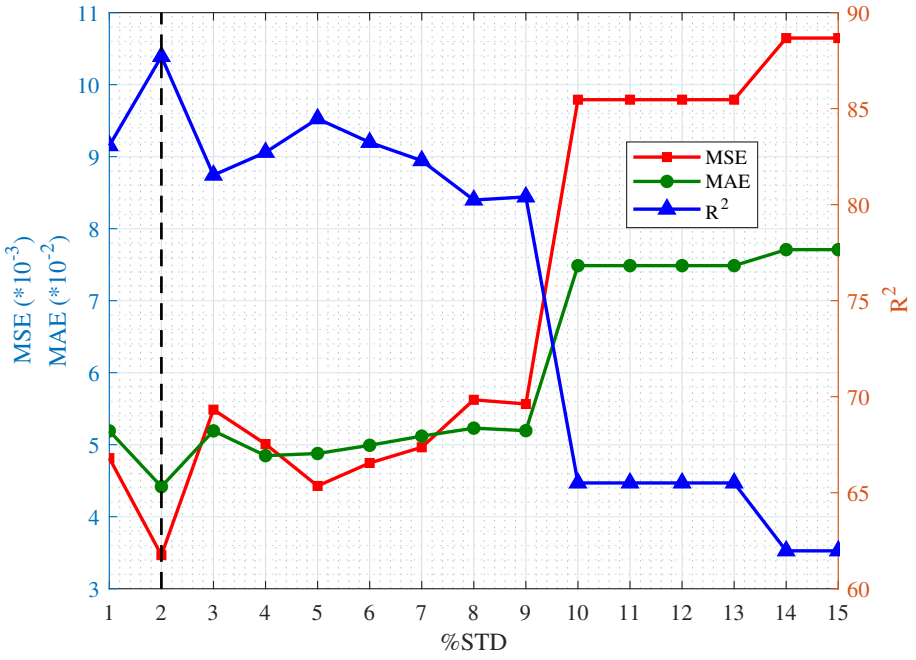


Figure 5.7 *Boston Housing* : the *LADR* measures of performance using the *%STD* method

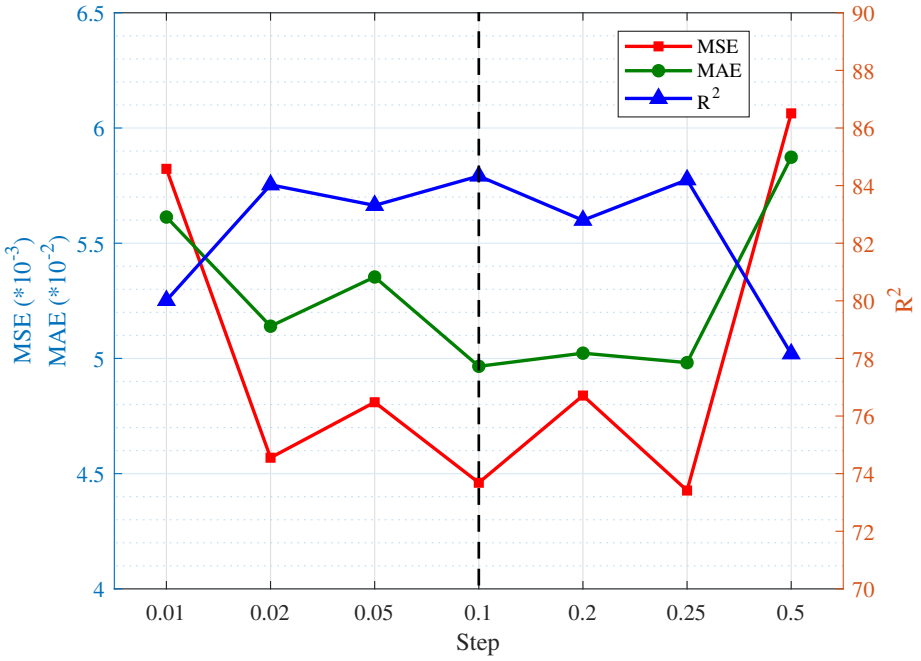


Figure 5.8 *Boston Housing* : the *LADR* measures of performance using the *QT* method

Table 5.6 The performance of the regression models for *Boston Housing*

Method	Threshold	MSE(*10^{-3})	R^2	MAE(*10^{-2})
LADR-*KM*	$\tau = 4$	3.08	88.98	4.15
LADR-*EW*	$\tau = 10$	3.31	88.35	4.29
LADR-*STD*	$\tau = 2\%$STD	3.47	87.72	4.42
LADR-*QT*	$\tau = 9$	4.46	84.33	4.97
LR	-	5.92	79.27	5.58
SVR	-	6.29	77.85	5.38
DTR	-	6.48	77.35	5.76
RF	-	3.72	86.86	4.37
PolyR	-	5.18	-	4.51

are Linear Regression (LR), Support Vector Regression (SVR), Decision Tree for regression (DTR), Random Forest (RF), and Polynomial Regression (PolyR). We apply these techniques using *R-studio software* [154]. Tables 5.6 to 5.11 show that *LADR* models that use the discretization methods *KM* and *EW* result in better values of the measures of performance compared with LR, SVR, DTR and PolyR. Their models score a higher value of R^2 and lower values for both MSE and MAE. Furthermore, both methods perform better than *QT* models. On the other hand, the %*STD* method is still competitive to the other methods. It outperforms LR, SVR and DTR in four out of six datasets as well as three datasets against *QT* models and there are slight differences in the results of the other datasets. The *LADR* technique obtains a predictive model that has a measure of performance values, which are competitive to the best model, RF. By comparing the results of these models, we conclude that the LADR-*KM* model at *KM*=5 has the best performance. In appendix B, the equation B.1 presents the detailed structure of this regression model. Similarly, the results of *Computer Hardware*, *Auto-mpg* , *Servo*, *Airfoil Self-Noise*, and *Concrete Strength* datasets are depicted in figures (5.9-5.12), (5.13-5.16), (5.17-5.20), (5.21-5.24), and (5.25-5.28) respectively. Tables 5.7 to 5.11 provide a comparison between the *LAD* models to obtain the best performance.In addition, the best model for each of these five datasets is depicted in equations (B.2-B.6).

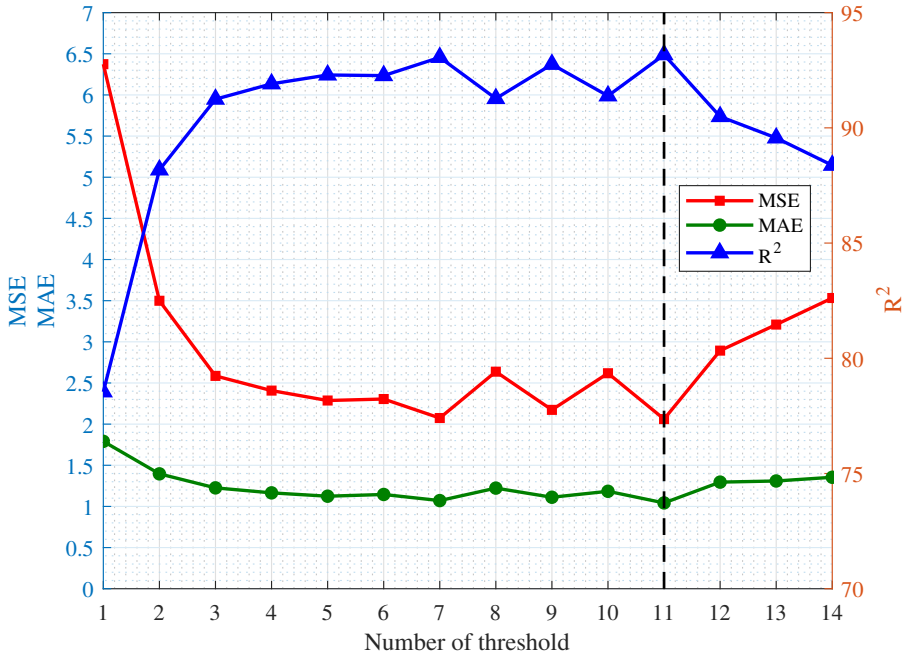


Figure 5.9 *Computer Hardware*: the *LADR* measures of performance using the *KM* method

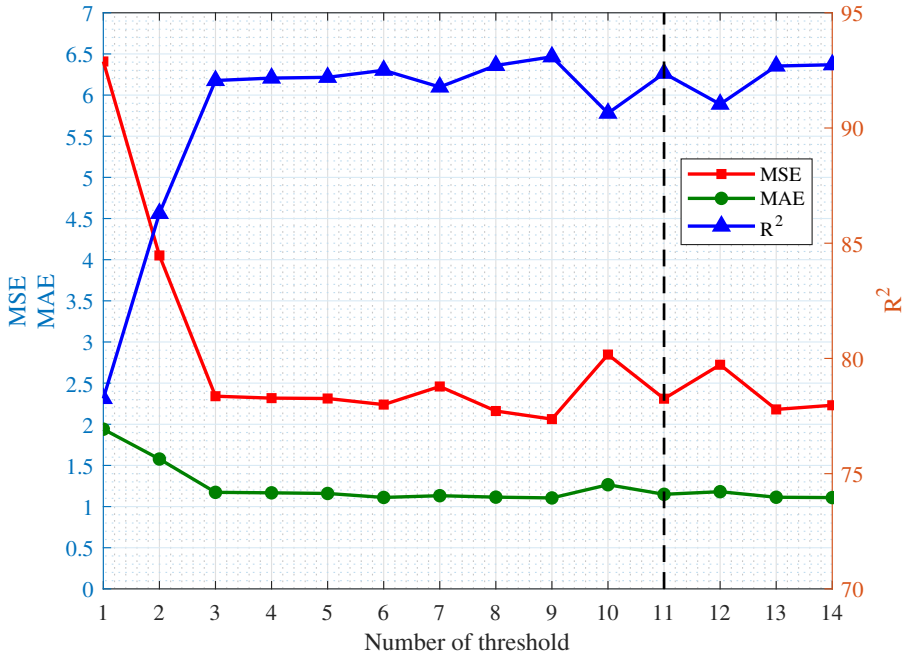


Figure 5.10 *Computer Hardware*: the *LADR* measures of performance using the *EW* method

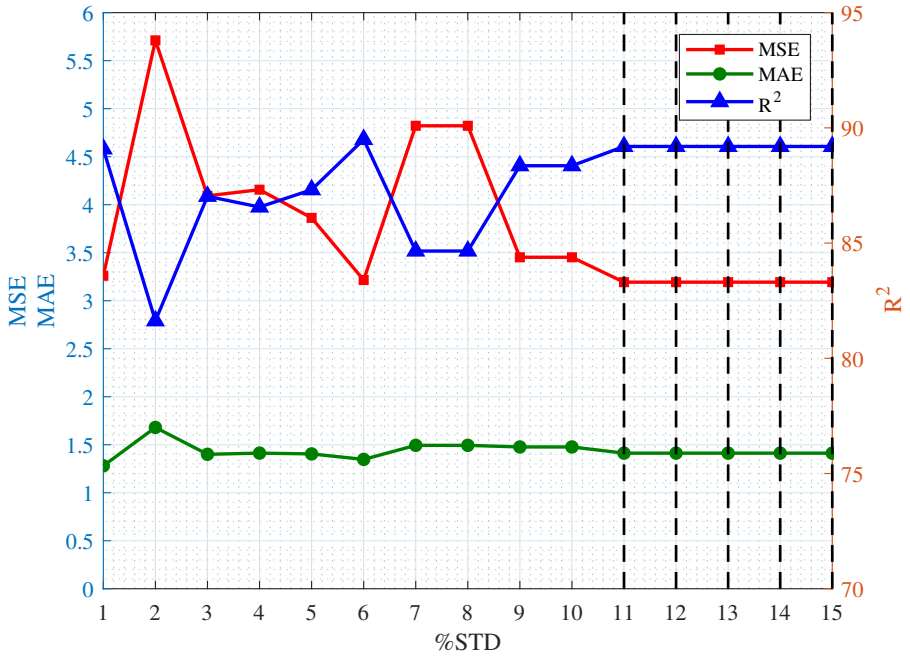


Figure 5.11 *Computer Hardware*: the *LADR* measures of performance using the *%STD* method

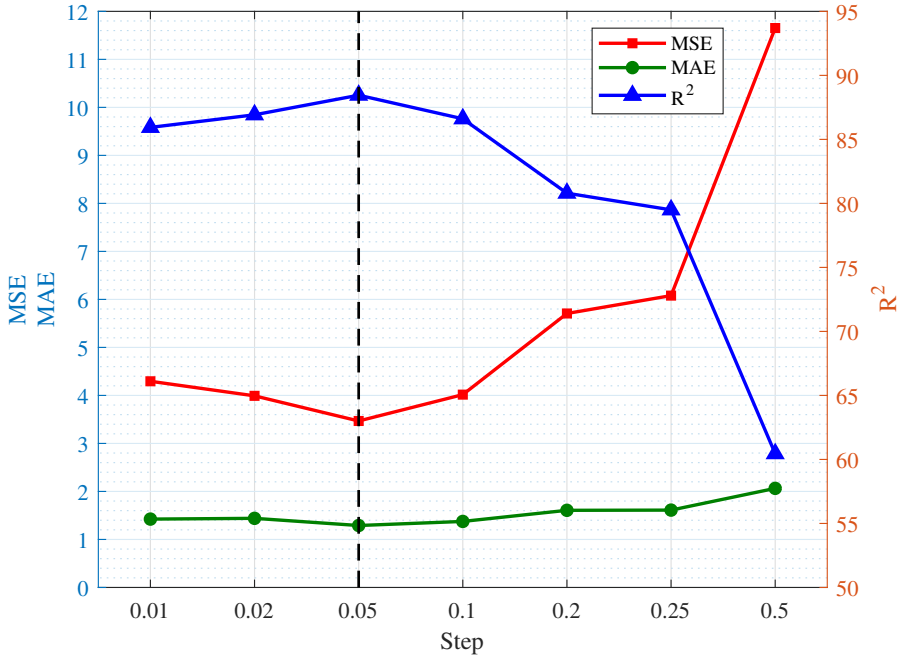


Figure 5.12 *Computer Hardware*: the *LADR* measures of performance using the *QT* method

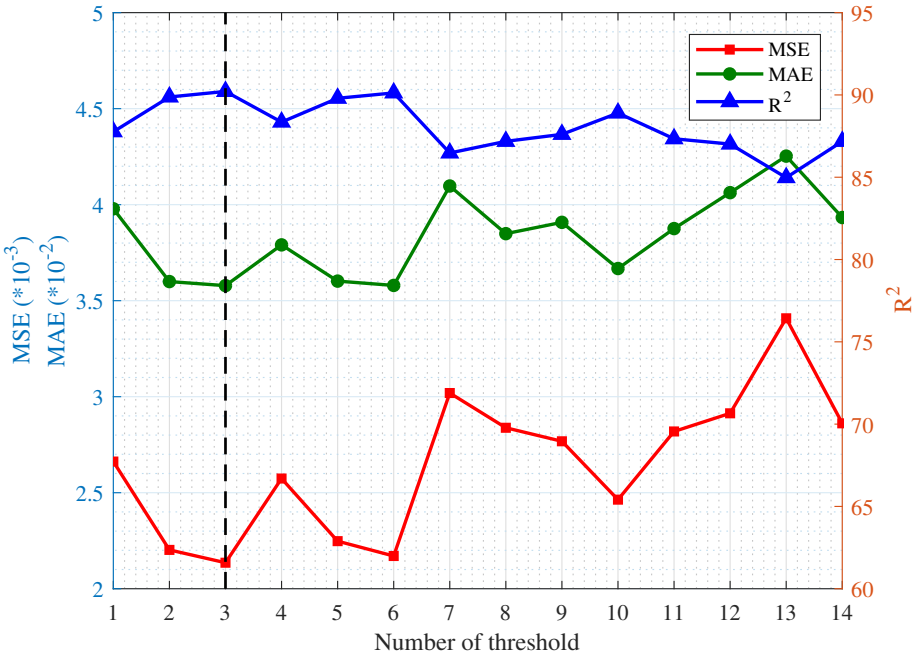


Figure 5.13 *Auto-mpg*: the *LADR* measures of performance using the *KM* method

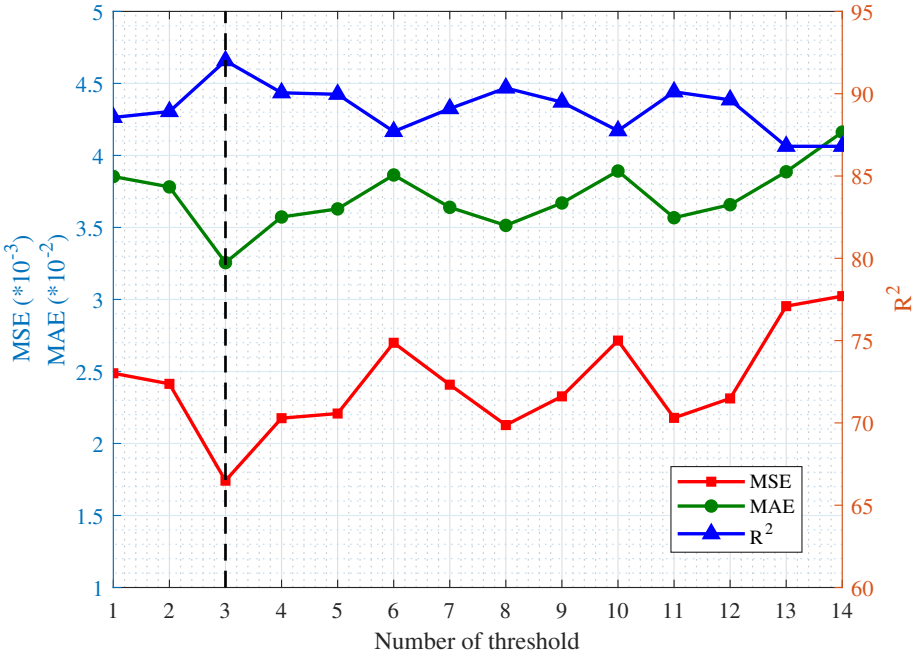


Figure 5.14 *Auto-mpg*: the *LADR* measures of performance using the *EW* method

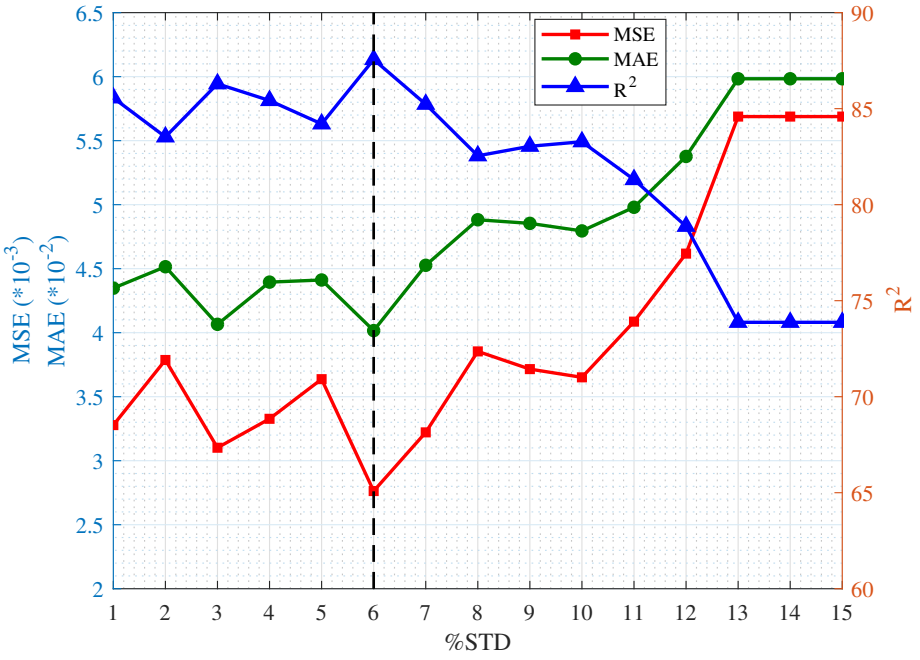


Figure 5.15 *Auto-mpg*: the *LADR* measures of performance using $\%STD$ method

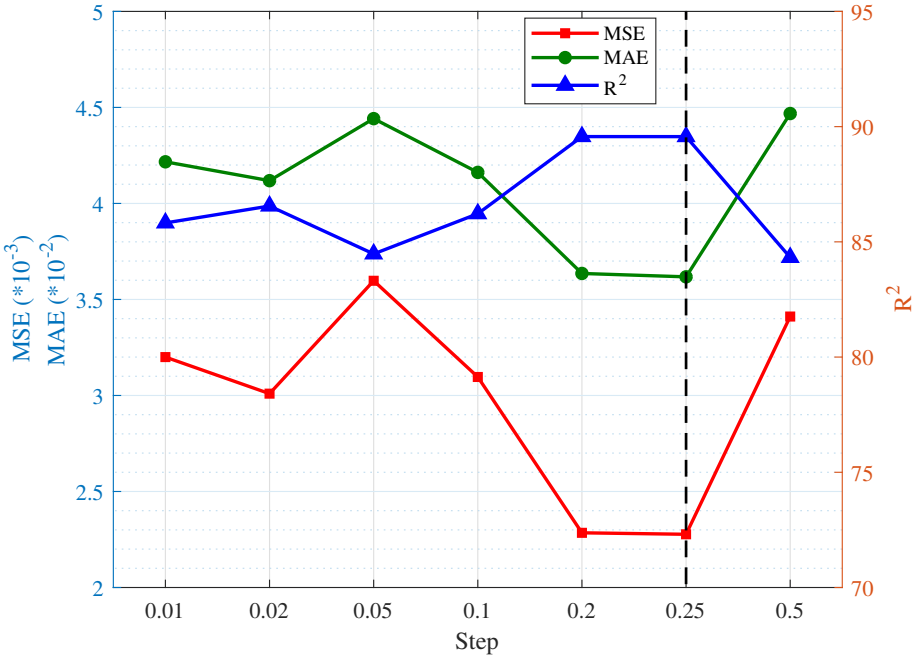


Figure 5.16 *Auto-mpg*: the *LADR* measures of performance using the QT method

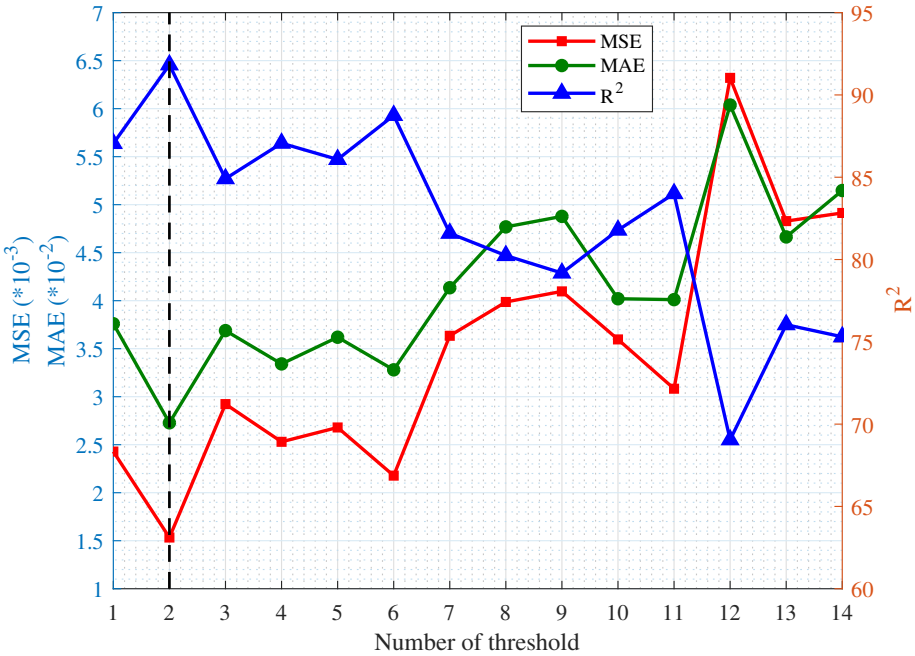


Figure 5.17 *Servo*: the *LADR* measures of performance using the *KM* method

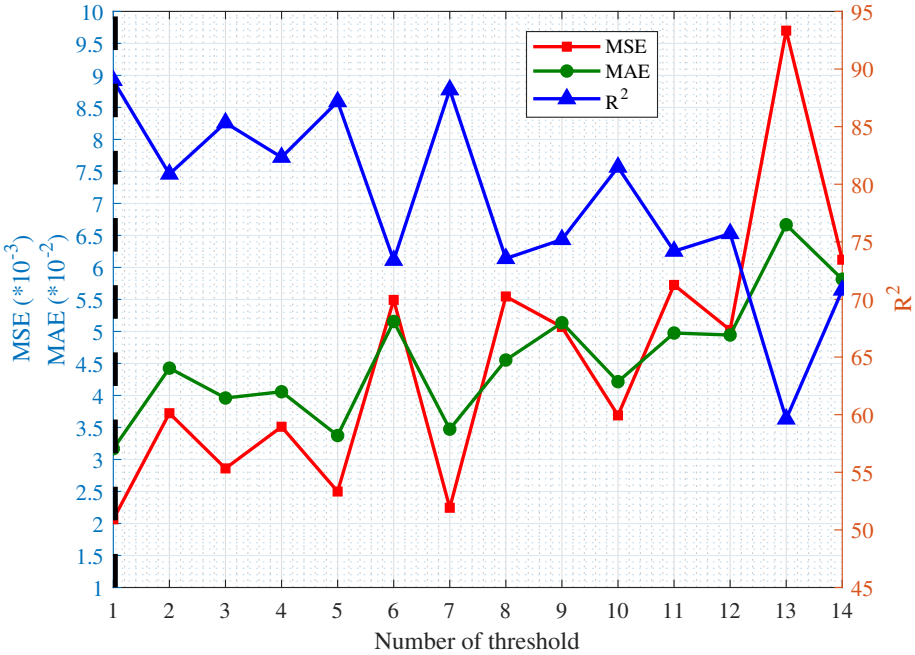


Figure 5.18 *Servo*: the *LADR* measures of performance using the *EW* method

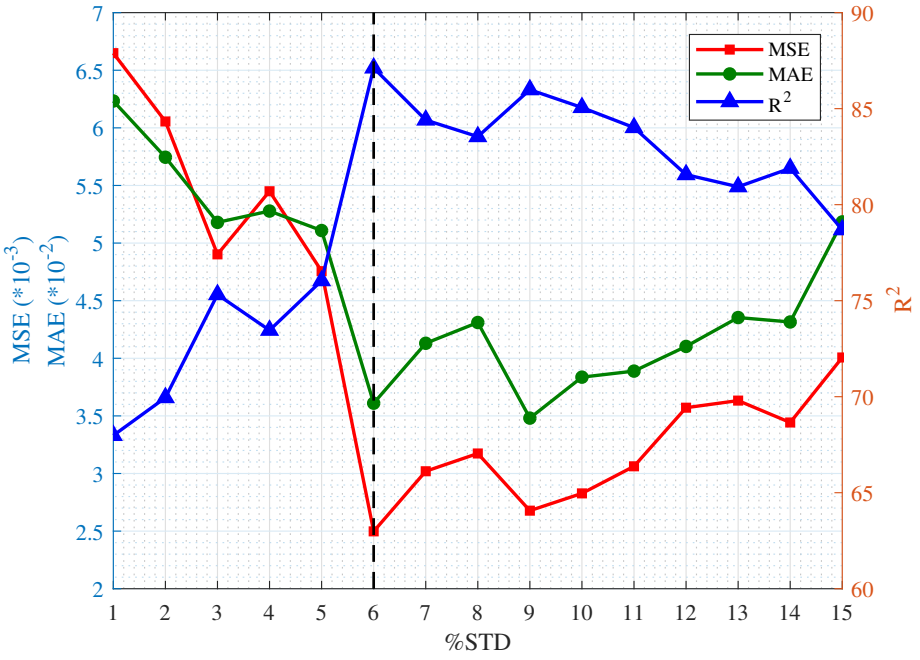


Figure 5.19 *Servo*: the *LADR* measures of performance using the %*STD* method

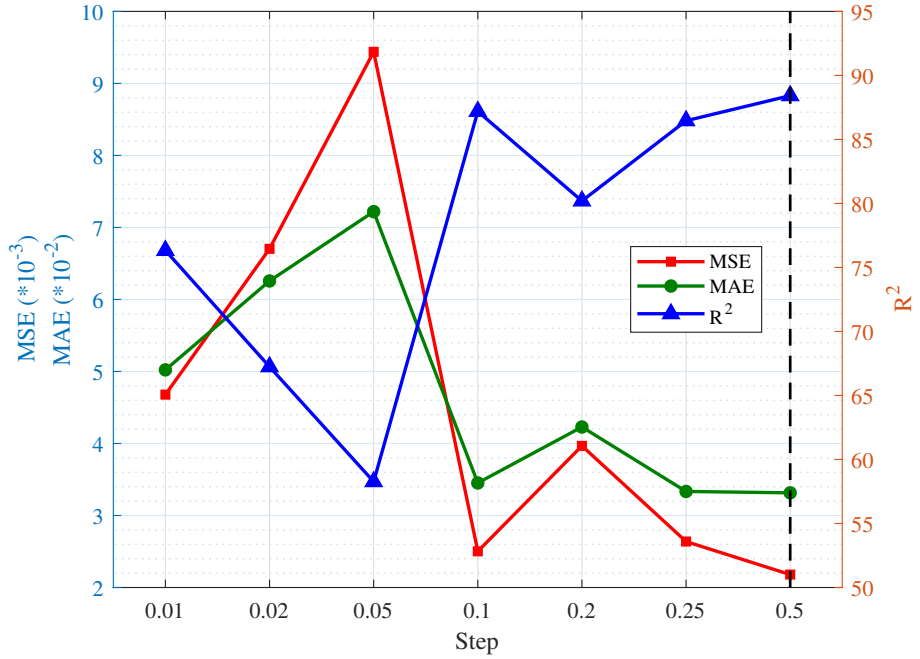


Figure 5.20 *Servo*: the *LADR* measures of performance using the *QT* method

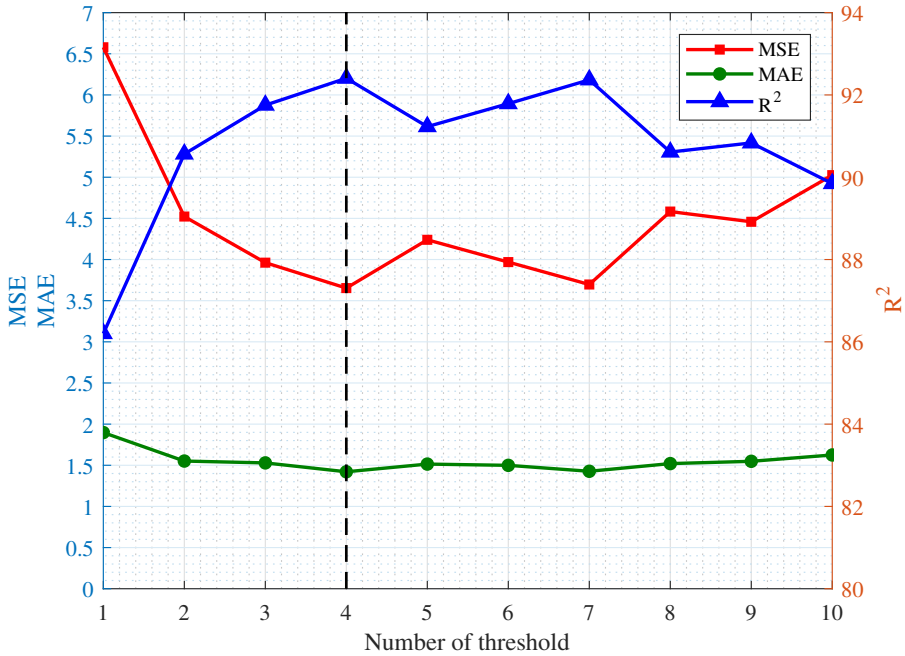


Figure 5.21 *Airfoil self-noise*: the *LADR* measures of performance using the *KM* method

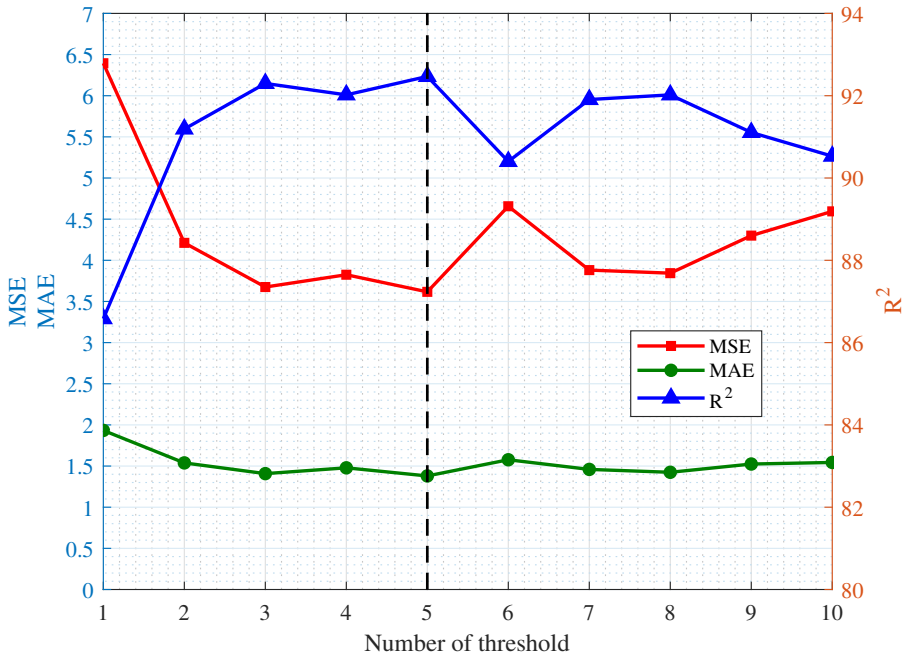


Figure 5.22 *Airfoil self-noise*: the *LADR* measures of performance using the *EW* method

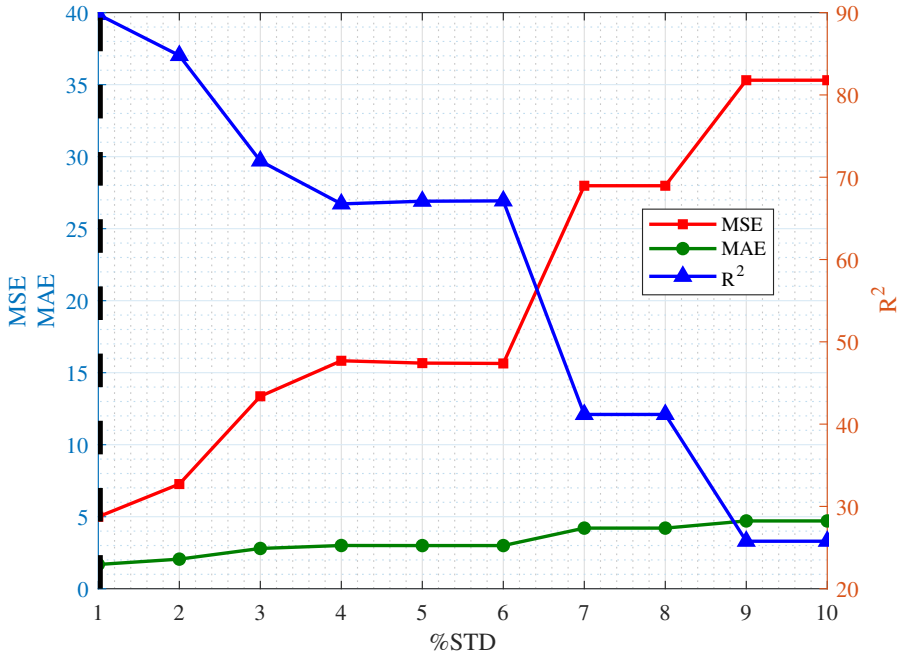


Figure 5.23 *Airfoil self-noise*: the *LADR* measures of performance using the $\%STD$ method

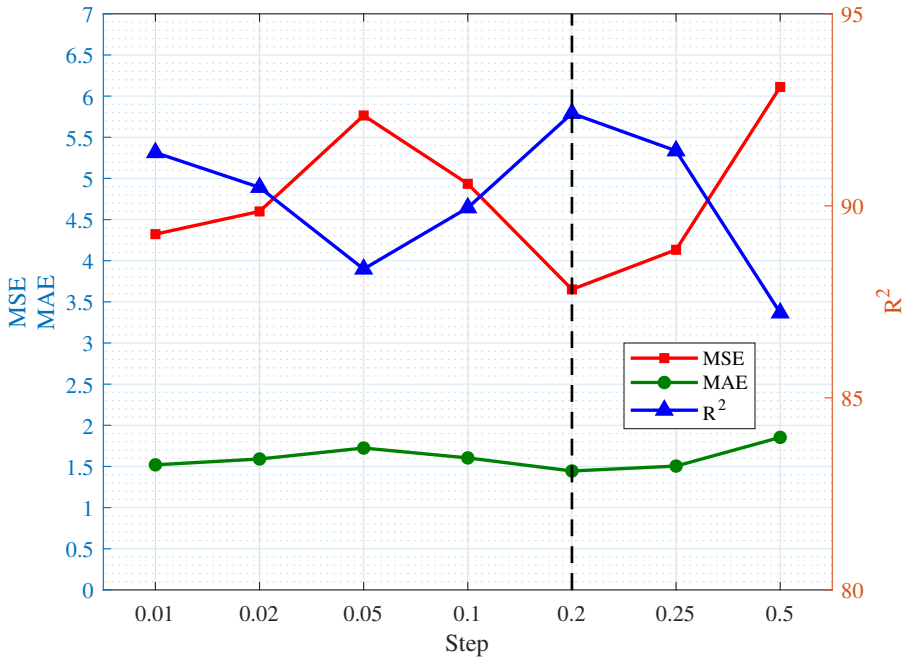


Figure 5.24 *Airfoil self-noise*: the *LADR* measures of performance using the QT method

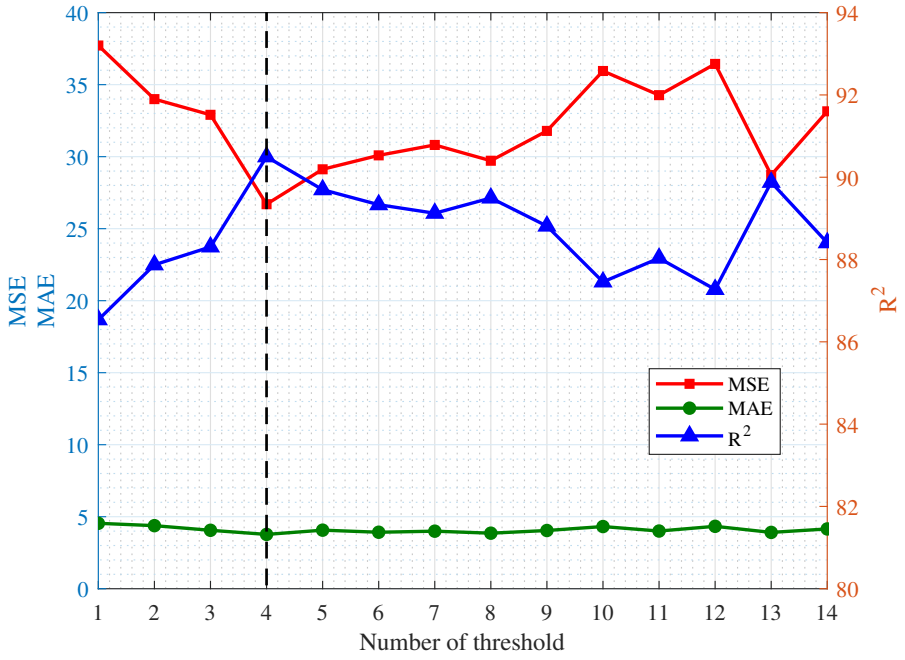


Figure 5.25 *Concrete strength*: the *LADR* measures of performance using the *KM* method

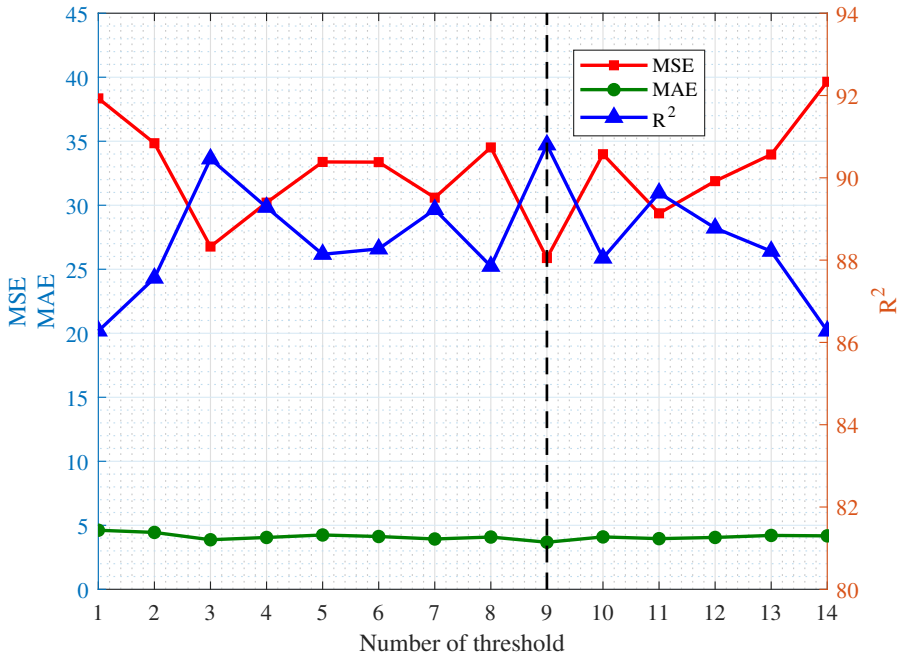


Figure 5.26 *Concrete strength*: the *LADR* measures of performance using the *EW* method

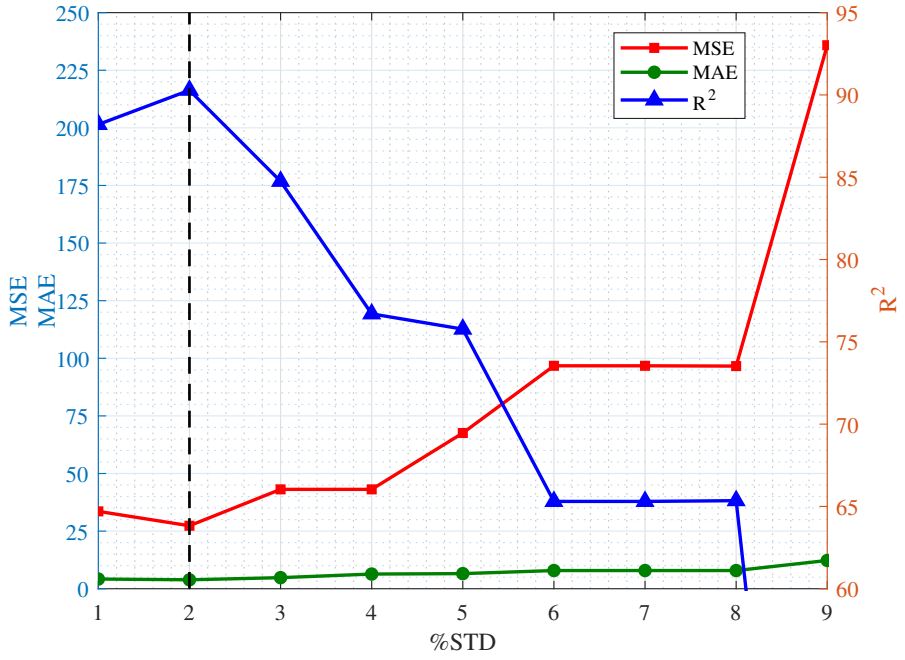


Figure 5.27 *Concrete strength*: the *LADR* measures of performance using the $\%STD$ method

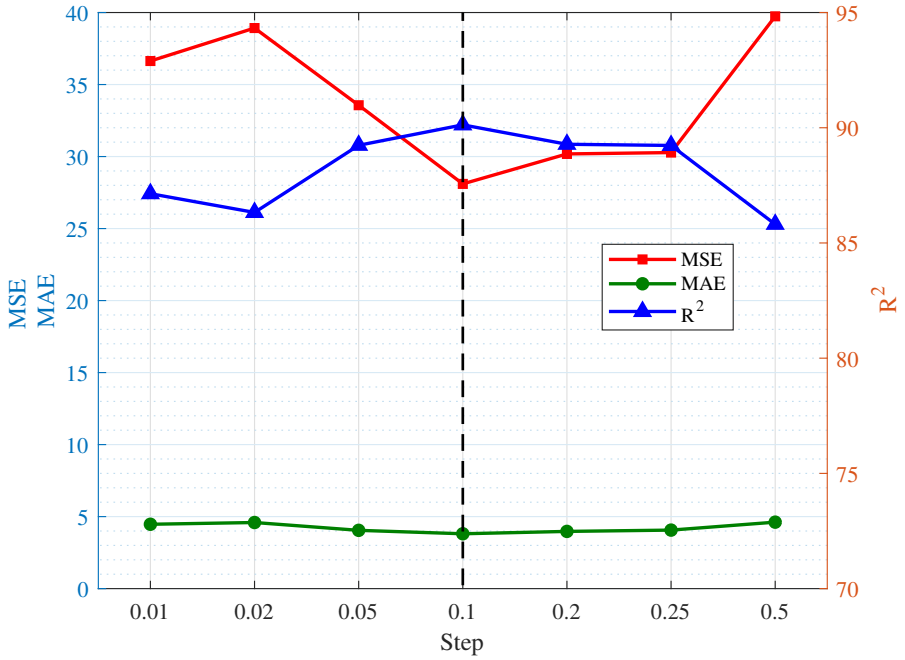


Figure 5.28 *Concrete strength*: the *LADR* measures of performance using the QT method

Table 5.7 The performance of the regression models for *Computer Hardware*

Method	Threshold	MSE	R^2	MAE
LADR-*KM*	$\tau = 11$	2.061	93.17	1.04
LADR-*EW*	$\tau = 9$	2.063	93.09	1.11
LADR-*STD*	τ = 11:15%STD	3.19	89.19	1.41
LADR-*QT*	$\tau = 19$	3.474	88.45	1.29
LR	-	2.63	89.36	1.28
SVR	-	2.70	89.64	1.28
DTR	-	5.46	81.95	1.61
RF	-	2.50	91.53	1.15
PolyR	-	2.62	-	1.14

Table 5.8 The performance of the regression models for *Auto-mpg*

Method	Threshold	MSE(*10^{-3})	R^2	MAE(*10^{-2})
LADR-*KM*	$\tau = 3$	2.14	90.21	3.58
LADR-*EW*	$\tau = 3$	1.74	92.00	3.26
LADR-*STD*	τ = 6%STD	2.76	87.56	4.02
LADR-*QT*	$\tau = 3$	2.28	89.57	3.62
LR	-	2.40	89.14	3.76
SVR	-	2.42	89.12	3.75
DTR	-	2.88	87.07	4.16
RF	-	2.01	90.90	3.38
PolyR	-	2.15	-	3.31

Table 5.9 The performance of the regression models for *Servo*

Method	Threshold	MSE(*10^{-3})	R^2	MAE(*10^{-2})
LADR-*KM*	$\tau = 2$	1.53	91.83	2.73
LADR-*EW*	$\tau = 1$	2.07	89.00	3.17
LADR-*STD*	τ = 6%STD	2.50	87.12	3.61
LADR-*QT*	$\tau = 1$	2.18	88.42	3.31
LR	-	7.26	63.81	6.94
SVR	-	7.48	64.50	7.11
DTR	-	3.10	83.53	4.13
RF	-	1.78	90.86	2.84
PolyR	-	3.29	-	3.84

5.5 Validation of the *LADR*

To test whether the differences between the values of the MSE, the MAE and the R^2 obtained with the best *LADR* models and those obtained by using the LR, SVR, DTR, *LADR-QT*,

Table 5.10 The performance of the regression models for *Airfoil Self-Noise*

Method	Threshold	MSE	R^2	MAE
LADR-*KM*	$\tau = 4$	3.65	92.41	1.42
LADR-*EW*	$\tau = 5$	3.62	92.47	1.38
LADR-*STD*	$\tau = 1\%$STD	5.00	89.78	1.68
LADR-*QT*	$\tau = 4$	3.66	92.40	1.45
LR	-	23.17	51.43	3.74
SVR	-	23.89	51.28	3.68
DTR	-	11.08	76.64	2.55
RF	-	3.21	93.57	1.30
PolyR	-	21.80	-	3.54

Table 5.11 The performance of the regression models for *Concrete Strength*

Method	Threshold	MSE	R^2	MAE
LADR-*KM*	$\tau = 4$	26.70	90.50	3.77
LADR-*EW*	$\tau = 9$	25.88	90.80	3.68
LADR-*STD*	$\tau = 2\%$STD	27.33	90.28	3.87
LADR-*QT*	$\tau = 9$	28.10	90.13	3.81
LR	-	109.37	61.19	8.30
SVR	-	117.21	60.14	8.20
DTR	-	56.57	80.38	5.54
RF	-	22.35	92.19	3.31
PolyR	-	38.20	-	4.61

RF, and PolyR are statistically significant, the Friedman–Nemenyi is used [155, 156]. The test is formulated as follows in equation (5.12):

$$Fr = \frac{12}{bM(M+1)} \sum_j Ra_j^2 - 3b(M+1) \tag{5.12}$$

Where b is the number of datasets, M is the number of competing models and Ra_j^2 is the square of the sum of the rank for the jth model. The null hypothesis (H_0) states that all of the models used have the same mean of MSE, while the alternative hypothesis (H_a) states that all of the models have a different mean of MSE. The $\chi^2_{(M\text{-}1,a)}$ is calculated with a degree of freedom M-1, at a significance level α=0.05. If $Fr > \chi^2_{(M\text{-}1,a)}$, the H_0 is rejected and the test is considered significant, and accordingly, some of the regression models lead to significantly different values of MSE. If the H_0 is rejected, we perform a post-hoc test by applying multiple pairwise comparisons between the models to identify the models that lead to significantly different values of MSE. We select the model that has the lowest mean rank. It is compared

with the mean ranks of the others using equation (5.13).

$$|Ra_j - Ra_j^*| \begin{cases} > D & \text{(Significant difference)} \\ < D & \text{(Insignificant difference)} \end{cases}$$

$$D = 2.2414\sqrt{\frac{bM(M+1)}{6}}$$

$$\forall j \& h = 1, \ldots, |M| \& j \neq j^* \tag{5.13}$$

Where j^* is the jth model that has the highest mean rank, D is the critical difference that is used to determine the significance of the differences. In this paper, D is estimated at a significance level of α_n =0.0125 of the standard normal distribution. If the difference between the rank sum of jth model and j^* model is greater than D, the performance of jth model is significantly different, otherwise it is not.

We apply the Friedman test to compare the values of MSE for the *LADR* using the three proposed discretization methods; *KM*, *EW*, and *STD* ; with the *LADR-QT* [157]. The performance of both *LADR-KM* and *LADR-EW* have significant difference from the *LADR-STD* and *LADR-QT* as in table 5.12. This emphasizes the same conclusion that both *LADR-STD* and *LADR-QT* models approximately have the same performance. On the other hand, We select the best *LADR* model for each dataset to represent the *LADR* technique. Table 5.13 shows that the SVR has the highest mean rank of MSE, so we compare its performance with the other techniques. It demonstrates that the performance of *LADR* and RF differ significantly from the performance of the other techniques. Moreover, the LR, SVR , DTR and PolyR techniques are not significantly different and have the same mean of MSE. The results show that the performance of *LADR* has better results and comparable to the most accurate regression models, RF.

Furthermore, in order to demonstrate the differences in the discretization methods, we hold a comparison with the combinatorial regression (CR) [138] for *Boston Housing* and *Computer Hardware* datasets. The differences are noticed by comparing the CR's results in table 5.14 with the previous tables. Thus, we apply the *LADR* for the raw data directly to be the worst case. Generally, the *LADR* outperforms the CR by a large margin, as is evident from the obtained results in table 5.14. For Boston Housing, the *LADR* model increases in R^2 by minimum 12.6% and decreases in MAE by minimum 20.5% as well as 41.7% and 51.1%, respectively, for Computer Hardware. We notice that using the three-discretization methods provide superior results in terms of MSE, R^2 and MAE that are to other regression methods.

The proposed technique shows better results with small values of MSE, and Friedman's

Table 5.12 Friedman test for the best *LADR* model using the four discretization methods for all datasets

Model	Ra_j	$\bar{Ra}_j$	$\lvert Ra_j - Ra_{LADR-QT/STD}\rvert$	Is there significant difference?
LADR-KM	9	1.5	12	Yes
LADR-EW	9	1.5	12	Yes
LADR-STD	21	3.5		
LADR-QT	21	3.5		
Fr	14.4			
$\chi^2_{(3,0.05)}$	7.81			
D	10.024			

Table 5.13 Friedman test for the best *LADR* model and other regression models for all datasets

Model	Ra_j	$\bar{Ra}_j$	$\lvert Ra_j - Ra_{SVR}\rvert$	Is there significant difference?
LADR	8	1.33	25	Yes
RF	10	1.67	23	Yes
PolyR	19	3.17	14	No
LR	27	4.50	6	No
DTR	29	4.83	4	No
SVR	33	5.50		
Fr	25.62			
$\chi^2_{(5,0.05)}$	11.07			
D	14.53			

test is carried out confirming that the accuracy is improved significantly. This affects the performance of the model in predicting accurately.

5.6 Numerical application

Process monitoring is essential for fault detection in industrial applications. Control charts are considered to maintain both process and product qualities [8, 9]. They monitor the process variations and determine whether they are in control or out of control, so corrective actions can be taken to reduce the variability and to improve the quality. In most industrial applications, the process has more than one variable. The streams of data may have high dependency and correlated variables, which affect the performance of the charts [67]. Moreover, when the number of variables in the process increases, there is a time delay in

Table 5.14 A comparison between the performance of *LADR* and CR

Method	*Boston Housing*		
	Threshold	R^2	MAE
LADR-*KM*	$\tau = 6$	89.74	2.18
LADR-*EW*	$\tau = 4$	89.32	2.26
LADR-*STD*	$\tau = 3\%$STD	87.20	2.36
CR	-	77.44	2.97

Method	*Computer Hardware*		
	Threshold	R^2	MAE
LADR-*KM*	$\tau = 9$	95.29	22.86
LADR-*EW*	$\tau = 12$	95.52	23.00
LADR-*STD*	$\tau = 15\%$STD	95.49	24.85
CR	-	67.24	45.04

detecting the mean shift in the process [68]. Subsequently, a large amount of missed detection occurs, in addition to false alarms. A regression adjustment was developed in [158] to regress a variable on the rest of the other variables. Then, a residual control chart monitors the variations of the regression model's residuals. As such, the regression model solved the problem of dimensionality, since it reduces the number variables to be monitored by a control chart. Furthermore, it monitors the residuals that do not suffer from the autocorrelation phenomenon. An important factor is to ensure the high accuracy of the regression model to improve the performance of the charts in fault detection. Therefore, the *LADR* technique is adopted as a regression adjustment since it has proven its efficiency in building a regression model. *LADR* obtains a relationship between the independent patterns with the dependent variable(s) of the process. For example, *LADR* is applied to the cascade process data given in [1]. The dataset contains 40 observations, where each observation has nine independent variables and two dependent variables. *LADR* is used to obtain a regression model for the first dependent variable to compare its performance with the least squares regression model obtained in [1]. In table 5.15, we present the results of both the *LADR* and LR models using 10-folds cross validation for 10 times. Again, MSE, MAE and R^2 are calculated to evaluate the performance of each technique. Although LADR-*STD* does not perform as well as the others in this dataset, the results show that LADR-*KM* and LADR-*EW* outperform the LR. The LADR-*KM* has the best performance, which is shown by the lowest value of both MSE and MAE, in addition to being higher than LR in terms of R^2. It can be concluded that both *EW* and *KM* are the best discretization methods for this example. The models' structures are depicted in equation (14).

Table 5.15 A comparison between the performance of *LADR* and LR

Method	Threshold	MSE	R^2	MAE
LADR-*KM*	$\tau = 2$	0.72	0.76	0.68
LADR-*EW*	$\tau = 3$	0.75	0.75	0.71
LADR-*STD*	$\tau = 12$: 15%STD	1.21	0.66	0.90
LR	-	0.96	0.67	0.77

Our future research will implement the residuals of the best model (LADR-*KM*) on the residual control chart. Therefore, the model can check the process stability by the detection of fault, if any. When a fault is detected, the developed model indicates the reason of the process abnormality through the used patterns. Accordingly, we will be able to take the corrective action(s) to retain the process in-control.

$$\begin{aligned} Y_{[EW]} =& 949.8136 + 1.7668X_{P_2} + 0.4513X_{P_3} + 0.6435X_{P_4} - 0.8506 \\ & X_{P_5} - 1.3392X_{P_7} + 0.8603X_{P_9} + 1.487X_{P_{11}} + 1.7805X_{P_{17}} + \\ & 1.2748X_{P_{20}} \end{aligned} \tag{5.14.1}$$

$$\begin{aligned} Y_{[KM]} =& 952.3113 - 1.5841X_{P_1} + 0.8017X_{P_4} - 0.8226X_{P_7} + 0.8413X_{P_{12}} \\ & + 1.2594X_{P_{15}} \end{aligned} \tag{5.14.2}$$

$$\begin{aligned} Y_{[\%STD]} =& 951.787 + 1.2729X_{P_3} - 1.0707X_{P_5} - 0.5346X_{P_6} + 0.5323X_{P_7} \\ & - 1.1123X_{P_{11}} - 1.5754X_{P_{13}} + 0.7708X_{P_{21}} + 0.3641X_{P_{22}} + \\ & 0.53X_{P_{25}} + 0.9851X_{P_{29}} - 0.3541X_{P_{35}} + 0.5777X_{P_{37}} - 0.491 \\ & X_{P_{38}} + 1.1585X_{P_{41}} + 0.7438X_{P_{49}} - 0.7382X_{P_{50}} + 1.26X_{P_{51}} - \\ & 0.7803X_{P_{52}} \end{aligned} \tag{5.14.3}$$

$$\begin{aligned} Y_{[LR]} =& 825.8853 + 0.4741X_1 + 1.4134X_2 - 0.1168X_3 - 0.0824X_4 - \\ & 2.3918X_5 - 1.2978X_6 + 2.1764X_7 + 2.9805X_8 + 113.217X_9 \end{aligned} \tag{5.14.4}$$

5.7 Conclusion

In this paper, we have constructed regression models to predict process performance based on patterns that are extracted by the data mining of sensor readings. We extended the existing

LAD data mining classification technique by developing a *LAD* regression model (*LADR*) to predict the response variable based on patterns found in the dataset. These patterns preserve the significant information and knowledge found in the data, and aggregate them into fewer patterns, instead of the entire set of observations. Binary variables that indicate the presence or the absence of these patterns in new unseen observations are used to predict the dependent continuous numerical response variable by using the regression model. This paper presents three main methods, the *EW*, *KM* and %*STD*, to create the independent variables. The best method provides the best performance metrics (MSE, R^2, and MAE). To ensure the accuracy of the regression model, we apply 10-fold cross validation 10 times to obtain the average of each performance metric. It is shown that *LADR* models, which use the *EW*, *KM* and %*STD*, have better results compared with the other well-known machine learning regression techniques. Moreover, *LADR* demonstrates significant improvement in the performance of the regression models.

For further research, the *LADR* model will be used in multivariate control charts to monitor and detect anomalies. Our hypothesis is that by using the *LADR* we will be able to decrease the false positive and false negative errors, which are inherent to any control chart. Moreover, it is expected that the extracted patterns will provide indications of the root causes for any out of control observations. Quality 4.0 will therefore be implemented, and manufacturers will be able to maintain and improve their quality management by using machine-learning techniques that use online and real-time sensor data to monitor the process performance.

CHAPTER 6 ARTICLE 3: ROOT CAUSE ANALYSIS OF AN OUT-OF-CONTROL PROCESS USING A LOGICAL ANALYSIS OF DATA REGRESSION MODEL AND EXPONENTIAL WEIGHTED MOVING AVERAGE

Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto

Submitted to:

Journal of Intelligent Manufacturing, 2022

6.1 Abstract

Control charts are widely used as a tool in process quality monitoring to detect anomalies and to improve the quality of a process and product. Nevertheless, their limitations have increased in the face of increasingly complex manufacturing processes. They do not have capability of handling large streams of non-normal and autocorrelated multivariate data, which is in most real applications. This may lead to an increase in false alarm signals and/or missed detection of anomalies. They are not designed to automatically identify the root causes of an anomaly when the process is out-of-control. Several machine-learning techniques were integrated with control charts to improve the sensitivity and specificity of anomaly detection. Nevertheless, some existing techniques still produce a high false alarm rate and/or missed detection. The root cause analysis is seldom performed. In this paper, we propose a new integration that combines the logical analysis of data regression technique (*LADR*) and the exponential weighted moving average (*EWMA*) as a new model-based control chart. *LADR* is based on the traditional *LAD* methodology, which is a supervised data mining technique for pattern generation. *LADR* transforms the original independent variables into pattern variables by using *cbmLAD* software to develop a regression model. The *LADR-EWMA* increases the sensitivity of anomaly detection in the process and uses the patterns to perform root cause analysis of that anomaly. We applied *LADR-EWMA* to a real application: a concrete manufacturing process. We compared its performance with Linear regression, Support vector regression, Partial Least Square regression, and Multivariate adaptive regression Spline. The results demonstrate that the *LADR-EWMA*, which is based on pattern recognition, performs better compared to the other techniques in terms of a reduction of false alarms and missed detection. In addition, *LADR-EWMA* facilitates interpretation and identification of the root cause of the detected anomaly.

Keyword: Process monitoring, Logical Analysis of Data Regression (*LADR*), anomaly detection, root cause analysis, LADR based EWMA control chart (*LADR-EWMA*), Quality 4.0

6.2 Introduction

Process quality monitoring and control are two important challenges that many manufacturers are interested in. Process monitoring plays a significant role in guaranteeing better performance of a process by detecting anomalies, abnormal variations, or degradation in performance, in addition to preparing for root cause analysis and corrective actions to be taken to return a process to normal condition. As such, process monitoring ensures that both the

quality of the process and the product are within the pre-specified statistical control limits.

The main concept of process monitoring techniques is to exploit the available data and develop an inferential model for the monitored process. It is of crucial concern not only to improve the sensitivity for anomaly detection, but also the identification and interpretation of the root cause of that anomaly.

This paper introduces a new model-based control chart, called Logical Analysis of Data Regression (*LADR*) based control chart to detect the anomalies and to perform root cause analysis for corrective actions. The approach combines *LADR*, which is a machine learning technique for pattern generation and regression based on combining the logical analysis of data (*LAD*) [159] with a control chart. The generated patterns identify the multidimensional zones that characterize different groups of observations in the original data. The patterns identify the root causes of the subgroups of anomalous observations. *LADR* uses these patterns as independent variables instead of the original independent variables to obtain a regression model describing the dependent variable(s). *LADR* is integrated with the control chart to monitor process quality. Once, an anomaly is detected, the *LADR* model uses the generated patterns to identify the possible causes, automatically and without human interference. Therefore, the main contribution is to use machine learning to obtain interpretable patterns, which describe the root causes of any anomaly. Once the process's anomaly has been detected, the developed technique analyzes the root causes via pattern recognition. The variables that form the patterns are those that contribute to the presence of the anomaly.

The remainder of the paper is organized as follows: The "Literature review" section provides a literature review of the model-based control charts and the previous machine learning techniques that are used for anomaly identification. The "Methodology" section elaborates the proposed *LADR* regression-based control chart and its role in identifying the root-cause of the anomalous observation. The "Numerical example" section evaluates the performance of the proposed approach using a numerical example indicating the accuracy of *LADR*, and compares the results with other approaches. Finally, the "Conclusion" section presents a summary of the contributions of this paper, areas for further research, and some concluding remarks.

6.3 Literature review

Statistical process control (SPC) is a data-driven process monitoring approach [160] that monitors a process to reduce its variability and improve its quality [6]. The process is considered statistically in-control when the variations are due to unavoidable or natural

causes which are called common causes. Conversely, assignable causes result in abnormal variability during manufacturing due to machine faults, non-conforming raw material, lack of calibration, or operator's errors. The process is statistically out-of-control [161]. Once SPC detects abnormal variability associated with the process, it provides a notification of the presence of abnormal variations. Hence, an investigation is launched and corrective action is taken to avoid loss of quality. In many industrial applications, control charts are one of the SPC tools that are used to monitor the variability of quality characteristics over time [8, 9]. They graphically represent the variations of the quality characteristics. Control charts determine the process condition, whether it is in-control or out-of-control. They are classified as univariate and multivariate control charts, depending on the number of observed variables [1].

The limitations of conventional control charts increase in the face of increasing complexity of manufacturing processes. Modern technological progress in various industrial applications has led to a continuous increase in the number of monitored variables, which makes monitoring the process's performance more cumbersome [16]. This is in addition to the presence of autocorrelated variables [162]. This leads to an increase in the time delay for detecting the mean shift in a process [69]. False alarm signals and missed detections are crucial problems that are inherent within the control charts [1]. Industrial processes may experience different types of anomalies during the operation of a system. Identification and interpretation of the root causes of these anomalies play an important role in taking the corrective actions to bring a process to an in-control state. The conventional control charts are not designed to automatically identify the root causes of an out-of-control signal [70]. For this reason, additional graphical techniques have been used, such as line graphs [163], boxplot charts [164], polyplots [165], and multivariate profile charts [166]. A statistical procedure was proposed in [167], based on a discriminant analysis that classified the observations as in-control or out-of-control in different groups [168]. Mason, Young, and Tracy (MYT) decomposition [169] is one of the more common statistical approaches that has been used for root cause detection. This approach is used when the number of the process variables are small, but it can not be used when the number of variables is large, because it suffers from extensive computational time problems. This is because it needs to carry out n! decompositions, where n is the number of the process variables. An adaptive step-down approach (ASD) [170] was suggested to overcome the drawback of MYT decomposition. The performance of root cause detection still has uncertainties when a large number of variables are responsible for the out-of-control state.

Recently, Quality 4.0 has been introduced as part of the new paradigm of Industry 4.0 [171]. It refers to the future of quality engineering within the context of Industry 4.0, the rapid

growth of technological advancement in internet applications, and the unprecedented rate of change that those advancements are bringing [172]. Quality 4.0 represents the impact of the digitalization of quality management functions on quality processes, tools, and people [10].

The Quality 4.0 framework consists of 11 axes; data analytic is one of them. It considers the improvement of performance in traditional quality tools by adding new capabilities [173]. Machine learning techniques are integrated with traditional quality tools to develop a model-based control chart. They are used to detect any abnormal variations in the process. This integration helps manufacturers that are striving to implement their own data analytic strategy to improve product quality and process performance. It also enables the automation of online monitoring to diagnose anomalies, as well as to predict quality characteristics. The implementation of machine learning not only monitors real-time processes, but it is also used to predict the process quality before the occurrence of an anomaly. The automatic identification and interpretation of abnormal variations allows the anomaly to be eliminated, and consequently the scrap and rework to be reduced, thereby improving the quality of the process.

The model-based control chart was originated by Mandel. It is a combination of a conventional control chart and linear regression [174]. Hawkins [158] developed a regression adjustment to monitor the residuals of the process variable of interest that is regressed on the others. The residuals represent the difference between the actual values of the dependent variable, which represents the monitored quality characteristic, and the predicted value of that variable. Support vector regression (SVR) was employed to obtain an SVR-chart [175]. A comparison was carried out when integrating Artificial Neural Networks (ANN), Support vector regression (SVR), and Multivariate adaptive regression splines ($MARS$) with control charts to monitor the mean process of the quality characteristic [162, 176]. A new statistical method combined with SVR and exponential weighted moving average ($EWMA$) was used to detect different types of anomalies at low levels of severity of the centrifugal chillers system [72]. A Partial Least Square (PLS) based on T^2 statistic and Q-statistic were not suitable in the detection of small variations because both statistics were memoryless, which means that no previous information was taken into consideration. The number of false alarms and missed detections increased as a result of their fixed control limits. Thus, PLS-based $EWMA$ was developed [80]. The PLS is used in process monitoring because it successively handles the multi-collinearity between the process variables. The $EWMA$ is a univariate control chart, which is very sensitive to small shifts.

Once an out-of-control observation is detected in the process, it is important to identify the root causes of anomalies. Several multivariate techniques are used to identify the behaviors

of the variables that contribute to an out-of-control process. Reconstruction-based methods were developed to obtain the anomaly detection index along the directions of the variables "Anomaly directions" based on an anomalous database [177]. The variable that contributes to the anomaly has the largest index in the reconstruction. They identify the variables that contribute to the detection of an anomaly. Practically speaking, the anomaly directions were unknown, and most often, the available data was not sufficient, which creates problems involving combinatorial optimization. Least absolute shrinkage and variable selection (*LASSO*) were used to overcome the shortage of the conventional reconstruction methods. It is used as a variable selection method using a penalized term. It tracks the propagation of the anomaly by shrinking the regression coefficients until these entire coefficient values are zero [17]. Elastic net (*EN*) based anomaly detection was developed to deal with strongly correlated variables [18]. Furthermore, a hybrid method was proposed that combined *ridge* and *LASSO* regression.

Different pattern recognition techniques have been integrated with a control chart. They are implemented to recognize patterns from the original data and diagnose the anomalous observation [178], for example, with Support Vector Machine (*SVM*) [8, 101], K-Nearest Neighborhood (*KNN*) [104], Decision Tree (*DT*) [99, 100], Random Forest (*RF*) [103] and Artificial Neural Network (*ANN*) [20, 106, 107]. Furthermore, hybrid techniques have been proposed when combining two or more classifiers such as *SVM-ANN* [108, 109] and Principal Component Analysis (*PCA*)-*SVM* [94]. Generally, it is necessary to randomly generate appropriate training data to determine the different types of anomalies and develop an efficient classifier.

The aim of integrating of machine learning with a control chart is either developing a regression model to overcome the drawbacks of the conventional control chart, or identifying the root cause of the detected out-of-control observation. This is the motivation behind the present paper. It is to develop a high-performance regression model based on extracted patterns from the original data. These patterns are exploited to conduct a root cause analysis of out-of-control observations, without generating any additional data for the training stage.

6.4 Methodology

We consider the historical data of a process of n, measured independent variables, $X_1, \ldots, X_n$, that indicate features that affect the process performance, and the output response Y represents the quality characteristics that measure this performance. *LAD* regression technique (*LADR*) extends the standard *LAD* classification methodology [132, 141] to obtain a data-driven regression model. It generates J patterns from the original dataset of independent

variables, such that all the observations are covered by at least one pattern. A regression model of J binary variables (0,1), X_{P_j}, j= 1, ..., J, is found, such that each observation is covered by pattern j (X_{P_j}=1), or not (X_{P_j}=0). The corresponding values of the quality characteristics are obtained by the regression model. The best regression model may be linear or non-linear. It has been shown in [179] that by using the generated patterns P_j and their corresponding binary variables X_{P_j} as the independent variables, instead of the original variables, X_1,...,X_n, the accuracy of the predicted values of the dependent variable increases. In the following sections, the patterns are used to perform a root cause analysis, and to find the possible causes of each out-of-control state.

Table 6.1 presents a simple dataset as an example to illustrate the methodology's steps. The dataset is generated based on a multivariate normal distribution of certain mean (μ) and covariance-variance matrix (Σ) using RStudio software [154]. It contains 30 observations, where each observation has two independent variables; X_1 and X_2; and one dependent variable; Y. The simulation is elaborated by considering:

$$\mu_{X_1} = \mu_{X_2} = \mu_Y = 0$$

$$\Sigma = \begin{bmatrix} 1 & 0.5 & 0.9 \\ 0.5 & 1 & 0.4 \\ 0.9 & 0.4 & 1 \end{bmatrix}$$

6.4.1 Overview of the *LADR* technique

For the sake of completeness, we provide a brief description of the *LADR* technique. The details are given in [179]. In sections 6.4.2 and 6.4.3, the *LADR* generated patterns are used to interpret the causes of the out-of-control state. *LADR* uses the *LAD* approach to generate patterns that differentiate and characterize different process states. It uses the *cbm-LAD* software [24] to obtain strong patterns, which have high coverage [159]. The following subsections show the steps to implement *LADR* technique as depicted in figure 6.1.

Classification of the response

The classification of a process's response Y is the cornerstone of *LADR*. It divides the data into N-classes to obtain the thresholds. Subsequently, *LAD* is implemented to generate the patterns that characterize each class. We consider a regression dataset $\Omega(n, Y, m)$ of m observations and n independent variables that indicate the process performance, and Y

Table 6.1 Illustrative example for the steps in the methodology

No.	X_1	X_2	Y	No.	X_1	X_2	Y
1	-0.63	1.33	0.53	16	-0.04	-0.71	-0.22
2	0.18	-0.09	0.12	17	-0.02	0.36	0.00
3	-0.84	0.35	-0.44	18	0.94	0.81	1.12
4	1.60	0.03	1.42	19	0.82	-0.07	0.72
5	0.33	-1.36	-0.38	20	0.59	0.91	0.70
6	-0.82	-0.46	-0.84	21	0.92	0.44	0.83
7	0.49	-0.37	-0.16	22	0.78	-0.57	0.45
8	0.74	-0.02	1.01	23	0.07	0.34	0.49
9	0.58	1.13	0.95	24	-1.99	-1.23	-2.58
10	-0.31	0.75	0.54	25	0.62	1.46	1.22
11	1.51	-0.09	1.42	26	-0.06	1.98	0.74
12	0.39	-0.23	0.08	27	-0.16	-0.37	0.00
13	-0.62	0.67	-0.16	28	-1.47	-1.12	-1.77
14	-2.21	0.45	-2.03	29	-0.48	0.55	-0.13
15	1.12	-0.63	0.45	30	0.42	-0.11	0.40

is the dependent variables, which are the quality characteristics of interest in the process. Three classification methods for the values of the dependent variable Y are used to create the best N-classes based on one of the following criteria: Equal width intervals (EW), K-means clustering (KM), percentage of standard deviation ($\%STD$) between the observations in each class. The EW classifies the values of the dependent variable in equal width space. The KM creates N-classes using the K-means technique [144]. While the $\%STD$ obtains the N-classes based on the percentage of the standard deviation of the Y values [144]. For more detail, see [24, 179].

For example in table 6.1 where m=30 and n=2, let us assume 8 classes (N=8) for observations that are created by using the KM method. The reason is given later in section 6.4.1. The 8-classes are equivalent to 7 thresholds, as depicted in table 6.2.

Classification of the data at each threshold

Each threshold τ_i classifies the dataset Ω into two classes. We call them positive class Ω_i^+ and negative class Ω_i^-, as is the convention in the traditional LAD. The Ω_i^+ (Ω_i^-) contains positive (negative) observations that have Y values equal to or greater (less) than the value of τ_i.

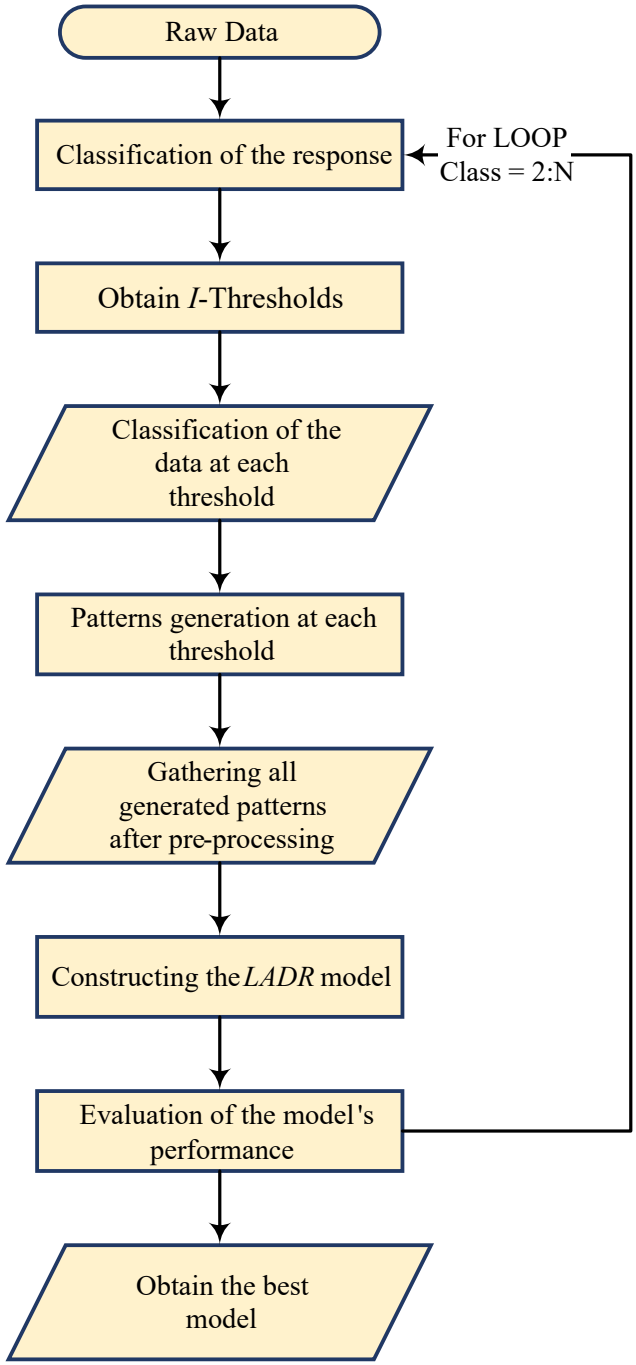


Figure 6.1 The $LADR$ methodology flow chart

$$\begin{aligned} \Omega_i^+ &= \{\omega \in \Omega \mid Y(\omega) \geqslant \tau_i\}, i = 1, ..., I \\ \Omega_i^- &= \{\omega \in \Omega \mid Y(\omega) < \tau_i\}, i = 1, ..., I \end{aligned} \tag{6.1}$$

Where Y (ω) is the dependent variable at the observation ω and i=1,. . . ,I is the thresholds

Table 6.2 The classes and thresholds using the KM method for the illustrative example

C0	C1	C2	C3
$\tau < \tau_1$	$\tau_1 \leq \tau < \tau_2$	$\tau_2 \leq \tau < \tau_3$	$\tau_3 \leq \tau < \tau_4$

C4	C5	C6	C7
$\tau_4 \leq \tau < \tau_5$	$\tau_5 \leq \tau < \tau_6$	$\tau_6 \leq \tau < \tau_7$	$\tau \geq \tau_7$

τ_1 = -2.03 , τ_2 = -0.84 , τ_3 = -0.44, $\tau_4 = 0$,
$\tau_5 = 0.4$, $\tau_6 = 0.7$, $\tau_7 = 1.12$

index. For the first threshold (τ_1) in table 6.2, the observations are in the Ω_i^+ when the value of the dependent variable Y, is greater or equal to -2.03, otherwise they are set in the Ω_i^- as shown in table 6.3. In this table, the positive and negative classes are denoted by (+) and (-), respectively. This procedure is repeated for all seven thresholds that are obtained by the KM method and that are shown in table 6.2.

Table 6.3 Defining the positive and negative classes for the first threshold (τ_1) using the KM method.

Class	X_1	X_2	Y	Class	X_1	X_2	Y
+	-0.63	1.33	0.53	+	-0.04	-0.71	-0.22
+	0.18	-0.09	0.12	+	-0.02	0.36	0.00
+	-0.84	0.35	-0.44	+	0.94	0.81	1.12
+	1.60	0.03	1.42	+	0.82	-0.07	0.72
+	0.33	-1.36	-0.38	+	0.59	0.91	0.70
+	-0.82	-0.46	-0.84	+	0.92	0.44	0.83
+	0.49	-0.37	-0.16	+	0.78	-0.57	0.45
+	0.74	-0.02	1.01	+	0.07	0.34	0.49
+	0.58	1.13	0.95	-	-1.99	-1.23	-2.58
+	-0.31	0.75	0.54	+	0.62	1.46	1.22
+	1.51	-0.09	1.42	+	-0.06	1.98	0.74
+	0.39	-0.23	0.08	+	-0.16	-0.37	0.00
+	-0.62	0.67	-0.16	+	-1.47	-1.12	-1.77
+	-2.21	0.45	-2.03	+	-0.48	0.55	-0.13
+	1.12	-0.63	0.45	+	0.42	-0.11	0.40

Generation of patterns at each threshold

Based on the previous step, the I–thresholds; τ_1, τ_2,... τ_I; creates I-datasets, Ω_i where i= 1,...,I and I=N-1 where N is the number of classes, for example I=7 and N=8 in table 6.2. For each threshold τ_i, the *cbmLAD* software [24] performs as a two-class classification. It generates patterns that differentiate between the two classes. The *cbmLAD* software uses ant colony optimization technique to generate strong patterns within the original data by solving the following optimization problem:

$$\max. \sum_{\phi \in \Omega^+} \prod_{\substack{k=1 \\ \phi_k \neq \psi_k}}^{d} (1 - y_k) \tag{6.2}$$

$$\text{s.t.} \sum_{\substack{k=1 \\ \gamma_k \neq \psi_k}} y_k \geqslant 1 \quad , \forall \gamma \in \Omega^- \tag{6.3}$$

$$y_k \in 0, 1 \quad , \forall k = 1, ..., d \tag{6.4}$$

Where d is the number of attributes in a pattern,
$y_k = 1$ if the kth attribute of the pattern equals ψ_k, 0 otherwise,
ψ_k : the value of the kth attribute in the ψ observation,
ϕ_k : the value of the kth attribute in the ϕ observation, where $\phi \in \Omega^+$
γ_k : the value of the kth attribute in the γ observation, where $\gamma \in \Omega^-$.

Consequently, we generate two sets of patterns at each threshold i, i=1,..., I. We call the class of the set observations in the class i 'Positive', while the class of the set of observations outside the class i is 'Negative'. A pattern identifies multidimensional zones that characterizes different groups of observations in the original data. It covers at least one observation in the I-class, while this pattern does not cover the observations of all other classes. Therefore, by gathering the two sets of patterns, Pi, i=1,..., I, is obtained including both positive and negative patterns at each threshold τ_i as depicted in equation (6.5).

$$Pi = \{Pi^+\} \cup \{Pi^-\}, i = 1, ..., I \tag{6.5}$$

Referring to our illustrative example in tables 6.1 and 6.3, the *cbmLAD* software [24] obtains a pattern set $P1$=$\{P1^+@(Y \geq -2.03)\}$ $\cup\{P1^-@(Y < -2.03)\}$ at the first threshold τ_1 as shown in table 6.4.

Table 6.4 The generated patterns at the first two thresholds of the KM method for the illustrative dataset and the binary values of the patterns' independent variables X_{P_j}, j=1,. . . ,5

No.	Threshold 1 (τ_1)			Threshold 2 (τ_2)	
	X_1<-1.73 X_2<-1.175	X_2>-1.175	X_1>-1.73	X_1<-1.155	X_1>-1.155
	X_{P_1}	X_{P_2}	X_{P_3}	X_{P_4}	X_{P_5}
1	0	1	1	0	1
2	0	1	1	0	1
3	0	1	1	0	1
4	0	1	1	0	1
5	0	0	1	0	1
6	0	1	1	0	1
7	0	1	1	0	1
8	0	1	1	0	1
9	0	1	1	0	1
10	0	1	1	0	1
11	0	1	1	0	1
12	0	1	1	0	1
13	0	1	1	0	1
14	0	1	0	1	0
15	0	1	1	0	1
16	0	1	1	0	1
17	0	1	1	0	1
18	0	1	1	0	1
19	0	1	1	0	1
20	0	1	1	0	1
21	0	1	1	0	1
22	0	1	1	0	1
23	0	1	1	0	1
24	1	0	0	1	0
25	0	1	1	0	1
26	0	1	1	0	1
27	0	1	1	0	1
28	0	1	1	1	0
29	0	1	1	0	1
30	0	1	1	0	1

Gathering all generated patterns

All of the positive and negative patterns generated at every threshold are gathered in one pattern set $\mathcal{P}$=$\cup Pi$, i=1,. . . ,I. Each pattern in this dataset $\mathcal{P}$ is represented by binary independent variable X_{P_j} where j= 1,...,J which is an indication of the existence of the pattern j. When the pattern j covers an observation in the original dataset Ω, X_{P_j} takes the value 1, and 0 otherwise. As in the illustrative example, the total number of generated patterns for the 7-thresholds is 31. For illustration, table 6.4 shows the generated patterns for the first two thresholds only. The X_{P_1}=1 covers the 24th observation only at threshold τ_1 because it is covered by pattern 1, and the X_{P_2} =1 for all observations except the 5th and 24th observations, which are not covered by pattern 2.

Data pre-processing is applied on the independent variables X_{P_j} j=1,. . . ,J in the dataset $\mathcal{P}$ to remove the irrelevant or redundant information that is present in the structure of the dataset. It is carried out in four steps to remedy the problems due to the presence of duplicated and

correlated X_{P_j}, the dependencies between X_{P_j}, and the multi-collinearity. For more details, see [179]. As in table 4, the correlation between the $X_{P_4}$and X_{P_5} is -1 where both patterns are complementary to each other. In this case, one of each pair is removed. Table 6.4 shows the resulting binary independent variables X_{P_j}, j=1,2,3,4 after removing the repeated patterns.

Computing the coefficients of the patterns in the regression model

The regression model represents the relationship, linear or non-linear between the remaining independent variables X_{P_j}, j=1,...,J and the dependent variable Y. Without loss of generality, in this paper, the linear regression algorithm is used to build the model. The coefficients of the regression model are obtained based on the minimization of the Mean Square Error (MSE). Moreover, The F-test is carried out to detect the insignificant variables whose p-value > 0.05. For more detail, see [179].

The $LADR$ model replaces the original independent variables with the patterns variables as shown in equation (6.6).

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} (\beta_j X_{P_j}) \tag{6.6}$$

Where $\hat{Y}$ is the predicted value of the $LADR$ model, β_0 is the intercept, and β_j is the coefficient of the pattern j, and X_{P_j} is a binary (0, 1) variable for all j= 1, ...,J, indicates whether the estimated value of an observation ,$\hat{Y}$, is correlated to X_{P_j}=1 or not when X_{P_j}=0, for pattern j. The $LADR-KM$ regression equation and the patterns that have X_{P_j}=1 are presented in equation (6.7). For this equation, all of the 7-thresholds τ_i, i=1,...,7 that are given in table 6.2 are taken into consideration. The details of the patterns, the covered zones, and the classes that they belong to, are depicted in table 6.5.

$$\begin{aligned}\hat{Y} = &- 0.1671 + 0.1957X_{P_2} + 0.6257X_{P_3} - 0.9414X_{P_6} - 0.3772X_{P_7} + 0.2785 \\ &X_{P_9} + 0.2585X_{P_{13}} + 0.3944X_{P_{15}} + 0.3278X_{P_{19}} + 0.2586X_{P_{21}} - 0.4719 \\ &X_{P_{23}} - 0.796X_{P_{26}} - 0.2655X_{P_{27}} - 0.3684X_{P_{28}}\end{aligned} \tag{6.7}$$

Evaluation of the model's performance

10-fold cross-validation for 10 replications is used to assess the performance of the model. Consecutive iterations are done by creating a different number of classes in order to find the

Table 6.5 The pattern's covered zones and classes

P_j	Covered zone	Class	P_j	Covered zone	Class
P_2	X_2>-1.175	C1,C2,C3,C4, C5,C6,C7	P_{19}	X_1>0.405 X_2>-0.17	C4,C5,C6,C7
P_3	X_1>-1.73	C1,C2,C3,C4, C5,C6,C7	P_{21}	X_1>0.025 X_2>-0.08	C5,C6,C7
P_6	X_1<-1.155	C0,C1	P_{23}	X_1<0.535 X_2<1.395	C0,C1,C2,C3, C4,C5
P_7	X_1<-0.725 X_2<-0.415	C0,C1,C2	P_{26}	X_1>-0.11 X_2>-0.1 X_2>0.78	C6
P_9	X_1>-0.725	C3,C4,C5, C6,C7	P_{27}	X_1<0.93 X_2<1.395	C0,C1,C2,C3, C4,C5,C6
P_{13}	X_1>-0.395 X_1<0.455 X_2>-0.67	C3,C4,C5, C6,C7	P_{28}	X_1<1.315 X_2<0.78	C0,C1,C2,C3, C4,C5,C6
P_{15}	X_2>0.71	C5,C6			

best model based on MSE. Consequently, the appropriate number of thresholds is determined for the data given in table 6.2. We divide the original dataset in the illustrative example into 8 classes (KM=8).

6.4.2 *LADR* regression-based control chart

In this section, we describe the process monitoring strategy, which is called the *LADR*-control chart. The *LADR* model is integrated with the control chart and acts as a modeling framework to monitor the dependent variable in that process. This combination of *LADR* and the control chart develops a new anomaly detection scheme. We consider the historical data of a process of n variables $X_1,\ldots,\ X_n$, which indicate the process performance as inputs. The output response that indicates the quality characteristic of interest in the process is the Y variable. The *LADR* technique creates a model that describes this data and predicts the output $\hat{Y}$. The independent variables of the *LADR* model represent the generated patterns' variables using the appropriate classification method at the best number of thresholds, as shown in section 6.4.1. The residuals of the *LADR* model are used to construct the control chart. The residual (E_t) is the difference between the measured value of the dependent variable and its corresponding predicted value based on the regression model given in equation (6.7) at time t; E_t= Y_t-$\hat{Y}_t$. The residual term is independent and normally distributed with $N(0,\sigma_E^2)$. Consequently, the control chart monitors the model's residuals to determine whether the process is in-control or out-of-control. In this paper, we use Exponential Weighted Moving Average (*EWMA*) control chart to monitor the process quality,

because it is considered an alternative approach to the Shewhart control chart in the case of small shifts in the process mean. It keeps a memory of the process' history through the recursive equation (6.8). The $EWMA$ is computed recursively for the available samples by:

$$z_t = \lambda \bar{E}_t + (1-\lambda) z_{t-1} \tag{6.8}$$

Where z_t , the value of $EWMA$ at the tth sample data, which is equivalent to time t, and the $\bar{E}_t$ is the mean value model's residuals at the tth sample or subgroup. λ represents a smoothing parameter ranging between 0 and 1 (i.e. $0 \leq \lambda \leq 1$). If the value of λ gets smaller, the smaller shifts will be quickly detected but false alarms may increase. Thus, an appropriate value should be selected to obtain accurate process monitoring. It is assumed that z_0=0, which is the target value. When the σ_E is unknown, the S_E is estimated by taking the standard deviation of the model's residuals in normal conditions. The $EWMA$ control chart graphically represents the $EWMA$ values z_t, t=1,. . . .m, with the m samples used to construct the control limits (UCL and LCL) and the centerline (CL) as the following:

$$UCL = L\sigma_E \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]} \tag{6.9}$$

$$LCL = -L\sigma_E \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]} \tag{6.10}$$

$$CL = 0 \tag{6.11}$$

Where, L is the width of control chart limits, UCL and LCL, the choice of design parameters of $EWMA$, L and λ are based on the desired Average Run Length (ARL) performance as described by Lucas and Saccucci [180] and Steiner [181]. ARL is the average number of samples that must be plotted in the control chart after an assignable cause has happened and before a sample mean falls outside the control limits, thus declaring the process to be out-of-control [1]. When the process is in-control, large in-control ARL (ARL_0) contributes to a reduction in false alarms. Conversely, small out-of-control ARL (ARL_1) is needed for out-of-control processes in order to rapidly detect the change [59]. Hence, the ARL_1 is much smaller than the ARL_0.

Referring to the illustrative example, the dataset represents an in-control process where the $LADR$ model's residuals follow $N(0,S_E^2)$. The estimated standard deviation of the model's residual for the in-control process operation is S_E=0.086. We select the in-control average

run length (ARL) of the chart at 370 based on error type (α= 0.0027). We determine the optimal design parameters at one standard deviation in the process mean using libraries of "*SPC*" and "*qcc*" in R-4.0.5 software [154]; $\lambda = 0.14$ and $L = 2.79$; to construct the control limits using equations (6.9-6.11). As an illustration, we generated the observation of table 6.6, where the means of X_1 and X_2 are shifted with magnitude δ equal to 1.

Table 6.6 Generation of special cause

No.	X_1	X_2	Y
31	0.46	0.47	0.45
32	2.21	2.40	2.56
33	2.16	0.84	1.89
34	1.70	0.86	1.33
35	2.59	0.98	2.02
36	1.56	1.74	1.48
37	-0.28	0.86	0.08
38	0.43	0.93	0.31

This shift is reflected in the dependent variable (Y_t), and accordingly the *LADR* model's residuals. The process is operating in the presence of abnormal variations which affect the process quality. The *EWMA* chart detects anomalies in the process, and the observations appear beyond the UCL from the 32nd residual point as depicted in figure 6.2.

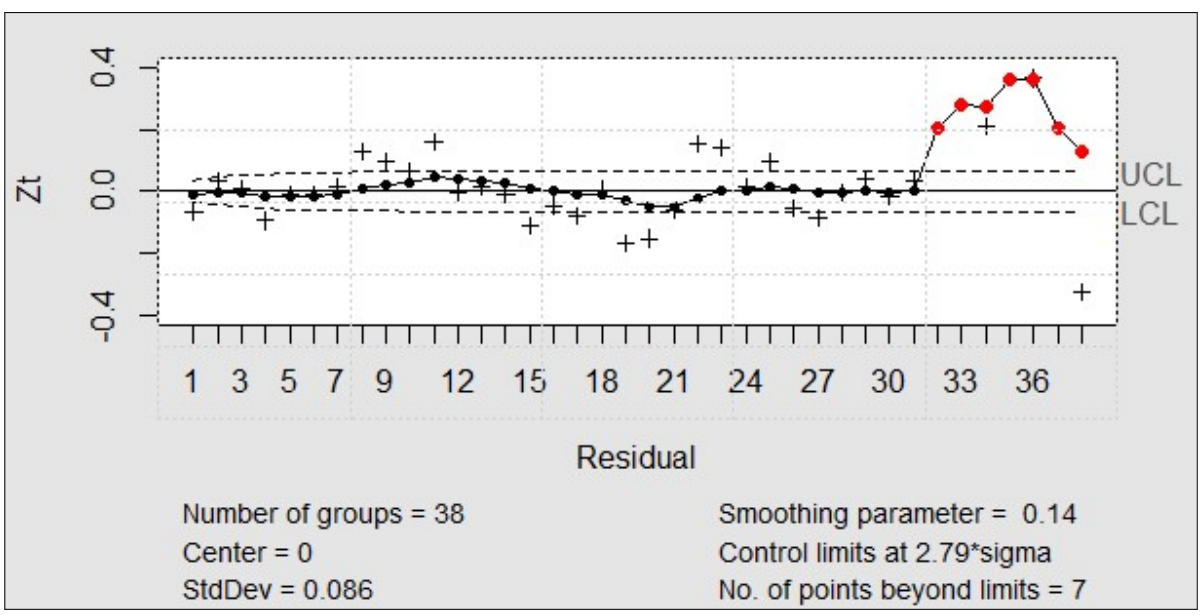


Figure 6.2 Anomaly detection using $LADR - EWMA$ chart

Once a residual goes beyond the control limits of $EWMA$ at the 32nd observation, the patterns covering this observation describe the causes of the out-of-control situation in terms of the independent variables X_1 and X_2. These causes determine the corrective actions that should be taken to return the process to in-control. The details of this procedure are shown along with an example in the next section.

6.4.3 Root cause identification of the out-of-control process

When the model's residual goes beyond the control limits of an $EWMA$ chart, the control chart detects an anomaly and provides an alarm signal. The $LADR$ model itself in equation (6.6) is used to determine the reason for that out-of-control situation using the covering patterns.

The $LADR$ model in equations (6.6) and (6.7) contains all of the patterns' variables for the generated patterns that are extracted from the original data as described in section 6.4.1. For each sample observation, some of these patterns cover it, which means X_{P_j}= 1, and others do not. The X_{P_j} =1 (X_{P_j} =0) in the model when the original independent variables are in (out) of the zone that is formed by the pattern j as explained in section 6.4.1. For the illustrative example, the $EWMA$ chart detects an anomaly at the 32nd observation in figure 6.2. Therefore, the model of the predicted dependent variable $\hat{Y}_{32}$ is as follows:

$$\begin{aligned}\hat{Y}_{32} = &- 0.1671 + 0.1957X_{P_2} + 0.6257X_{P_3} + 0.2785X_{P_9} + 0.3944X_{P_{15}} \\ &+ 0.3278X_{P_{19}} + 0.2586X_{P_{21}} - 0.796X_{P_{26}}\end{aligned} \tag{6.12}$$

When an observation is out-of-control, we analyze the model at that out-of-control observation in terms of the covered patterns, the patterns' prevalence, and the location of the anomaly value in the control chart whether beyond the UCL or the LCL. The prevalence of the pattern is the proportion of observations that were covered by each pattern in the training phase of pattern generation as in equation (6.13). It represents the robustness of this pattern. High prevalence is an indication of a robust pattern.

$$Prevalence(P_j, C) = \frac{\mid Cov(\omega_{P_j,C}) \mid}{\mid Cov(\Omega_{P_j}) \mid} \quad , C = 0 : N \tag{6.13}$$

Where $Cov(\omega_{P_j,C})$ is the coverage of the jth pattern for the set of observation in class "C", $Cov(\Omega_{P_j})$ is the coverage of the jth pattern for all the dataset Ω, and the N is the number of classes.

The model consists of patterns' binary variables, X_{P_j}, that have negative and/or positive coefficients. In equation (6.12), the original independent variables of the 32nd observation is covered by the zones formed by the patterns X_{P_2}, X_{P_3}, X_{P_9}, $X_{P_{15}}$, $X_{P_{19}}$, $X_{P_{21}}$, and $X_{P_{26}}$. All the X_{P_j} in the $LADR$ model have positive coefficients except the $X_{P_{26}}$.

The $LADR$ technique uses the classification process to partition the values of the dependent variable, Y, into classes as shown in section 6.4.1. Each class represents a range of Y-values. The covered patterns are arranged in descending order based on their prevalence (coverage) in each class. When these robust patterns are presented in a class, the $LADR$ model predicts the value of the dependent variable within the interval of that class. Conversely, when the patterns with lower prevalence in a certain class exists, the $LADR$ model predicts the value of the dependent variable beyond the interval of that class. A lower prevalence indicates a robust indication of the anomaly and its causes deduced from the patterns. For our illustrative example, there are 8-classes in table 6.2; C0 to C7; and 13 significant patterns as determined in section 6.4. The prevalence of each pattern per class is presented in figure 6.3.

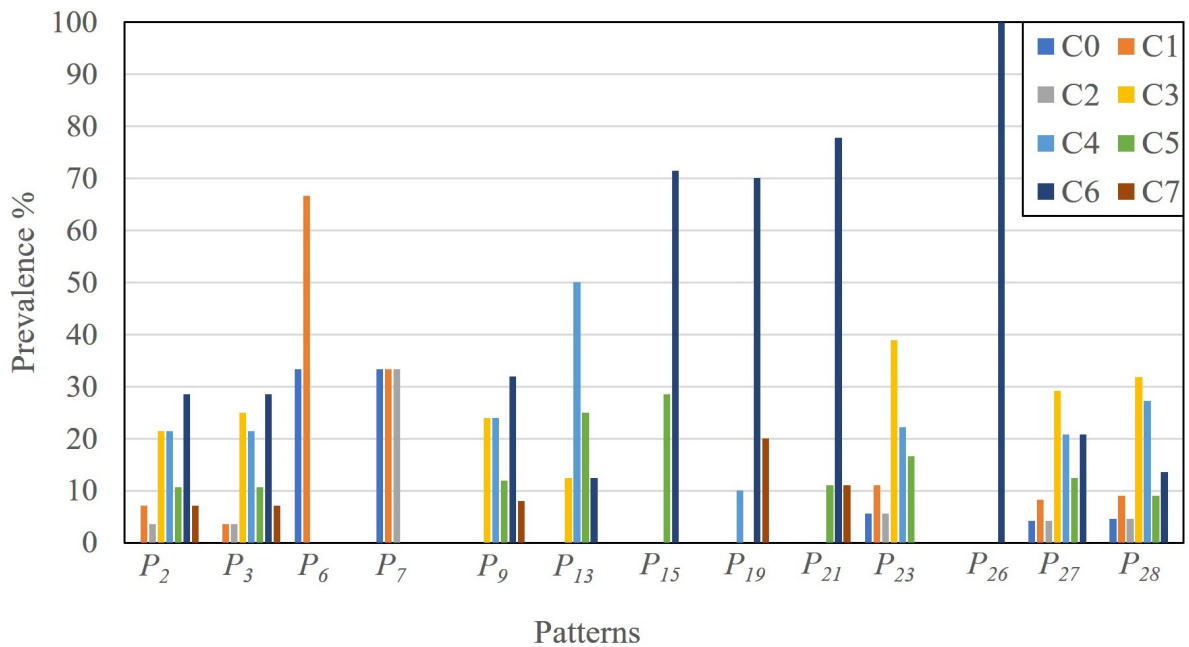


Figure 6.3 The prevalence for each pattern per each class

When the $z_t < LCL$, the residual point at time t which is the difference between the measured dependent variable and the predicted value obtained by the $LADR$ model is negative. In other words, the actual dependent variable is lower than the predicted value. Thus, the actual dependent variable belongs to class h and the predicted value belongs to class h', where $h \leq h'$. Consequently, the reason for this is the covering patterns that have positive

coefficients of lower prevalence in the $LADR$ model. Conversely, when the $z_t > UCL$, the residual point at time t is positive. Therefore, the actual dependent variable is greater than the predicted value. Subsequently, the actual dependent variable belongs to class h and the predicted value belongs to class h', where $h \geq h'$. Consequently, the reason is the covering patterns that have negative coefficients of lower prevalence in the $LADR$ model as depicted in algorithm 1.

As for the illustrative example, the actual value of the 32nd original dependent variable in table 6.6 is 2.56, and belongs to C7 (table 6.2). The predicted value is 1.1176 obtained by equation(6.12), which is also in C7. Moreover, the $LADR$ mode in equation (6.12) has binary independent variables $X_{P_{15}}$ and $X_{P_{26}}$ that have positive and negative coefficients, respectively, which are not in C7 at all; see figure 6.3. Furthermore, the coefficient of the $X_{P_{26}}$ is the highest negative. It can therefore be concluded that the patterns $X_{P_{26}}$ is the root cause of the anomaly for two reasons. The first reason is the $z_{32} > UCL$ which means the actual dependent variable (Y_{32}) is greater than the predicted one from the $LADR$ model ($\hat{Y}_{32}$). Therefore, the residual point has a positive value. The $X_{P_{26}}$ is the only covered pattern that has a high negative coefficient in the $LADR$ model. The second reason is that the P_{26} does not cover C7 as shown in figure 6.3. The description and covered zone of the pattern P_{26} is shown in table 6.5. Consequently, the corrective actions would violate this zone to eliminate the anomaly and to keep the process in-control. Thus, the variables X_1 and/or X_2 have to be changed in order to leave the zone (X_1>-0.11, X_2>0.78) by decreasing X_1 to <-0.11 and/or X_2 to <0.78; in this way $X_{P_{26}}$ would be eliminated.

Therefore, when an anomaly is detected in the $EWMA$ control chart, we determine the class of the actual dependent and predicted variable at time t, and the covered patterns in the model for that instance. We identify the covered patterns, which do not cover the actual class or have very low prevalence in that class. Accordingly, the corresponding values of X_1,.., X_n are considered for root cause analysis of the out-of-control signal. This allows for the corrective actions that should be taken to retain the process to in-control to be determined. The details of this procedure are demonstrated through an example in the following section.

6.5 Numerical example

6.5.1 Dataset description

Concrete manufacturing process is a highly complex and sensitive process responsible for producing high-performance concrete. The process is based on mixing cement with various ingredients to satisfy the construction materials standards. There are common types of con-

Algorithm 1 *LADR* Technique - Root cause

Read τ ▷ The I-thresholds
PV= *Prevalence* (P_j, C) ▷ Refer to equation 6.13
Read E_t, Y_t, and $\hat{Y}_t$ ▷ The out-of-control observation at time t
X_P =[ ] ▷ The pattern set covered by the observation at time t
Co =[ ] ▷ The corresponding coefficients of the pattern set at time t
if Cover(P_j,ω_t) **then** ▷ If the P_j covers the observation ω at time t
X_{P_j}=P_j $\forall$ j=1,...,J
Co_j = coefficient $(LADR(X_{P_j}))$ $\forall$ j=1,...,J
else
    X_{P_j}=0 , Co_j=0
end if
for $i \leftarrow 1$ **to** I **do**
    if Y_t<τ_i **then** C=i-1
    else
        C=I
    end if
end for
if $z_t > UCL$ **then**
    A=find $(X_{P_j}, PV(X_{P_j}, C), Co_j(X_{P_j})<0)$ ▷ Stored the patterns which have negative coefficients, their prevalence, and their coefficients' values at the out-of-control observation ω at time t in table "A"
    A= Sort A $(PV(X_{P_j}, C))$ ▷ Sort table A based on the prevalence of patterns in ascending order
    Υ = Min A_{i2} ▷ Obtain the indices of the patterns of minimum prevalence of patterns in the 2nd column ($i2$) of A
    ξ=Min$_{\upsilon \in \Upsilon}$ A_{i3} ▷ Obtain the index of the pattern of minimum coefficient corresponding to the minimum prevalence index (υ) in the 3rd column ($i3$) A
    RC= $A_{i1}(\xi)$ ▷ Root cause of the out-of control observation ω at time t
end if
if $z_t < LCL$ **then**
    A=find$(X_{P_j}, PV(X_{P_j}, C), Co_j(PV(X_{P_j}))>0)$
    A= Sort A
    Υ = Min A_{i2}
    ξ=Max$_{\upsilon \in \Upsilon}$ A_{i3}
    RC= $A_{i1}(\xi)$
end if
Print A
Print RC

crete such as plain, precast, reinforced, etc. High-performance concrete is characterized by high workability, which ensures high strength, high stability, and high durability [182]. Consequently, quality monitoring of the concrete manufacturing process is necessary to ensure the

concrete production is systematically controlled and complies with pre-specified characteristics of the construction materials standards. It involves data collection for the measurements of the variables affecting this process that lead to the presence of anomalies. So, a dataset of a concrete production process is taken from the "UCI Machine Learning Repository", which describes a variety of concrete designed mix [151]. The training set consists of 78 measurements, which are considered as the in-control process.

Each measurement in the dataset has seven ingredients that represent the independent variables and three dependent variables. In this paper, we focus on only one dependent variable which is the compressive strength of concrete in MPa. On the other hand, the independent variables are: cement content in the mixture (X_1), blast furnace slag (X_2), fly ash (X_3), water content in the mixture (X_4), Superplasticizer or high range water reducers in the mixture (X_5), fine sand aggregate (X_6), and coarser sand aggregate (X_7). All of these variables are measured in Kg per cubic meter of concrete.

The concrete strength is specified by a designer, while the concrete producer determines the proportions in the mixture. It is influenced by the proportion of these ingredients in the mixture. When the water content (X_4) increases too much in the mixture, voids occur in the concrete, which decrease its strength. Therefore, the ratio between the water (X_4) and cement (X_1) behaves inversely with strength. However, too high of a proportion of cement (X_1) in the mixture may lead to the occurrence of cracks. Sometimes superplasticizers (X_5) are added to the mixture to reduce the high water content (X_4) and to increase its strength. Supplementary cementitious materials such as blast furnace slag (X_2) and fly ash (X_3) enhance the concrete's cohesiveness. The fine sand aggregate (X_6) and coarser sand aggregate (X_7) reduce the strength of the concrete when they represent a high proportion of the mixture [183]. Therefore, the proportion within the mixture must be accurately designed to obtain the desired strength and the characteristics of the concrete.

Generally, the testing procedures of the concrete strength inevitably contain experimental errors. In this paper, the proposed $LADR-EWMA$ is implemented to monitor the variation of the concrete strength based on the desired strength to ensure the process quality. So, a $LADR$ model is developed to model the concrete strength using patterns generated from the original ingredients. Then, the $EWMA$ chart monitors the residuals, obtained by the $LADR$ model, to detect any anomaly experience in the process. The variations in the values of one or more inputs may lead to a shift in the process resulting out-of-control conditions. This is associated with the predicted dependent variable of the model $\hat{Y}$ leading to a mean shift beyond the pre-specified limits of the concrete strength. To assess the performance of the $LADR$ based control chart in the detection of the process mean shift of residuals, a

simulated environment has been carried out to generate an additional set of 40 measurements in which the first 30 represent the in-control process. The last ten create a variation in the ingredients, which affects the concrete strength, which leads to an out-of-control process. Furthermore, we compare the results of our proposed *LADR-EWMA* with the integration of four other different regression techniques based EWMA. These use Linear Regression (*LR-EWMA*), Support Vector Regression (*SVR-EWMA*), Partial least square regression (*PLS-EWMA*), and Multivariate adapative spline regression [184] (*MARS-EWMA*). We implement the techniques using "*e1071*", "*pls*", "*earth*"packages in R-4.0.5 software [154].

6.5.2 Development of the *LADR* regression models

Based on the simulated data, we develop $LADR$ models using KM and EW classification methods to obtain a regression model that describes the relationship between the X_1 to X_7 and Y variables. In this example, we obtain a regression model for concrete strength (Y) as a function of generated patterns from all concrete ingredients in the dataset. In other words, Y is the dependent variable, and generated patterns are the independent variables. After consecutive iterations, the best model, which has the lowest value of MSE using 10-fold cross-validation for 10 replications, is determined for each method using "caret" package in R-4.0.5 software [154]. We select the best model, which in this case is $LADR - EW$ as depicted in table 6.7. In addition, Table 6.7 demonstrates the results of the other regression models. The $LADR - EW$ has better performance than the other regression models in terms of MSE, R^2, and MAE. Nevertheless, the $MARS$ model is still competitive. The structure of the $LADR - EW$ regression model is shown in equation (6.14).

Table 6.7 The performance of the $LADR$ models for the Concrete manufacturing process

Method	Threshold	MSE	R^2	MAE
$LADR - KM$	5	4.26	93.63	1.72
$LADR - EW$	12	3.08	95.75	1.37
LR	-	7.21	88.42	2.05
SVR	-	6.91	88.74	2.03
PLS	-	8.18	86.36	2.10
$MARS$	-	3.15	94.90	1.42

$$\begin{aligned}\hat{Y} =& 37.273 - 3.271X_{P_7} - 3.447X_{P_8} - 2.646X_{P_{15}} - 1.110X_{P_{27}} - 1.974X_{P_{29}} - \\ & 1.891X_{P_{30}} + 1.317X_{P_{32}} + 1.634X_{P_{34}} - 0.953X_{P_{36}} - 1.957X_{P_{38}} - 1.289 \\ & X_{P_{39}} - 3.659X_{P_{40}} - 1.935X_{P_{41}} + 1.934X_{P_{45}} - 1.537X_{P_{50}} - 2.470X_{P_{54}} - \\ & 1.455X_{P_{61}} + 0.967X_{P_{62}} - 1.214X_{P_{64}} + 2.865X_{P_{65}} + 1.616X_{P_{67}} + 3.942X_{P_{71}} \\ & + 6.503X_{P_{72}} + 3.912X_{P_{82}}\end{aligned} \tag{6.14}$$

6.5.3 Construction of the control charts

In order to construct the EWMA chart, we define two key values: 1) the mean $\bar{E}$ and standard deviation (S_E) or standardized residual ($S_{E_{MODEL}}$) for the regression model's residuals, E, in the in-control operation, 2) the parameters (λ, L) of EWMA chart. In this example, we use the 78 historical data to estimate mean and standardized residuals for the $LADR-EW$ model, which are $\bar{E}$=0, $S_{E_{LADR-EW}}$=1 respectively. We select the in-control average run length (ARL) of the chart at 370 based on error type (α =0.0027). We are targeting a shift of one standard deviation (δ=1) in the $\bar{E}$. Therefore, the optimal design parameters of the EWMA chart are determined; $\lambda = 0.14$ and $L = 2.79$; to construct the control limits.

6.5.4 Results of the $EWMA$ charts

A residual based $EWMA$ chart is developed to monitor the process quality and detect any anomalies that affect the process or the measuring equipment. Therefore, in this process the residuals obtained by the $LADR-EW$ based on the patterns are implemented on the EWMA chart. When the value of EWMA at t^{th} sample lies within the UCL and LCL, the concrete manufacturing process is in-control, otherwise, it implies a mean shift in the process. Figure 6.4 shows the 78 measurements that are used in the EWMA chart for phase I. In phase II, we used the LADR model to predict the concrete strength of the 40 simulated measurements. When we implement the predicted values using $LADR-EW$ on the $EWMA$, minimum variations in the residual are shown within the control limits in figure 6.5. In other words, the strength of the produced concrete is within the limits that are satisfactory to the desired strength for the first 30 measurements. After the 31st residual point of concrete manufacturing process no. 31, it deviates beyond the LCL, which shows a negative trend. In this section, the results demonstrate that the proposed integrated approach satisfactorily detects anomalies in the process without/reducing any false alarms.

Figures (6.6 - 6.9) present the integration of other regression models with the $EWMA$. In

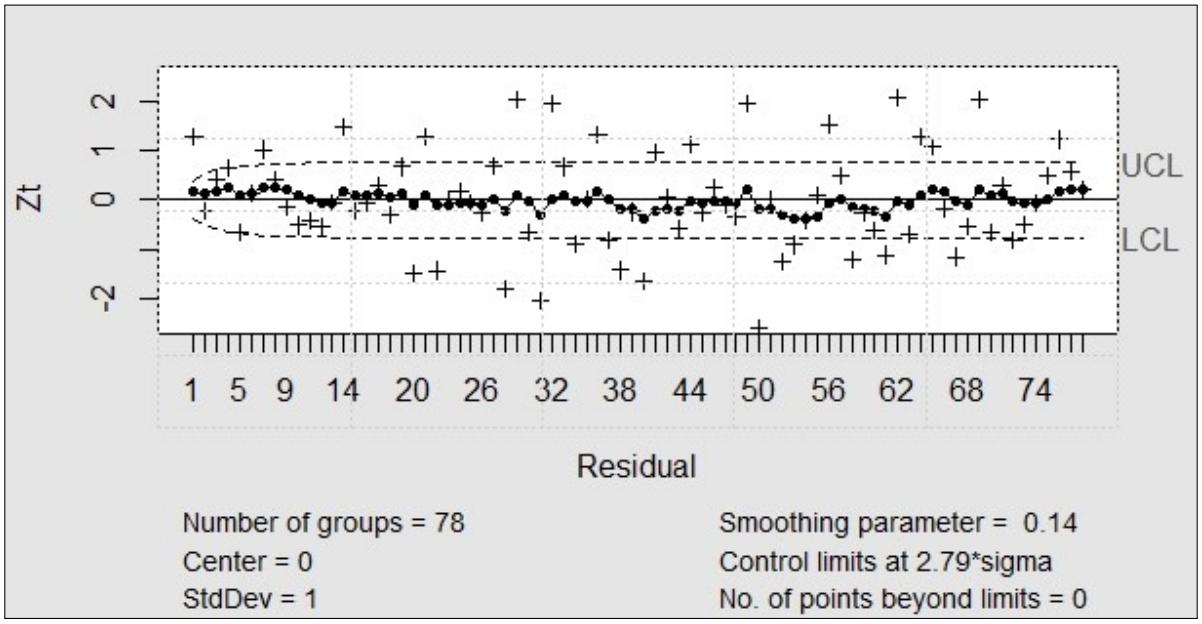


Figure 6.4 Concrete manufacturing: *LADR-EW* based *EWMA* chart for Phase I

order to compare the detection performance using the $LADR$ model or other models, we use two metrics to evaluate the performance of monitoring the process quality, the false alarm rate (FAR) and missed detection rate (MDR) [185]. The FAR represents the ratio between the number of the false signals to the total number of normal samples. On the other hand, the MDR is the ratio of the missed detection samples to the total number of anomalous samples. Both FAR and MDR are formulated as follows:

$$FAR = \frac{FP}{FP + TN} \tag{6.15}$$

$$MDR = \frac{FN}{TP + FN} \tag{6.16}$$

Where FP (FN) is false positive (negative) when a normal (an anomalous) observation is detected as anomalous (normal) one, TP is true positive when an anomalous observation is detected correctly, TN is true negative when the observation is correctly normal. In addition, $(FP+TN)$ represents the normal observations while $(TP+FN)$ is the anomalous observations.

The comparison in table 6.8 reveals that the *LADR-EW*-EWMA has the ability to correctly detect the mean shift in the process without any false alarms, as well as *LR-EWMA*, *SVR-EWMA*, and *PLS-EWMA* in figures 6.6, 6.7, and 6.8, respectively. However, the proposed

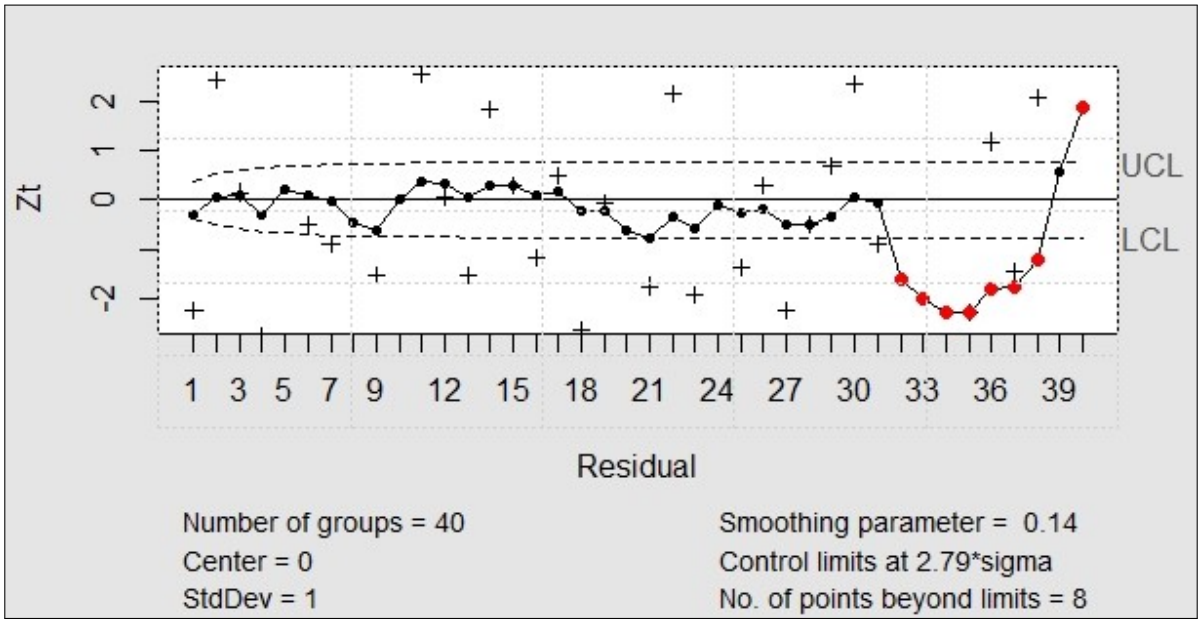


Figure 6.5 Concrete manufacturing: *LADR-EW* based *EWMA* chart for Phase II

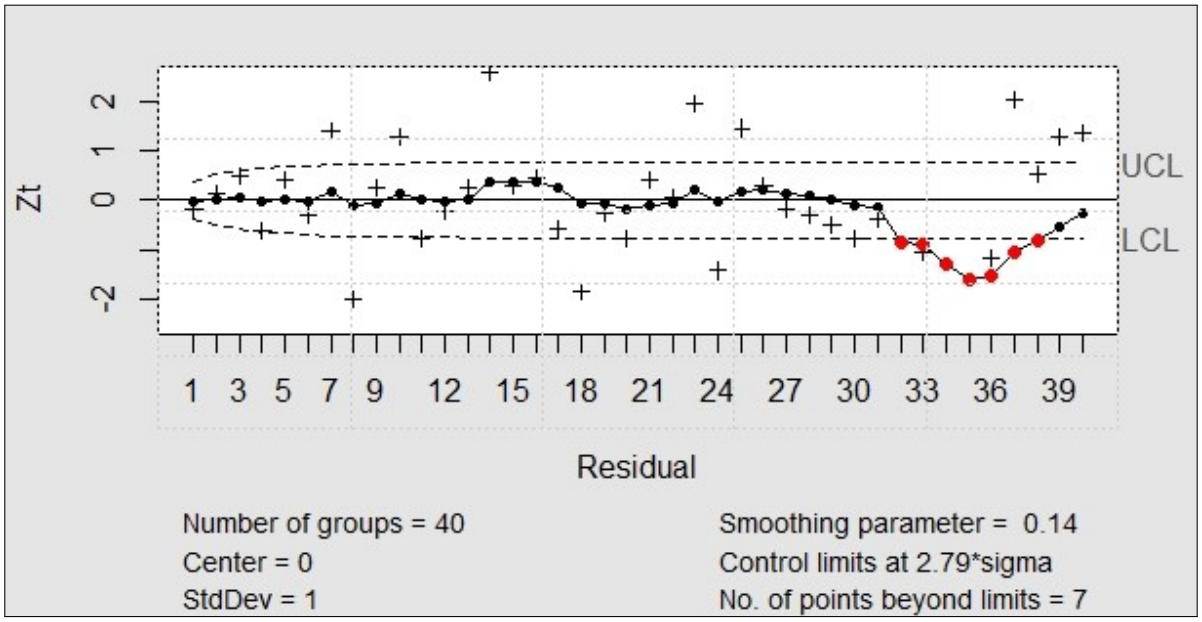


Figure 6.6 Concrete manufacturing: *LR* based *EWMA* chart

method has a slightly lower missed detection rate compared to the others. Conversely, the *MARS-EWMA* in figure 6.8 contributes to high false alarms, which affect the manufacturing process and increase downtime and economic losses, although the method is free of missed detection instances. Consequently, the *LADR-EW-EWMA* performs better than the other methods. The satisfactory mean shift detection ability of the *LADR-EW-EWMA* is due to

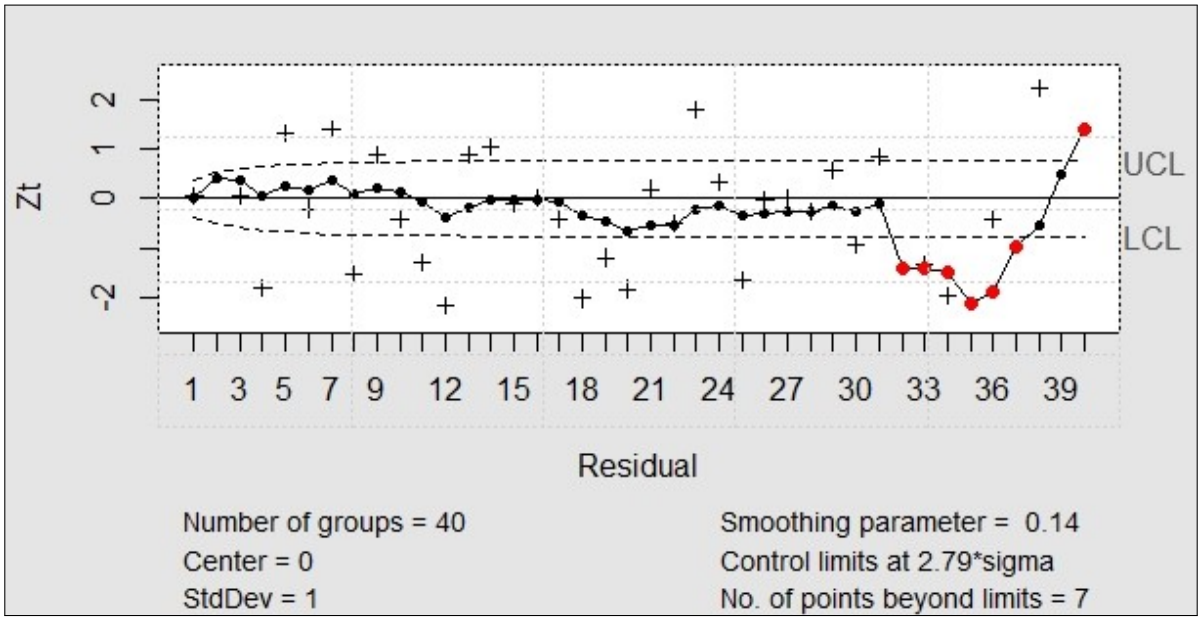


Figure 6.7 Concrete manufacturing: SVR based $EWMA$ chart

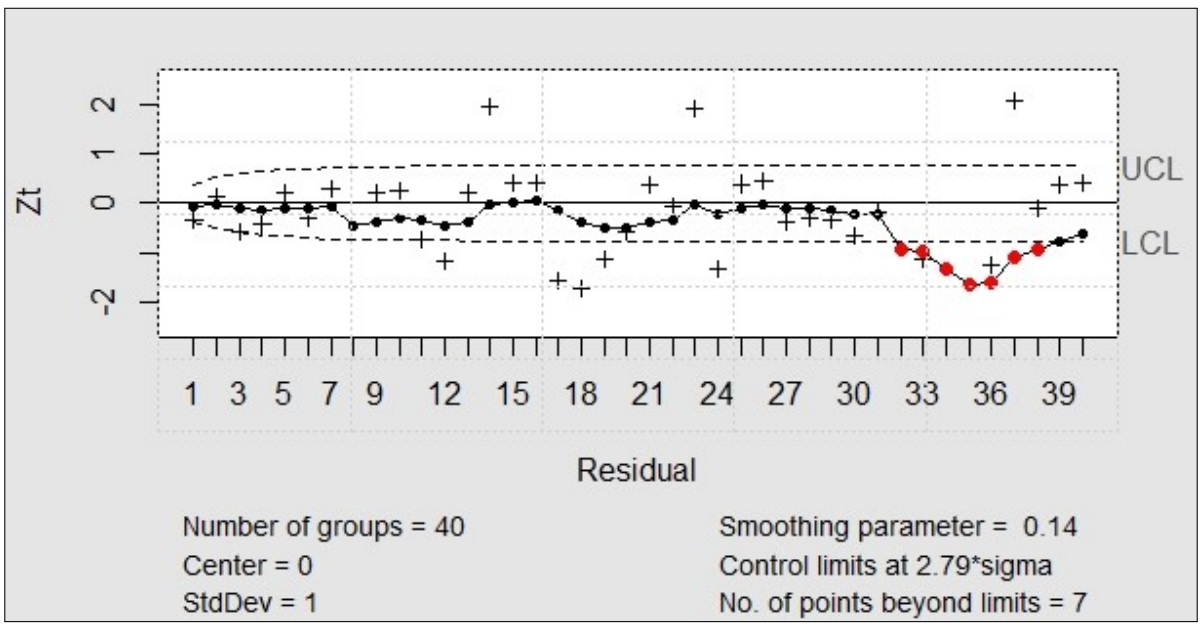


Figure 6.8 Concrete manufacturing: PLS based $EWMA$ chart

the high performance of the developed model based on the $LADR$, and the merits of the $EWMA$ to detect small shifts occur in the process.

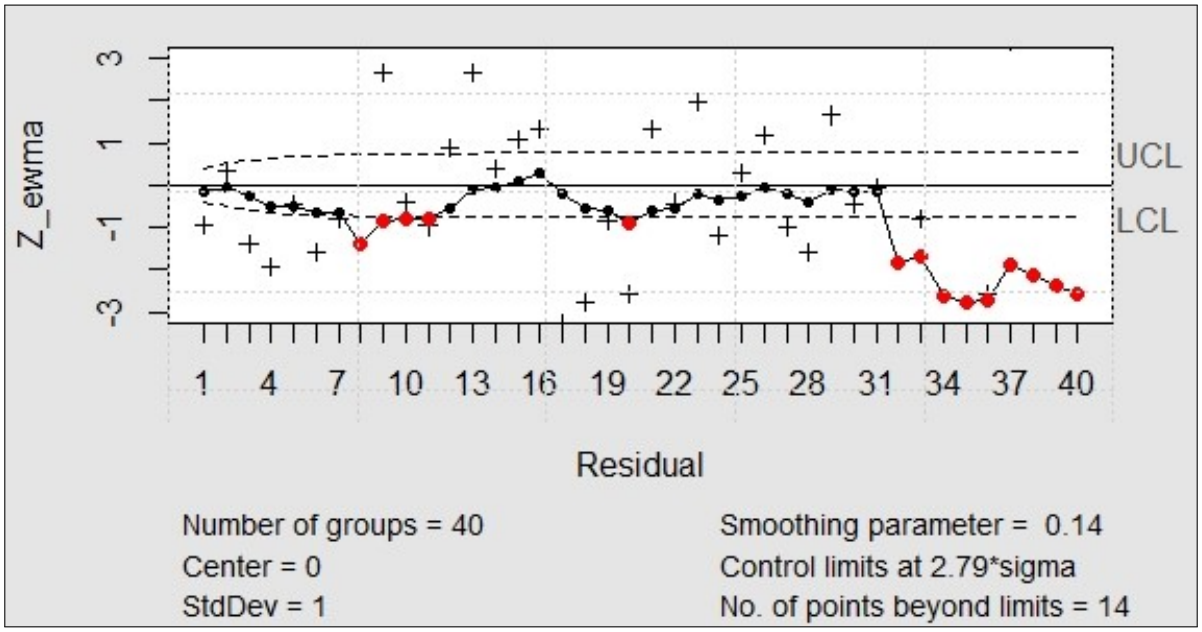


Figure 6.9 Concrete manufacturing: $MARS$ based EWMA chart

Table 6.8 The $FAR\%$ and $MDR\%$ for the regression based EWMA methods

Method	*FAR*		*MDR*	
	FP	*FAR %*	*FN*	*MDR%*
LADR-EW-EWMA	0	0	1	10
LR-EWMA	0	0	2	20
SVR-EWMA	0	0	2	20
PLS-EWMA	0	0	2	20
MARS-EWMA	5	17	0	0

$FP + TN$=30 and $TP + FN$=10

6.5.5 Diagnosis of the root cause of the out-of-control signals

The regression-based $EWMA$ chart monitors the final output process (Y) under different conditions, taking into account the independent variables that affect the (Y) and reduce the *FAR%* and *MDR%*. When the $EWMA$ chart provides an alarm, it means that the data has changed. Thus, it is important to interpret the root cause of the anomaly. All of the previously mentioned integrations require additional tools to analyze the reason for the alarm. Hence the importance of the proposed *LADR-EW-EWMA* is to interpret the root cause of the anomaly without resorting to the use of any additional tools.

Once the anomaly is detected, we investigate the causes of this alarm signal in the $EWMA$

chart. The strength of the *LADR* model is derived from using the generated patterns as independent variables X_{P_j} instead of the original variables. These patterns are used to determine the root causes of the out-of-control observations. The mix proportion of the 32nd measurement sample are ($X_1 = 286$, $X_2 = 17$, $X_3 = 151$, $X_4 = 169$, $X_5 = 12$, $X_6 = 990$, $X_7 = 713$) and the desired concrete strength is (Y_{32}=29.96). In equation (6.14), the patterns' variables X_{P_j} that do not cover the 32nd measurement sample and their values are eliminated. Equation (6.17) gives the resulting regression model for the 32nd measurement sample after eliminating the patterns' variables with zero values. The patterns' variables that have a value equal to 1 are $X_{P_{34}}$ and $X_{P_{65}}$. For the 32nd measurement sample, table 6.9 defines the full description of these patterns and summarizes the zone that is covered by each pattern.

$$\hat{Y}_{32} = 37.273 + 1.634X_{P_{34}} + 2.865X_{P_{65}} \tag{6.17}$$

Table 6.9 The pattern's covered zone and class for the 32nd measurement sample

P_j	Pattern description	Covered zone
P_{34}	X_1>155 X_3>48.5 X_7<779	C5,C6,C7, C8,C9,C11
P_{65}	X_1>274.5 X_3>115.5 X_4<191.5 X_7<809.5	C7,C8,C9, C10,C11,C12

The classification process is the key building block of the *LADR-EW* model as described in section 6.4.1. The original dataset is divided into 13 classes, which are defined by 12 thresholds as depicted in table 6.10. The prevalence of each pattern in each class is based on these thresholds. We arrange all of the patterns for each class in descending order based on their prevalence in that class, as in section 6.4.3. At the 32nd measurement sample, the value of the concrete strength is 29.96, which belongs to class (C4), as in table 6.10. On the other hand, the predicted value obtained by the *LADR* model is 41.77, which belongs to class (C7).

It is observed in equation (6.17) that two patterns' variables; $X_{P_{34}}$, and $X_{P_{65}}$; have only positive coefficients. Moreover, $z_{32} < LCL$, which means that the strength predicted by the LADR model is greater than the measured strength. Consequently, the lower prevalence pattern with positive coefficient in the measured strength's class that covers the 32nd measurement sample in equation (6.17) is the reason for that anomaly. Both patterns have positive coefficients and lower prevalence in the class (C4) which should not exist in the model, as in

Table 6.10 Concrete manufacturing: The classes, zones, and thresholds using the EW method

C0	C1	C2	C3	C4	C5	C6
$\tau < \tau_1$	$\tau_1 \leq \tau < \tau_2$	$\tau_2 \leq \tau < \tau_3$	$\tau_3 \leq \tau < \tau_4$	$\tau_4 \leq \tau < \tau_5$	$\tau_5 \leq \tau < \tau_6$	$\tau_6 \leq \tau < \tau_7$

C7	C8	C9	C10	C11	C12
$\tau_7 \leq \tau < \tau_8$	$\tau_8 \leq \tau < \tau_9$	$\tau_9 \leq \tau < \tau_{10}$	$\tau_{10} \leq \tau < \tau_{11}$	$\tau_{11} \leq \tau < \tau_{12}$	$\tau \geq \tau_{12}$

$\tau_1 = 20.37$, $\tau_2 = 23.55$, $\tau_3 = 26.73$, $\tau_4 = 29.91$, $\tau_5 = 33.09$, $\tau_6 = 36.09$, $\tau_7 = 39.45$, $\tau_8 = 42.63$, $\tau_9 = 45.81$, $\tau_{10} = 48.99$, $\tau_{11} = 52.17$, $\tau_{12} = 55.35$

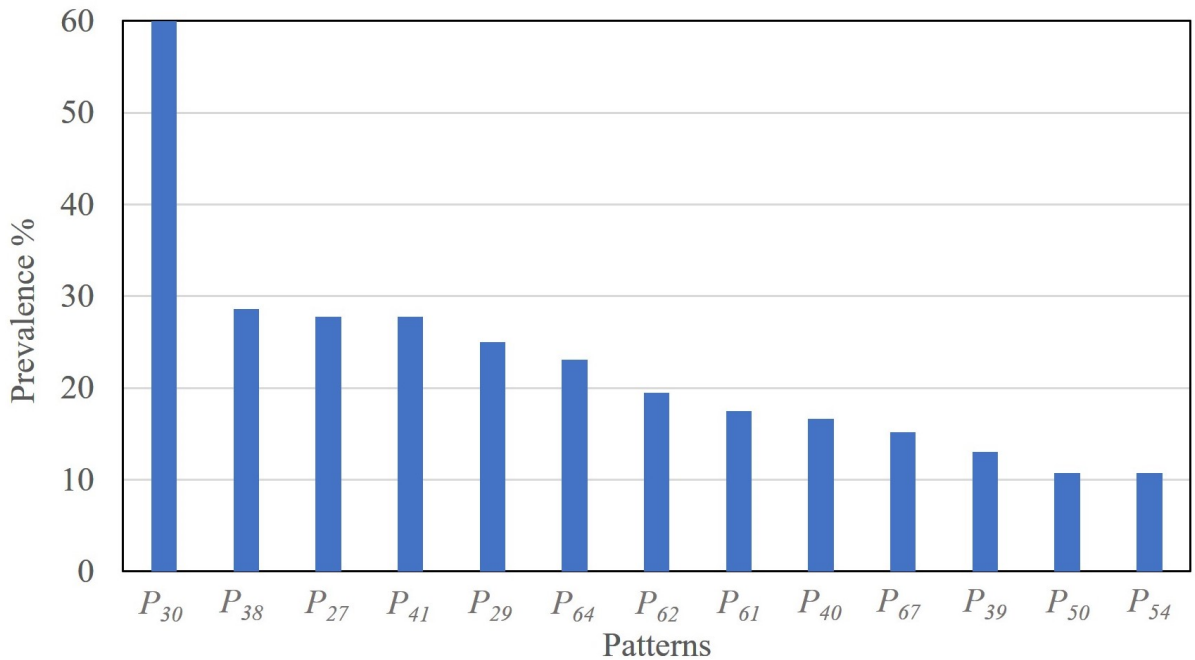


Figure 6.10 Concrete manufacturing: The prevalence for each pattern in class C4

the figure 6.10. Any pattern that is not shown in the figure 6.10 does not exist in the class (C4). Nevertheless, the $X_{P_{65}}$ has a higher coefficient than that of the $X_{P_{34}}$ in equation (6.17). Consequently, the $X_{P_{65}}$ causes the out-of-control alarm in the process. It provides an interpretation of the reason for that alarm. The conjunction of process ingredients X_{1to7} in the mixture satisfies the pattern P_{65}. The P_{65} states that the mixture has cement content>274.5, fly ash>115.5, water content<191.5, and fine aggregate<809.5. This demonstrates that the mixture has higher cementitious materials - cement and fly ash - compared to water content. Accordingly, the water/cementitious ratio in the 32nd measurement sample is about 0.37, which is an indication of lower water content in the mixture, which should produce higher strength. In order to treat this mixture, it should be adjusted to violate the pattern P_{65} to

produce the desired concrete strength. Therefore, the residuals will be within the control limits in the $EWMA$ chart, and the process will be in-control.

6.6 Conclusion

In this paper, we have constructed $LADR-EWMA$ to monitor quality and detect anomalies in the process. The $LADR$ technique is an extension of the standard LAD methodology that obtains a regression model based on generated patterns from the original dataset. These patterns have a physical meaning that maintain significant knowledge and information in that dataset. Therefore, the key feature is to obtain better independent variables that are more interpretable to understand process conditions. $LADR$ is based on three different classification methods: Equal Width intervals (EW), K-Means clustering (KM), percentage of standard deviation ($\%STD$). They transform the process response Y into N-classes to generate the patterns that characterize each interval. Consequently, $LADR$ obtains a regression model that describes the relationship between these patterns and the process response Y. Then, we implement the output of the process, which is the predicted dependent variable of the regression model, on the control chart. The $LADR$ increases the performance of the control chart to detect the process anomalies by reducing both the false alarm rate and missed detection rate. Once the control chart detects out-of-control observation, the $LADR$ analyzes the root causes of the process abnormality through generated patterns included in the model. Accordingly, appropriate corrective action can be taken to maintain the stability of the process. Generally, the proposed technique is used for anomaly detection and diagnosis of the root cause for the anomaly. Integrating the $LADR$ technique with control charts represents a reliable and robust approach for online applications.

For further research, the $LADR$ model will be adapted for decision-making to remedy the anomalous process and to maintain stability in the process. We hypothesize that when the $LADR-EWMA$ determines the pattern(s) responsible for the detected anomaly, it will be capable of specifying the values of the independent variables that violate these patterns and an action can be taken to attain an in-control process .

CHAPTER 7 ARTICLE 4: CONDITION MONITORING AND WARNING MECHANISM IN THE BELT DRIVE SYSTEM BASED ON *LADR* BASED RESIDUAL CONTROL CHART

Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto, Yasser Shaban

Submitted to:

Mechanical Systems and Signal Processing, 2022

7.1 Abstract

The belt drive system is commonly used to transmit power in different industrial systems to maintain high performance and safety. Online condition monitoring techniques (CMTs) are used to monitor the operational conditions of such systems. Vibration-based monitoring techniques (VMT) are among the CMTs that are used in the analysis and diagnosis of state of a belt drive system. Machine learning techniques are integrated with the VMT based on Industry 4.0 aspects for vibration analysis and fault diagnosis. Most of these techniques are based on the collection of vibration data from the belt drive system under known normal and different known faulty operations. This enables a fault to be diagnosed when it is detected during the operation of a system. In this paper, a new condition monitoring and warning mechanism is proposed to monitor operational conditions of a belt drive system. The mechanism is based on an integration of a Logical Analysis of Data Regression (*LADR*) with a residual Control Chart (*RCC*). It uses vibration data from the belt drive system under normal operation only. This mechanism exhibits better performance in fault detection and also in interpreting the root cause of the faults in a belt drive system. Experimental investigations on a belt drive test rig have been carried out to collect vibration data based on a design of experiment for operational factors during normal operation. The *LADR-RCC* is implemented to monitor the operation of the belt drive system and to detect the faulty state. The accuracy of *LADR* is compared with Multiple Linear Regression (*MLR*) based *RCC*, Support Vector Regression (*SVR*) based *RCC* and Random Forest (*RF*) based *RCC*. The *LADR-RCC* demonstrates significant enhancements in fault detection. The advantage of *LADR-RCC* over other model-based RCC is that it finds the root cause of a fault that is experienced in the system.

Keyword: Belt drive system, fault detection, Condition monitoring, Logical Analysis of data regression (*LADR*), Residual Control Chart (*RCC*)

7.2 Introduction

The belt drive system is extensively used as power transmission in various industrial applications, such as machine tools, conveyors, fans, and motors [186]. This system consists of a belt,bearings, driver and driven pulleys, which are connected to rotating equipment as a motor through shafts. The system is lubrication-free, has low noise operation, and has easy maintenance in addition to high-efficiency [187]. The belt is the machine element that transmits the power based on its friction with pulleys. Different types of belts, such as V-belts, timing, flat, and round belts, can be used in this system. They are classified according to the

section shape, the relative position of the shafts, the number of shafts simultaneously driven and the transmission ratio [193].

Several sources of vibration may develop different types of faults in the belt drive system, such as cut or damage in the belt, the presence of misalignment or unbalance problems, the occurrence of defects or cracks in the shafts or bearings [188–190]. When several belt drive systems are used in an industrial process, economic loss due to a loss in power transmission may be significant. Consequently, maintaining high reliability of a belt system and a mechanism for detecting of abnormal conditions during operation are necessary [194].

Online condition monitoring techniques (CMT) are implemented to measure various parameters of a system such as vibration, temperature, pressure, etc. [195]. They are used to determine the mechanical conditions of the belt drive system during operation. When a fault is developed, the CMT allows proper system analysis and diagnosis actions to take place [196]. Vibration-based monitoring techniques (VMT) are the most used techniques that continuously measure vibration signals of a belt drive system. These signals are collected from sensors attached to the system during operation. Statistical features are extracted from these signals, where their values are used as valuable diagnostic information. Generally, the values of these features represent the vibration signature of the system when the system operates under normal conditions. The vibration signature of a system is considered the characteristics of its generated vibration signals during normal operation of the system. Therefore, the values of the statistical features extracted from the current operation of the system are compared to those of the vibration signature in the system in order to analyze and diagnose the system's condition [197]. When the difference exceeds a pre-specified limit, a fault is detected [198]. Consequently, the VMT decreases the downtime of the system, facilitates its maintenance, avoids/reduces consequential failure, and reduces economic loss.

Conventional VMT has three approaches for the fault detection and diagnosis of a belt drive system: (1) Time domain approach, (2) Frequency domain approach, and (3) Time-Frequency approach. The belt drive system was experimentally investigated to analyze the system behavior under healthy and faulty conditions using the time and frequency domains [199]. Three different faulty conditions have been created on the belt which are side-cut-out, side-cut-in, and both side-cut-out and loose in the belt. A comparison was conducted to describe the belt behavior in various conditions and its influence on the system response. Different faults can be experienced in a belt drive system during operation. Therefore, a model was established for a belt drive system using *ABAQUS* software that simulated and studied the effect of three belt-drive defects, worms or cogs missing in the belt, and misalignment in the system [200]. The simulation model analyzes the vibration signals of the system in time

and frequency domains. The model was validated with experimental results, and both have approximately the same dynamic response. Moreover, the Root Mean Square (*RMS*) is a significant feature for fault detection in the system. Furthermore, another simulated model using *ABAQUS* was carried out to analyze the vibration response of the belt drive system under unsteady operation due to a misalignment problem [201]. Additional faults in the belt drive system were explored, such as unbalance and resonance in the system [202].

Recently, Industry 4.0 aspects have strongly insisted on the implementation of vibration analysis for a belt drive system-based condition monitoring using machine learning techniques. A Principal Component Analysis (*PCA)* model was developed to detect and diagnose five different types of faults experienced in a two-stage reciprocating compressor [203]. The *PCA* model was used to select the statistical features that were extracted from the vibration signal in the time domain. A combination between Wavelet packet decomposition (*WPD*) and support vector machine (*SVM*) was proposed as a diagnostic approach and a condition monitoring for the belt conveyor system [204]. The vibration signals transmitted from the belt conveyor system were decomposed using *WPD* into energy at each frequency band. Furthermore, the energy and the statistical features extracted in the time domain were trained by *SVM* to obtain a model that detected and diagnosed the faults in the system. An intelligent diagnosis system was established for fault detection of the timing belts based on the vibration signals in the three domains [205]. The extracted features for each domain were considered to be the input of an artificial neural network (*ANN*). In other words, there were three *ANN*s and the classification accuracy of each *ANN* was combined using the Dempster–Shafer theory of evidence. This led to the final fault detection and classification of the belt's defect. Furthermore, an ANN model-based time-domain vibration signal was developed to diagnose five different types of faults in the pulley belt system [206].

Most of the current VMTs used to detect the faults in the belt drive system are based on the collection of the vibration data from the system under normal and abnormal operations. Practically speaking, many experiments need to be carried out to offer appropriate training data to determine different types of faults that can be experienced in the system. In this paper, we propose an online condition monitoring and warning mechanism for fault detection and root cause identification in the belt drive system. This mechanism does not require any faulty data to monitor, analyze, and indicate the current operating condition of the system. The mechanism implements a Logical Analysis of Data Regression (*LADR*) [179] based control chart to develop a regression model based on the extracted statistical features (vibration signature) in the time and frequency domains only during the normal operation of the system. The *LADR* model is a machine learning technique that constructs a regression model based on patterns generated by the standard Logical Analysis of Data (*LAD*) methodology. These

patterns define multidimensional zones that distinguish between various groups of observations in the original training data. The residuals of the regression model are implemented in a Residual Control Chart (*RCC*). Once the residuals go beyond pre-specified limits of the *RCC*, the mechanism detects a fault in the system and sounds a warning alarm to indicate that the appropriate decisions needed to be made. Since *LADR* model is developed using patterns, these patterns are used to analyze the reason for that fault.

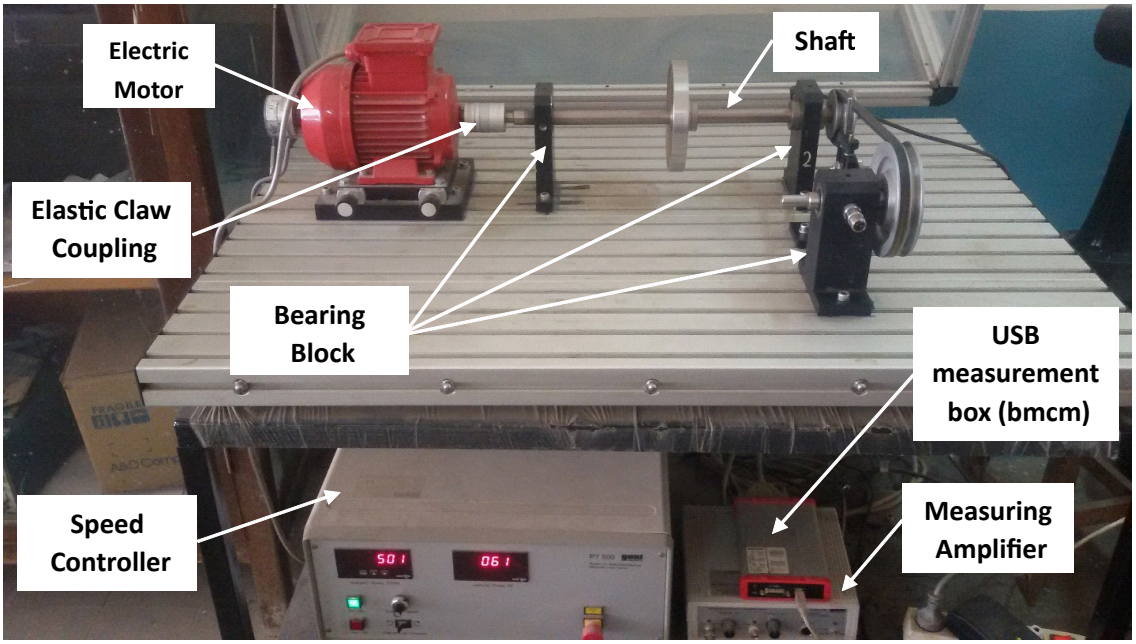
The paper is organized in the following sections. Section 7.3 presents a description and setup of the test rig experiment, in addition to data acquisition and feature extraction of the vibration signal. Section 7.4 describes in detail the proposed condition monitoring and warning mechanism based on the concept of the *LADR* regression-based Residual control chart and its implementation for to monitor and diagnose faults in the operation of a belt drive system. It discusses its strength for detecting faults and performing root cause analysis. Next, section 7.5 provides a scenario to evaluate the performance of the *LADR-RCC* compared with different integrations of regression techniques; Multiple Linear regression (*MLR*), Support Vector Regression (*SVR*), and Random Forest (*RF*), with *RCC*. Then, Section 7.6 discusses the results of the *LADR-RCC* compared to the other regression technique-based control chart. Finally, conclusions and future research are made about the proposed mechanism in Section 7.7.

7.3 Experimental Study

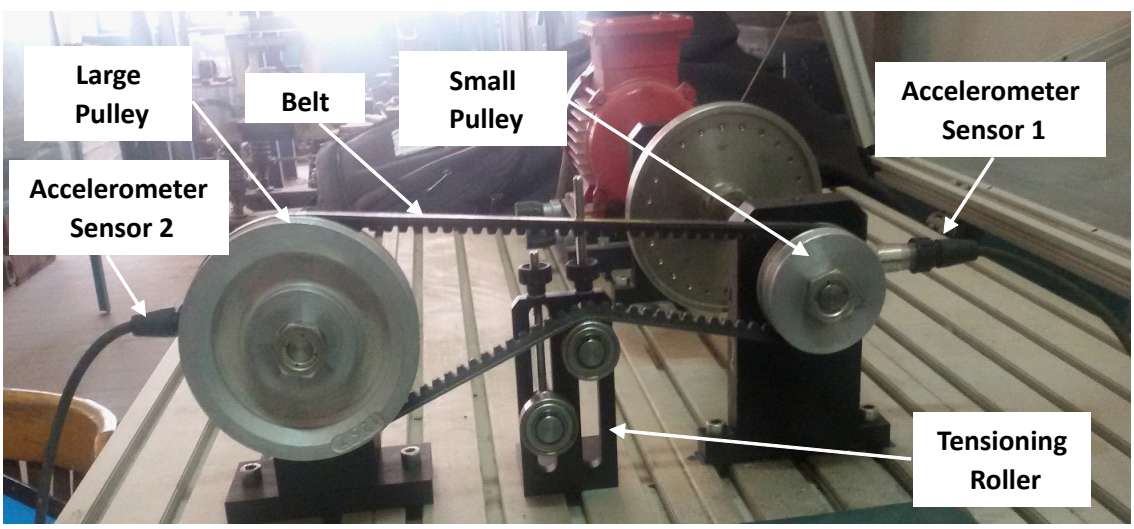
7.3.1 Experiment test-rig Description

In this paper, an experiment is performed on a belt drive system to monitor the belt condition by collecting and analyzing the vibration signals generated during normal operation. The experiments are conducted using the belt drive kit (PT 500.14) [191] of the G.U.N.T. machinery diagnostic system (PT 500) [192] as depicted in figure 7.1. G.U.N.T. (PT 500) is the base unit of the experiment. Its key components are an electric motor, two bearing blocks, and a shaft. The belt drive kit consists of a small driver pulley and a large driven pulley connected with a pre-tensioned belt using tensioning rollers. The electric motor is the drive motor with a variable speed "N" that is controlled by a speed controller. It is connected to the shaft through an elastic coupling to increase shaft flexibility and avoid its misalignment. The shaft is supported by two ball bearing blocks and connected to the small driver pulley with a 63 mm diameter. The power is transmitted by a V-belt of SPZ type with 10 mm width and 912 mm length to the large driven pulley of 125 mm diameter. The tensioning roller adjusts the pretension of the V-belt "T" by moving the roller screw with

(tensioning) or against (looseness) the belt. The "T" is measured by a pretension gauge by slowly pressing the middle of the belt until the lever of the gauge clicks at the "T" value, as shown in figure 7.2. The "T" values in the experiments are 70, 110, and 150 N.



(a)



(b)

Figure 7.1 Test-rig Description

The Data acquisition system contains two piezoelectric accelerometers (*IMI 603C01*). They are attached to the ball-bearing blocks in horizontal direction using studs to measure the vibration signals generated by the system during the experiment. Accelerometers #1 and

Figure 7.2 The pretension gauge

#2 are attached to the bearing block of the driver and driven pulley, respectively. Then, these signals are converted to electrical signals that are amplified by an amplifier. Subsequently, a USB data acquisition card (*bmcm*) digitalizes the collected signals and transfers them to a laptop through *LabVIEW* software [207] to analyze these signals.

7.3.2 Measurements and Data Description

The Design of Experiment (DOE) for collected vibration signals in the time domain from the GUNT test rig is based on the full factorial design with three controllable factors. These factors are the motor speed "N" in revolutions per minute, the unbalanced weight "W" in grams, and the pretension of the belt "T" in Newtons. Since the original experiments were carried out for different objectives, we used the "W" equals zero (absence of unbalance weight) in this paper, as experiments concerned the generated vibration signals during normal operation. The values of "N" vary from 400 to 2000 revolutions per minute by step 100, thus it has 17 levels. "T" has three levels: 70, 110, and 150 Newtons. Each experimental run is repeated three times; the total number of experimental runs is 153.

The sample value of vibration signal at time t is defined by "x_t". In this paper, we record the vibration signal in time domain using a sample size of 10000 per 100 seconds as depicted in figure 7.3a. Generally, each vibration signal at each speed "N" and pretension "T", which are measured by the accelerometers, holds information or features about the belt drive system during normal operation. These features are extracted from the signal that is measured to represent the vibration signature of normal operation. When variation of the extracted features from the signal changes significantly under the same operational conditions, "N" and "T", a fault is detected. Thus, statistical features are extracted from the vibration signals to map them from the time and frequency domains into another space in order to reveal the signal information [208]. There are ten extracted features in the time domain, as follows:

(A) Peak (Pe): The maximum absolute value of the signal.

$$Pe = max.(\mid x_t \mid) \tag{7.1}$$

(B) Range (R): Defined as the difference between the maximum and minimum values over all samples in the signal.

$$R = max.(x_t) - min.(x_t) \tag{7.2}$$

(C) Mean (M): The average value or the central tendency of all samples in the signal.

$$M = \frac{1}{k}\sum_{t=1}^{k} x_t \tag{7.3}$$

(D) Median (Me): The parameter that determines the center value of all samples in the signal at which half of the values have a larger value than the median and the other half have a lower value than the median.

$$Me = \begin{cases} x_{\frac{k}{2}} & \text{if } k \text{ is even} \\ \dfrac{x_{\frac{k-1}{2}} + x_{\frac{k+1}{2}}}{2} & \text{if } k \text{ is odd} \end{cases} \tag{7.4}$$

(E) Standard Deviation (SD): The measure of the energy content of all samples in the signal.

$$SD = \sqrt{\frac{1}{k-1}\sum_{t=1}^{k}(x_t - M)^2} \tag{7.5}$$

(F) Variance (V): The square of the "SD" and it is defined as the spread of the values over all samples in the signal.

$$V = SD^2 \tag{7.6}$$

(G) Root Mean Square (RMS): This feature represents the energy level of all samples in the signal and is the square root of the mean of the squared values in that signal.

$$RMS = \sqrt{\frac{1}{k}\sum_{t=1}^{k} x_t^2} \tag{7.7}$$

(H) Skewness (Sk): The third centered moment of all samples in the signal that determines the asymmetry of the signal distribution. It has a positive (negative) value when the signal distribution is right (left)-skewed.

$$Sk = \frac{\sum_{t=1}^{kn}(x_t - M)^3}{(k-1)SD^3} \tag{7.8}$$

(I) Kurtosis (Ku): Known as the normalized fourth-order moment. It measures the steepness of all samples in the signal distribution. When the Kurtosis value is negative, the signal distribution is flat, relative to normal distribution.

$$Ku = \frac{\sum_{t=1}^{k}(x_t - M)^4}{(k-1)SD^4} \tag{7.9}$$

(J) Crest Factor (CF): The Pe to RMS ratio of the signal.

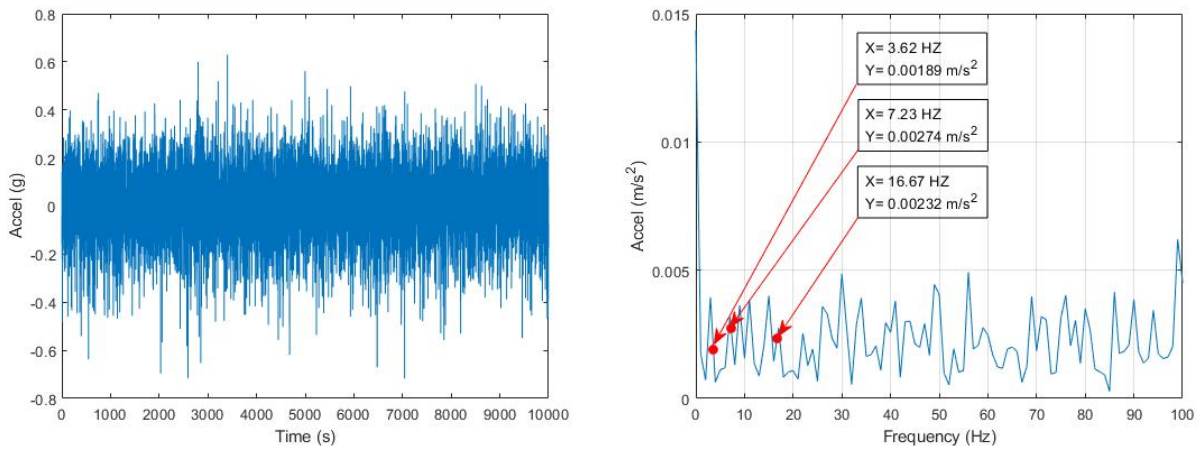
$$CF = \frac{Pe}{RMS} \tag{7.10}$$

Frequency analysis relies on converting a vibration signal into the frequency domain. Therefore, the vibration signal collected in the time domain during normal operations and at each "N" and "T" are transformed into the frequency domain or spectrum using Fast Fourier transform (FFT). Figure 7.3b shows the implementation of the FFT using *MATLAB* software. In this paper, four features are determined. The magnitude of acceleration amplitude corresponding to fundamental driver frequency of belt (at driver pulley) "f_d" at a certain drive speed "N" is obtained as a statistical feature. For example, when $N = 1000$ rpm, which is equivalent to f_d=16.67 HZ, and T=70 N in the system, the magnitude of acceleration amplitude is $0.00232 m/s^2$ as illustrated in Figure 7.3b. Similarly, the magnitude of acceleration amplitudes are obtained at belt frequencies, f_b and $f_{b'}$ (equivalent to $2f_b$) as depicted in figure 7.3b. The last feature is the magnitude of acceleration amplitude at the fundamental driven frequency of belt (at Driven pulley), f_{dn}.

Ten statistical features are extracted from the total number of experimental runs: peak, range, mean, median, standard deviation, variance, root mean square, skewness, kurtosis, and crest factor. For each condition N and T, we take the average values of each feature. Consequently, a dataset is constructed that contains 51 observations and 10 extracted features in addition to the fundamental belt frequencies for normal operation of the belt drive system, as depicted in table 7.1.

7.4 Methodology

In this section, the methodology of the new condition monitoring and warning mechanism is introduced as an enhancement in system monitoring and fault detection and diagnosis.



(a) Time domain (b) Frequency domain

Figure 7.3 Time and Frequency domains for the normal operation of belt drive system at N=1000 RPM and T=70 N

The mechanism is based on the integration of *LADR-RCC* to monitor the condition of the belt drive system through the vibration signals generated to state the condition of the belt. The *LADR-RCC* is a model-based control chart [209]. It is a combination of the *LADR* technique [179] and residual control chart [1] that is used to monitor the operation of a system. The *LADR* develops a regression model based on the generated patterns from the original data as independent variables. The *LADR* develops a regression model based on the generated patterns from the original data as independent variables. Each pattern defines a multidimensional zone of the features' values that distinguishes between various groups of observations in the original training data. These features provide indications of the system's state. The *LADR* model describes the relationship between these patterns and the response variable. The *LADR* model acts as a regression adjustment, in which the residuals ,E, are monitored by the *RCC* to detect any fault in the system during operation. The residual term is defined as the difference between the actual value of a system's response and the value that is predicted by the model. Once the value of the "E" goes beyond the *RCC* limits, a fault is detected in the system.

7.4.1 LADR regression technique

The cornerstone of the *LADR* technique is the standard Logical Analysis of Data (*LAD*) methodology. *LAD* is non-statistical data mining technique to generate patterns that is based on Boolean logic and combinatorial optimization [210, 211]. It is a supervised learning

Table 7.1 Samples of the data used

N	T	Pe_1	Pe_2	R_1	R_2	M_1	M_2	Me_1	Me_2	RMS_1	RMS_2
400	70	0.1806	0.1944	0.3144	0.3133	0.0173	0.0045	0.0175	0.0047	0.0352	0.0286
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	70	1.4097	4.8131	2.7369	6.9883	0.0134	-0.0057	-0.0031	0.0013	0.3654	0.5769
400	110	0.292	0.218	0.501	0.431	0.017	0.004	0.016	0.003	0.043	0.041
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	110	1.769	2.742	3.361	5.156	0.018	0.004	0.014	0.006	0.374	0.635
400	150	0.217	0.427	0.372	0.738	0.017	0.004	0.016	0.004	0.033	0.035
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	150	2.143	2.998	3.728	5.661	0.011	0.004	-0.002	0.004	0.402	0.619

N	T	SD_1	SD_2	V_1	V_2	Sk_1	Sk_2	Ku_1	Ku_2	CF_1	CF_2
400	70	0.0307	0.0283	0.0009	0.0008	-0.1465	-0.1854	3.6835	3.8187	5.1288	6.7846
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	70	0.3651	0.5769	0.1333	0.3328	0.1465	-0.0981	2.9396	3.5181	3.8585	8.3429
400	110	0.040	0.041	0.002	0.002	0.401	0.092	6.123	3.908	6.716	5.263
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	110	0.374	0.635	0.140	0.403	-0.019	-0.028	3.432	3.065	4.723	4.319
400	150	0.028	0.034	0.001	0.001	-0.087	-0.361	3.956	10.07	6.603	12.35
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2000	150	0.402	0.619	0.161	0.383	0.209	0.016	3.383	3.239	5.336	4.847

N	T	f_b	$f_{b'}$	f_{dn}	f_d
400	70	0.0003	0.0005	0.0002	0.0006
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2000	70	0.0061	0.0032	0.0055	0.0075
400	110	0.0011	0.0010	0.0005	0.0004
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2000	110	0.0037	0.0042	0.0108	0.0070
400	150	0.0004	0.0006	0.0007	0.0006
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2000	150	0.0097	0.0053	0.0115	0.0077

technique that is originally applied to classification problems [132]. More detail about the LAD methodology are found in [141, 212]. The $LADR$ technique transforms the standard LAD methodology to solve regression problems. It uses LAD to extract the hidden patterns, P_i , j=1,. . . ,J, where j is the identifier of the pattern. The patterns cover zones of values in the dataset, such as the union of zones cover the totality of the dataset. The $LADR$ technique maps the features' values in Q-dimensional $\{0, 1\}^Q$, where the Q is features covering pattern. It builds a model for response variable Y, and J independent binary variables X_{P_j} , j=1,. . . ,J.

The X_{P_j} equals 1 when an observation "ω" is covered by the pattern P_j, otherwise it equals 0. The *LADR* model is applied to both linear and non-linear regressions between response and independent variables although the relationship between the response variables and the patterns' variables is linear, as shown in equation (7.11).

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} (\beta_j X_{P_j}) \tag{7.11}$$

Considering a dataset Ω of n-independent variables (X_1,...,X_n), one dependent variable (Y), and m-observations (ω=1,...,m). The *LADR* technique is implemented in four main steps: response classification, pattern generation, data preparation and processing, and regression modeling and validation [179] as depicted in Algorithm 2.

The *LADR* technique starts by sorting the Ω based on the Y-values in ascending order. Subsequently, it classifies the Y-values into $\mathcal{N}$-classes. I-thresholds; τ_1,...,τ_I where I=$\mathcal{N}$-1; are derived from these classes. The τ_i threshold where i = 1,..., I is defined by the starting value of $(i+1)^{th}$ class [179]. In this paper, we use two different classification methods: Equal width intervals (EW) and K-means clustering (KM). They classify the Y-values into $\mathcal{N}$-classes based on having the same width, and by using the K-mean technique, respectively [143]. For each τ_i, the Ω=$\{\Omega_i^+, \Omega_i^-\}$ where $\Omega_i^+$is called the positive set of observations, for which the Y-values are greater than or equal to the τ_i-value. While Ω_i^- is called the negative set of observations, for which the Y-values are less than the τ_i-value. This is considered a two-class classification problem for the *LAD* methodology. It is implemented using a *cbmLAD* software [24] to extract hidden patterns at τ_i (P_i) that distinguish between these two classes. The generated patterns P_i =$\{P_i^+\} \cup \{P_i^-\}$, where P_i^+ (P_i^-) is the positive (negative) patterns that cover at least one observation of the Ω_i^+ (Ω_i^-) at the τ_i. We repeat the same procedures for all of I-thresholds and gather all of the generated patterns in a single dataset $\mathcal{P}$=$\cup P_i$. Therefore, $\mathcal{P}$ contains J patterns that are included in P_i at all of the thresholds.Each pattern P_j in $\mathcal{P}$ is considered as a binary independent variable X_{P_j} where j=1,...,J which indicates that the pattern j exists. When an observation in the Ω is covered by P_j, the X_{P_j} is equal to 1, and 0 otherwise.

The prevalence of each pattern P_j is calculated. It is defined as the percentage of observations in the original data that are covered by each P_j during the training phase. When a P_j has a higher prevalence in a class, it is considered more significant. As such, the predicted value of Y by the developed *LADR* model is likely to be within the range of Y-values in the class. On the contrary, when the P_j has a lower prevalence in a class, it is less significant, and the prediction may be outside the range of Y-values of that class.

Algorithm 2 *LADR* Technique

```
Read (Ω)
Sort (Ω)                                          ▷ Ascending order based on Y
for 𝒩 ← 2 to 20 do                                ▷ 𝒩-Classes
    τ= Discretize (Ω)                             ▷ Using the EW or KM method
    I=Length (τ)                                  ▷ I=𝒩-1
    𝒫=[ ], MSE=[ ], R²=[ ]
    for i ← 1 to I do
        for ω ← 1 to m do
            Ω_i^+={Ω_ω ∈ Ω | Y(ω) ≥ τ_i}
            Ω_i^-={Ω_ω ∈ Ω | Y(ω) < τ_i}
        end for
        P_i= cbmLAD (Ω_i^+,Ω_i^-)                 ▷ Generate pattern at τ_i
        𝒫=𝒫 ∪P_i
    end for
    X_P=[ ]
    if Cover(𝒫_j, ω) then X_{P_j}=1     ∀ j=1,..., Length(𝒫)
    else
        X_{P_j}=0
    end if
    Ω′=(X_{P_j}, Y)
    Ω′=Data processing (Ω′)
    Model= Linear regression (X_{P_j}, Y)
    MSE_i= MSE(Model) & R_i^2= R²(Model)          ▷ Using 10-cross validation for 10 times
end for
min(MSE), max (R²), & optimum-classes.
```

7.4.2 Residual Control Chart

Residual Control Chart (RCC) is a statistical tool that graphically monitors the residual terms associated with the regression model [213–215]. In RCC, a regression technique is used to fit the in-control data of the process to develop a regression model. The residuals "E_t" obtained from that model at each time "t" are implemented in the RCC as in equation (7.12).

$$E_t = Y_t - \hat{Y}_t \tag{7.12}$$

Where Y_t is the actual response at time t, and $\hat{Y}_t$ is the response predicted by the regression model at time t.

The merit is that the residuals from the regression model are typically uncorrelated even though autocorrelation is presented in the original data [216]. E_t is independently and normally distributed (iid) where $E_t \sim N(0,\sigma_E^2)$. In this paper, we use the $LADR$ technique to develop the regression model of Y and X_{P_j} as discussed in the previous subsection.

In order to determine the RCC parameters, the data or vibration signals are collected during the normal operation of the system. The control limits of the RCC are calculated as the following:

$$UCL = \mu_E + 3\sigma_E \tag{7.13}$$

$$LCL = \mu_E - 3\sigma_E \tag{7.14}$$

Where the μ_E is residuals' mean and the centerline of the RCC, which equals 0, and σ_E is the standard deviation of the residuals obtained by the model in normal operation of the system.

Thus, the E_t are monitored instead of the original observations (Y_t) in the control charts. Consequently, the operating condition of the system is identified to be in-control when E_t is within the control limits, otherwise, it is out-of-control. Once the value of E_t goes beyond the RCC limits, a fault is detected. The model is analyzed at that out-of-control observation based on two terms: 1) whether the $E_t > UCL$ or $E_t < LCL$, 2) the prevalence of the P_j. By using RCC, we determine which class the actual Y-value of the out-of-control observation is in. At that observation, the $LADR$ model consists of X_{P_j} with positive and negative coefficients. When $E_t < LCL$, E_t has a negative value where the predicted value obtained by the model is greater than the actual dependent variable. Consequently, the covering X_{P_j} that

have the positive coefficients with the lowest prevalence in that observation's class contribute to that fault. Conversely, when $E_t > UCL$, E_t has a positive value where the predicted value obtained by the model is lower than the actual dependent variable. Consequently, the covering X_{P_j} that have the negative coefficients with the lowest prevalence in that observation's class contribute to that fault. For special cases in which the *LADR* model has X_{P_j} of positive coefficient(s) only and $E_t > UCL$ at an observation, the X_{P_j} that have positive coefficients with higher prevalence in that observation's class contribute to that fault and vice versa.

7.4.3 Condition Monitoring and Warning Mechanism

The condition and warning mechanism is proposed to monitor the operational conditions of the belt drive system. The mechanism is based on the integration of the Logical Analysis of Data Regression (*LADR*) with the Residual Control Chart (*RCC*). The structure of the proposed mechanism is as depicted in figure 7.4. It is divided into two-stages, offline training and online monitoring.

The vibration signals generated from the experiments that were carried out on the belt drive system during normal operation using a healthy belt at different speeds "N" and pretension "T" are considered to be the historical data. In this paper, we use the signals that are measured from the accelerometers attached to the driver pulley. In the offline stage, the statistical features of the time and frequency domains are extracted from the historical data, as previously mentioned in section 7.3.2. The *LADR* technique acts as a regression adjustment to develop a regression model. The independent variables of that model are the controllable input variables, speed "N" and pretension "T", and the time domain features. The f_d is the dependent variable that is extracted from the frequency domain via the FFT method in *MATLAB*. By training the *LADR* model during normal operation, the model's residuals "E_t" represent the difference between the measured f_{d_t} and the predicted $\hat{f}_{d_t}$. Consequently, the mean and standard deviation of the model's residuals are obtained. Moreover, the control limits are calculated based on equations (7.13) and (7.14). Accordingly, the *RCC* is constructed to monitor the residual, detect, and diagnose any fault experiences in the system during operation.

Considering the belt drive system operates at certain values of "N" and "T", in the online stage of monitoring the system, the vibration signals are collected by the accelerometer #1 attached to the driver pulley and transferred to the laptop as was previously mentioned in section 3.1. The mechanism analyzes the signal and extracts the statistical features in the time and frequency domains. Subsequently, the *LADR* model is used to predict the value of the $\hat{f}_{d_t}$. Then, the residual E_t is implemented on *RCC*. When the E_t lies within the control

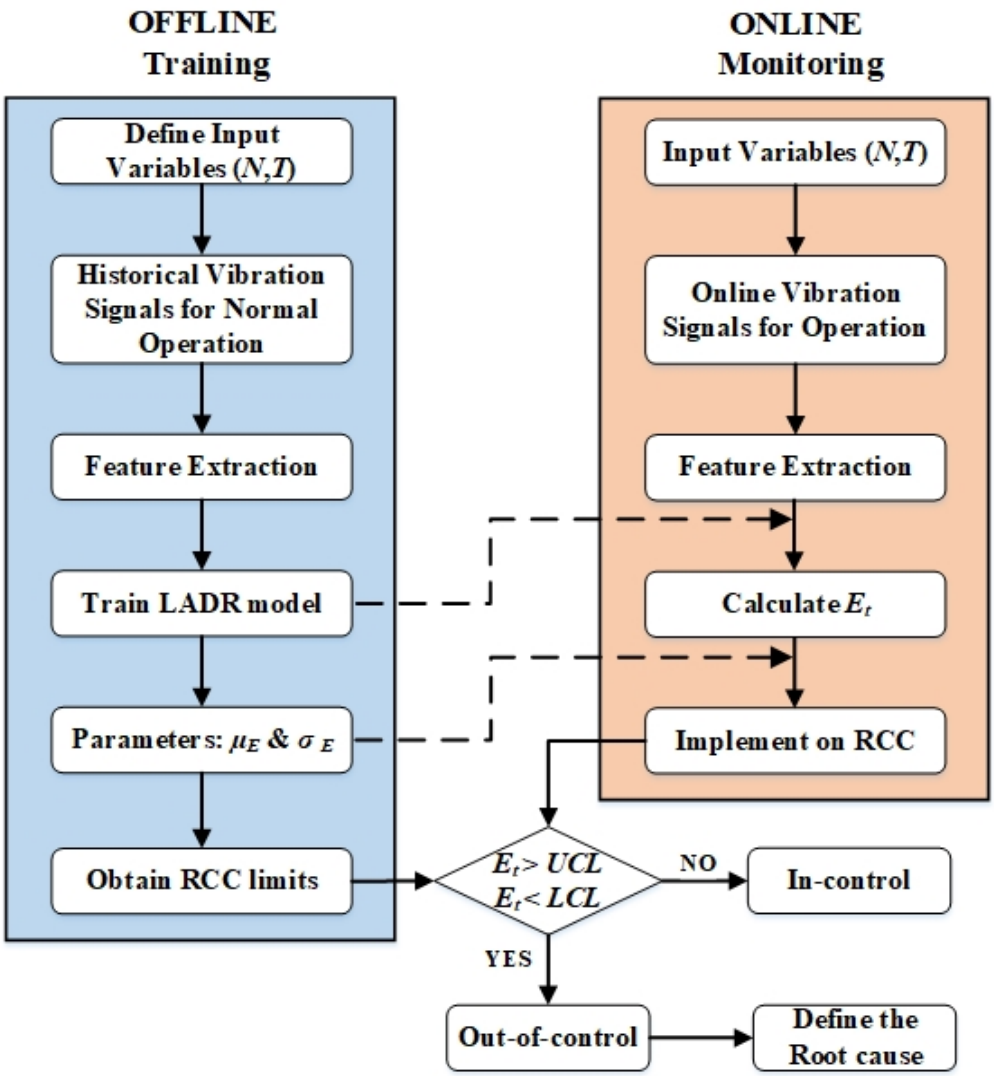


Figure 7.4 Schematic of proposed condition monitoring and warning Mechanism

limits, the system operates normally and has an in-control status. A fault is detected when E_t goes beyond the control limits and the mechanism provides a warning and the root cause of the fault is determined. Therefore, this allows an appropriate decision to be made to reduce damage caused by this fault.

7.5 Experimental case study to evaluate the mechanism

An experiment scenario was conducted to assess the performance of the proposed mechanism using *LADR-RCC* to detect and diagnose a fault that occurs during the operation of the belt drive system. The experiment was that the system operated at a constant speed, N=1350

RPM, and a healthy belt was pre-tensioned with T= 70 N. The vibration signals are acquired by accelerometer#1 and recorded for 100 seconds by the data acquisition system at a sampling rate of 100 per second each time. Thus, the sample size of one sample is 10000. The measurements are repeated every 200 seconds. We recorded 144 samples, and the statistical features in the time and frequency domains were extracted for each sample. The system was operating normally until sample 128, after which a persistent abnormal behavior was noticed in the system due to a cut or damage on the belt cog.

The *LADR* was already trained by the historical data that was generated in section 7.4.1. Then, *LADR* was integrated with the *RCC* to monitor the system online. Furthermore, we implemented three different regression techniques that are integrated with the *RCC* in order to compare their performance with the proposed approach. These techniques are described in the following sections.

7.5.1 Multiple linear regression (MLR)

The *MLR* is a parametric and supervised learning approach that describes the linear relationship between the independent and dependent variables [217]. It uses least square estimation to find out the coefficients of the model [218]. It is used widely due to its simplicity. Nevertheless, four assumptions must be associated with the *MLR* to obtain an appropriate model: 1) the linearity relationship between the independent and dependent variables, 2) The homoscedasticity where the residuals' variance is constant, 3) The residuals obtained by the model are independent and do not correlate, and 4) The normality of the residuals [219].

7.5.2 Support Vector Regression (SVR)

The *SVR* is based on the principle of the Support Vector Machine (*SVM*) approach that can be adapted to regression problems [162]. The main idea of SVR is to obtain a loss function *f(x)* that can predict the dependent variables with ϵ-error during the training stage [220]. The *f(x)* should be as flat as possible [221]. Unlike *MLR*, the *SVR* can handle the non-linearity relationship using different kernel functions such as polynomial, radial,..., and so on. In this paper, we use the radial kernel function using *"E1071"* in the *RStudio* package [154] to map the space of the input feature to a new space. More details about the algorithm are available in [175].

7.5.3 Random Forest Regression (RF)

The *RF* is considered a supervised ensemble learning approach that is used for developing a regression model. It combines many decision trees, n_{tree}, to more accurately predict the dependent variable. The n_{tree} is determined based on the sample sets that are drawn from the training data. Each tree is split into several nodes that represent the statistical features extracted from the vibration signal of the system. For each sample set, m_{try} of the independent variables are selected randomly, and accordingly, the best split is defined from them [222]. The prediction of *RF* is the average from the aggregating the prediction for the whole the n_{tree}, as illustrated in figure 7.5 [223]. We implement the *RF* approach using the *randomForest* package in *RStudio* software. See details about the approach in [222].

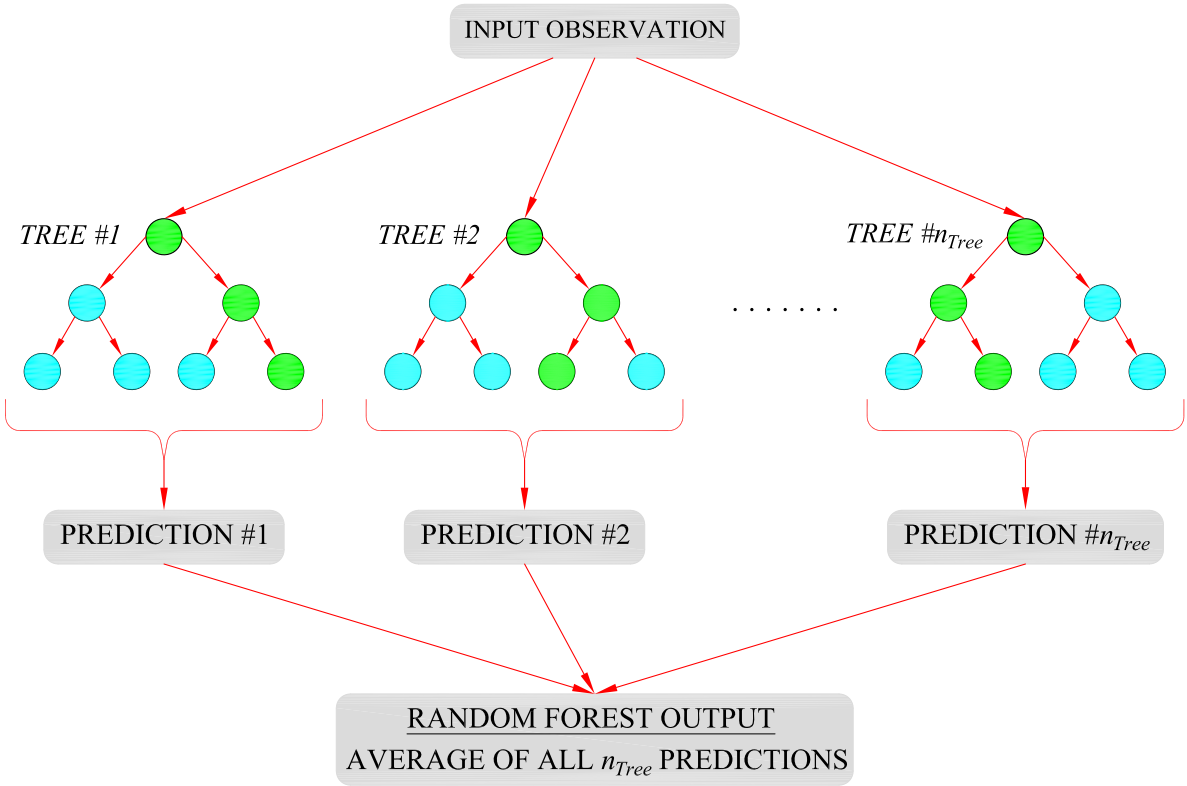


Figure 7.5 Random Forest Regression model

7.6 Results and Discussions

Referring to the original experiments, we develop a *LADR* model for the Y-variable as a function of patterns instead of the original independent variables during the training stage. We use the integration of this model with *RCC* to monitor the operating conditions of the belt drive system in the scenario in section 5. In addition, we compared three different model-based control charts (*MLR-RCC*, *SVR-RCC*, and *RF-RCC*) with the proposed approach. The comparison is carried out based on the performance of the regression model and fault detection during a system's operation.

The comparison between the performance of the regression approaches is based on two metric terms: the MSE and R^2 using 5 cross-validation for 10 replications. Table 7.2 presents the comparative results of the different approaches.

Table 7.2 The performance of regression models

Approach	Threshold	MSE*10^{-4}	R^2
LADR-KM	4	0.003	93.288
LADR-EW	3	0.004	91.762
MLR	-	0.013	71.618
SVR	-	0.012	76.154
RF	-	0.008	81.507

The *LADR* models, *LADR-EW* and *LADR-KM*, using the two classification methods *EW* and *KM*, respectively have better performance compared to the other approaches. Moreover, the *LADR-KM* is the best-fitted model that has the lowest MSE and the highest R^2. The *RF* model produces better results compared to the *SVR* and *MLR* and is competitive with the *LADR-KM*. The *MLR* has the lowest performance. Although the *LADR-KM* model is obtained using the multiple linear regression approach, it provides higher performance compared with the *MLR*. The performance of *RF* relies on determining n_{Tree} and m_{Try}. Using 5 cross-validation for 10 replications, the optimum $m_{Try} = 10$ where ten independent variables are randomly selected to be candidates at each split. The n_{Tree} is 500 where the average of their predictions is the prediction of the *RF*.

The *LADR-KM* model partitions the data into 4 classes, which are equivalent to 3 thresholds as shown in table 7.2, $C_0 : C_3$.The structure of the model is in equation (7.15). Table 7.3 describes the patterns that are significant independent variables in the model, as in equation (7.15). Furthermore, figure 7.6 indicates the prevalence of each pattern in each class.

$$\hat{f}_d = 0.0028 - 0.0009X_{P_1} - 0.0003X_{P_2} + 0.0006X_{P_5} - 0.0009X_{P_7} + 0.0003 X_{P_9} + 0.0010X_{P_{12}} + 0.0013X_{P_{13}} + 0.0009X_{P_{14}} + 0.0015X_{P_{17}} \tag{7.15}$$

Table 7.3 The Patterns of the *LADR-KM* model

P_j	Pattern Description	P_j	Pattern Description
P_1	RMS_1<0.1435 f_b>0.000345 f_b<0.00204	P_{12}	Sk_1>-0.139 Ku_1<3.768 f_b>0.004225 $f_{b'}$>0.003355 f_{dn}>0.00144 f_{dn}>0.00485
P_2	RMS_1<0.1435 Sk_1>-0.1535 f_b>0.000345 $f_{b'}$<0.002325	P_{13}	Sk_1>-0.1075 Sk_1>-0.089 Ku_1>3.8795 $f_{b'}$>0.001255 f_{dn}>0.00144
P_5	RMS_1>0.1285 $f_{b'}$>0.001015 f_{dn}>0.001285	P_{14}	RMS_1>0.317
P_7	f_b>0.000345 $f_{b'}$<0.003355	P_{17}	N>1750 Ku_1<3.5025
P_9	T>90 RMS_1<0.317 Sk_1<0.276 $f_{b'}$<0.006985 f_{dn}<0.007625		

We integrate the *LADR* model as a regression adjustment method with the *RCC* to monitor the operating conditions for the belt drive system. Then, we compare the performance of the proposed *LADR-RCC* with the other regression models based *RCC* in fault detection. The mean and standard deviation of the residuals are determined for each regression model based *RCC* during the training stage of the normal system's operation. Therefore, the E_t of the *MLR*, *SVR*, *RF*, and *LADR-KM* models follow $N(0,0.00160^2)$, $N(0,0.00153^2)$, $N(0,0.00145^2)$, and $N(0,0.0014^2)$, respectively. Subsequently, the control limits of *RCC*, *UCL* and *LCL*, are constructed as in equations (7.13) and (7.14). The regression models predict the f_d each time for the scenario mentioned in section 7.5. Afterward, the obtained residuals are implemented in the *RCC* to detect any fault in the system during its operation. The results demonstrate

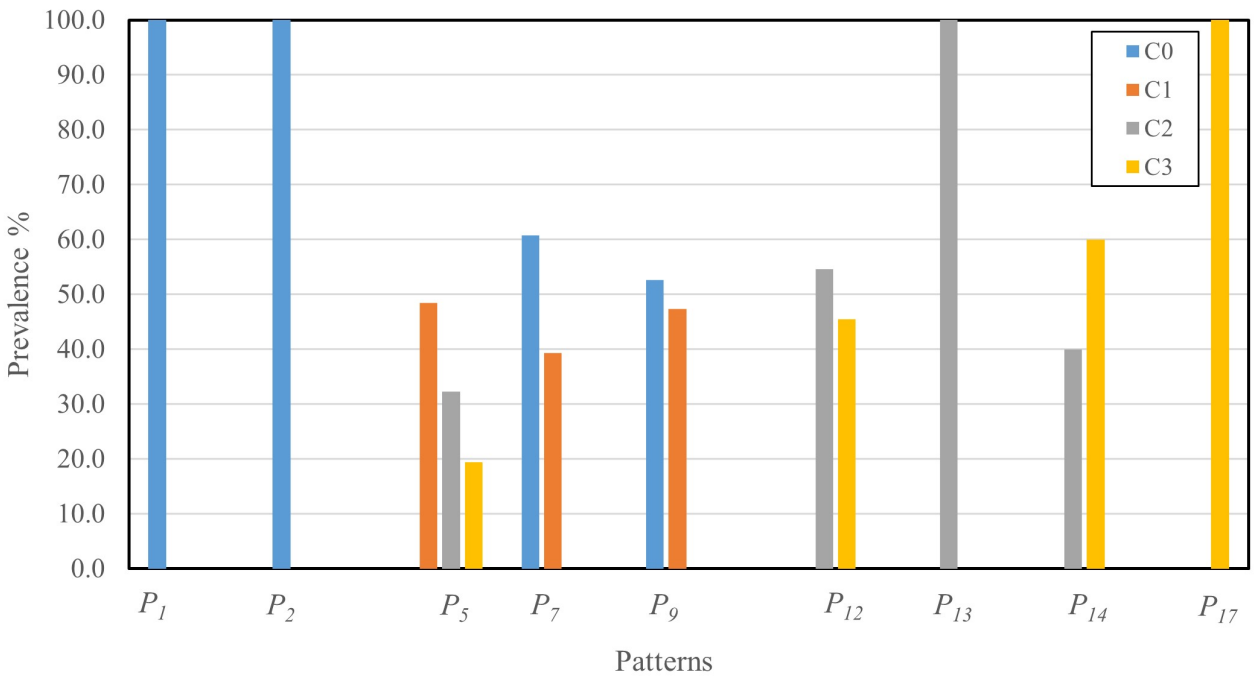


Figure 7.6 The pattern's prevalence for each class

that both *LADR-KM-RCC* and *SVR-RCC* perform the best in monitoring and detecting a fault in the system, as depicted in figures 7.7 and 7.8. Several observations are greater than the *UCL*, which represent out-of-control conditions. Both approaches provide alarms at the 130th sample, which is the first out-of-control point. There are more missed detection samples in the *MLR-RCC* where it detects the fault at the 134th sample, as depicted in figure 7.9. Moreover, the *RF-RCC* in figure 7.10 does not realize that a fault is experienced in the system where no observation goes beyond the *RCC* limits. Nevertheless, many consecutive points are located on one side of the *RCC*, which is considered abnormal behavior.

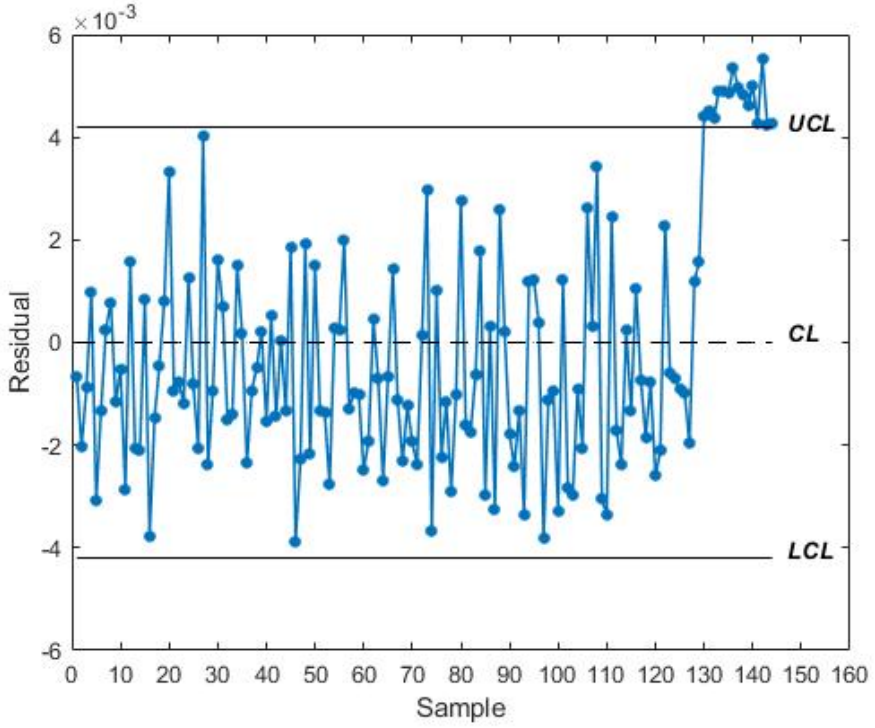


Figure 7.7 The *LADR-KM-RCC* for belt drive system

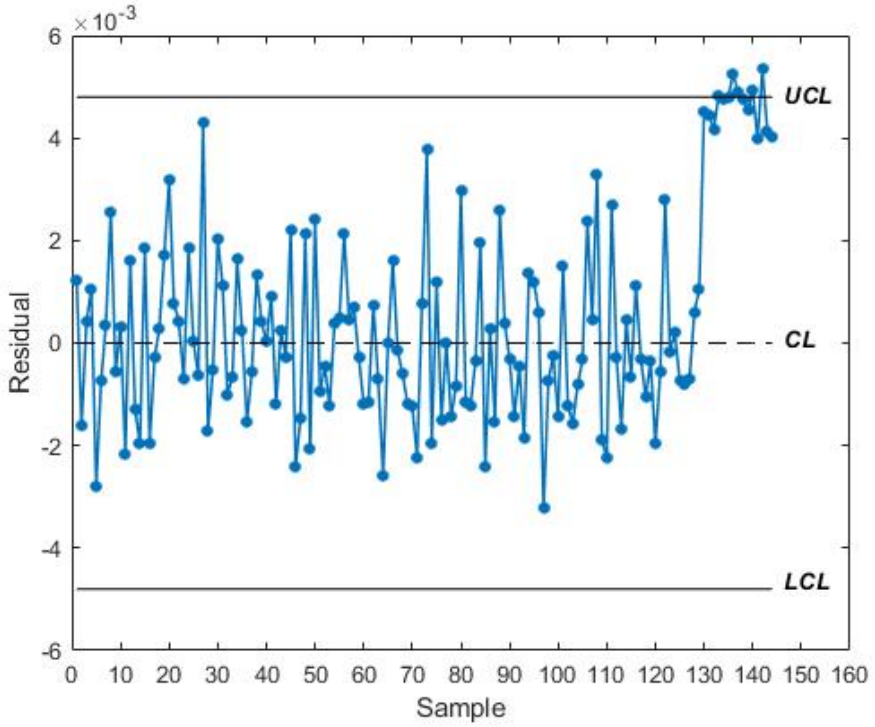


Figure 7.9 The *MLR-RCC* for belt drive system

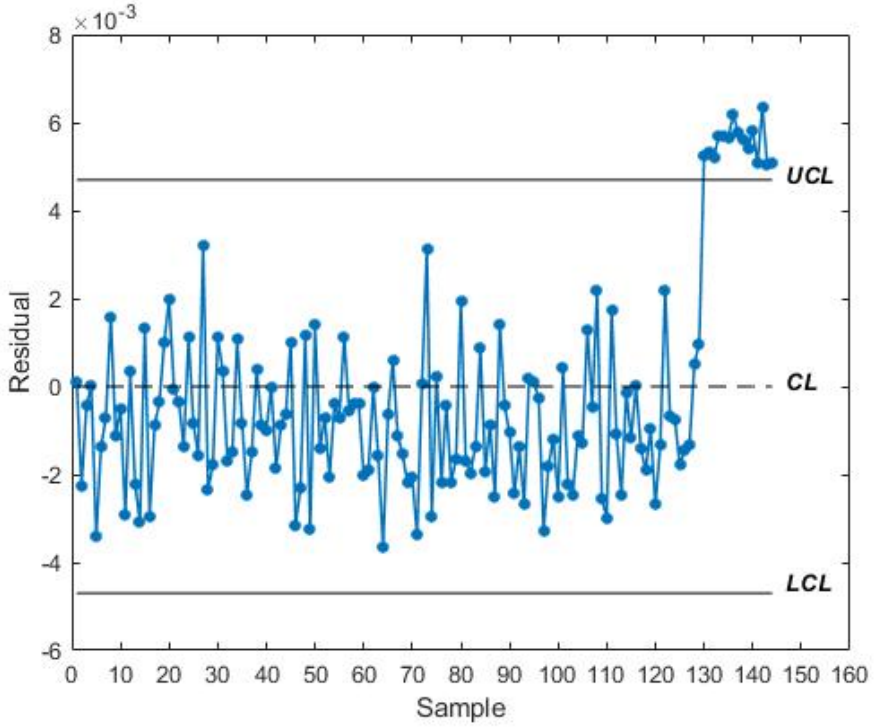


Figure 7.8 The *SVR-RCC* for belt drive system

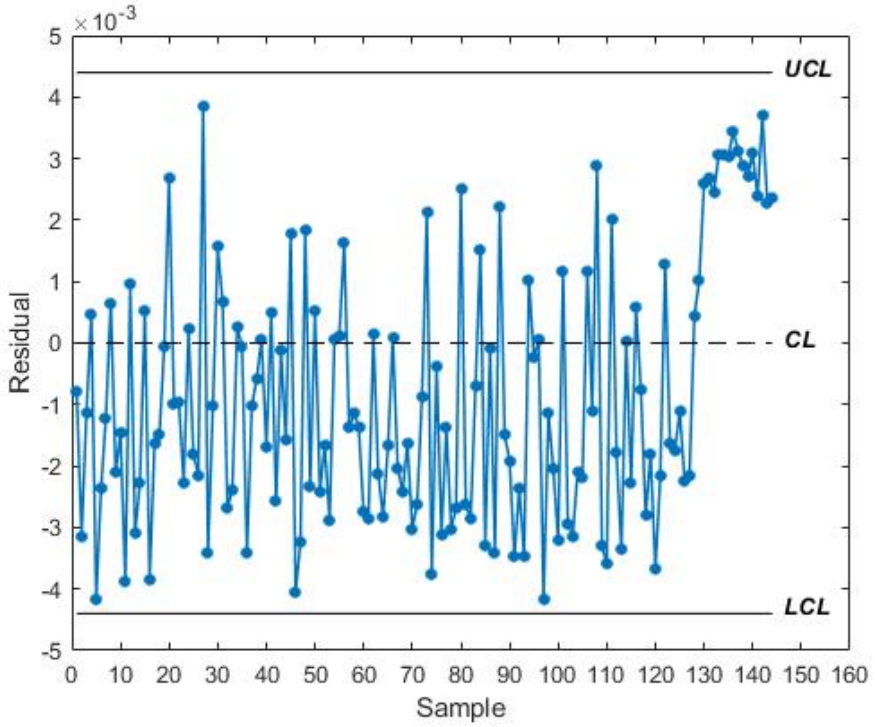


Figure 7.10 The *RF-RCC* for belt drive system

The strength of the *LADR-KM* model is developed using the generated patterns that determine the root cause of the out-of-control observation. Conversely, the other regression models require an additional tool to interpret the reason for the alarm. Furthermore, it is necessary to generate training data randomly or to carry out many experiments to offer enough data that describes the operating condition of the belt drive system under the presence of different types of faults that can be experienced in the system. Once an observation is detected as a fault, a warning signal is alerted and that observation is investigated. The 130th observation goes above the *UCL* in which the covered patterns in the *LADR-KM* and the uncovered patterns are eliminated, as shown in equation (7.16).

$$\hat{f}_d = 0.00280.0006X_{P_5} + 0.0009X_{P_{14}} \tag{7.16}$$

The actual value of the 130th sample belongs to class C_3 while the one predicted by the *LADR-KM* belongs to class C_2. Since the $E_{130}>UCL$, the reason for the faulty observation is the presence of a X_{P_j} with negative coefficient and lower prevalence in class C_3. Nevertheless, the *LADR-KM* contains X_{P_j} with positive coefficients only. Referring to section 7.4, the X_{P_j} with positive coefficient and higher prevalence in that class is the root cause. Therefore, it can be concluded that the root cause of the warning alarm is the pattern $X_{P_{14}}$. As shown in table 7.3, the $X_{P_{14}}$ is RMS_1>0.317. The current signal has a higher *RMS* value of the belt at the driving pulley where higher vibration energy is induced in the system. This indicates peak amplitudes appear in the signal during current operating conditions. It can be noted that the peak of this signal is much greater than in normal conditions. This indicates fatigue or damage to the belt of the system and this was confirmed when the belt of the system was investigated.

7.7 Conclusion

Fault detection has a great impact on the safety and reliability of complex manufacturing and industrial operations. A new online condition monitoring and warning mechanism has been proposed in this paper, which adopts the integration of the *LADR* technique and the *RCC*. The new mechanism monitors the manufacturing system to detect any abnormal behavior during operation. Furthermore, it interprets the root cause of that behavior to make an appropriate decision and to avoid any economic loss. The *LADR* technique extends the role of the standard *LAD* methodology to develop a regression model based on extracted hidden patterns in the original data. The *RCC* is used to monitor the residuals obtained from that model to detect any fault in the system. We use the proposed mechanism to monitor the

belt drive system, which is used widely in different industrial applications. A DOE scheme is designed to apply extensive experiments with different levels of controllable variables in the system: speed, pre-tension of the belt, and unbalance loading. Thus, the statistical features of the time and frequency domains are extracted from the vibration signals of the system during normal operation.

The proposed mechanism is used to obtain the *LADR* model and the parameters that are used to construct the *RCC* during the offline training stage. *LADR* is considered a regression adjustment to describe the relationship between the f_d as a function of the controllable variables, the motor speed and tension of the belt, and the extracted features from the signals during normal operation. During the online stage, the *RCC* in the mechanism monitors the difference between the actual f_d and predicted value during operation. When the residuals go beyond the *RCC* limits, a fault is detected. Subsequently, the *LADR* model interprets the reason for that fault to take corrective actions and to return the system to normal condition.

A comparison is carried out between the proposed mechanism and the other model-based *RCC* in terms of performance of the developed model and fault detection. Not only does the mechanism exhibit better performance in fault detection, but the interpretation of the root cause of the belt drive system as well. However, the *SVR-RCC* is competitive to the *LADR-RCC*, but a classifier tool in addition to enough faulty data is required to interpret the reason for the detected fault. Generally speaking, the proposed mechanism contributes to improvements in the performance of fault detection, and it is considered a robust and reliable approach for online industrial applications.

Our future research will implement our proposed technique to detect and diagnose other types of faults in the belt drive system, such as bearing defects, bent and cracked shaft, and eccentric pulleys.

CHAPTER 8 GENERAL DISCUSSION

This thesis provides data management and analytics that serve process monitoring and quality control in the industrial field. Most of the data that is exchanged and collected through the sensors and controllers within a process, or a system, includes diverse and heterogeneous data. Different data structures with different formats are captured with low quality in terms of data redundancy, incompleteness, and inconsistency. Therefore, the stakeholders take a long time to get their required data top prepare their reports and further analyses. In thesis, we focus on managing the data within a process to conduct the key performance indices for all stakeholders, especially the quality team. On the other hand, the control charts have several limitations to monitor the process variability to make sure that the process and/or the product is in-control. High rates of false alarms and missed detection are provided, which lead to economic losses. Moreover, control charts are not designed to diagnose or identify an anomaly when it is detected in a process.

Based on the above mentioned problems and challenges, these were the motivation to develop a research roadmap that is compatible with the aspects of the Quality 4.0 paradigm. Data modeling using Entity-Relationship modeling (*ERM*) has been developed to manage the data within a process in a database with predefined structures. Then, an accurate machine learning technique based on pattern recognition is called Logical Analysis of Data Regression (*LADR*). *LADR* is used to obtain a regression model that describes the key performance indices of process quality. *LADR* has been integrated with conventional control charts as a new model-based control chart. The new integration is used for both anomaly detection and identification, unlike the other well-known machine learning techniques.

The *ERM* provides structured and high-quality data that is visually represented and easily understood by the business users. Each stakeholder receives the required information and data in a predefined structure that facilitates preparing reports. Business organization is capable of data exploitation for further analysis using Artificial Intelligence/Machine learning techniques. This leads to new insights that improve the productivity performance in the future. Furthermore, the stored data in the database represents the base of creation of digital twin.

The *LADR* is considered as an extension of standard LAD methodology to be implemented for regression problems. The predicted response of the *LADR* is a function in binary independent variables. The independent variables represent the extracted patterns from the original dataset. We proposed a clear methodology for implementation of *LADR* using three

new classification methods: equal width (EW), K-means Clustering (KM), and % standard deviation ($\%STD$). The LADR provides better and significant results compared to the other well-known regression techniques.

The $LADR$ is integrated with the control charts to solve the drawbacks of the existed model-based control. Since the $LADR$ constructs a regression model of high performance, this increases the accuracy of anomaly detection. Unlike other integrations, $LADR$ does not require sufficient training data that describes different anomalies. Since the LADR model is based on extracted patterns, we exploit these patterns to identify the reasons for the detected anomaly. The results of the proposed integration show a reduction in false alarm and missed detection rates. Generally, the proposed integration is used for anomaly detection and diagnosis of the root cause of the anomaly. It represents a reliable and robust approach for online-monitoring different applications.

A condition monitoring and warning mechanism has been developed based on integrating the $LADR$-residual control chart to monitor the operation of the belt drive system. It is considered a new vibration-based monitoring technique. It monitors the statistical features that are extracted from the vibration signals of the system. When these features deviate significantly from those defined in the normal operation, a fault is detected. Not only, the mechanism provides better results for detecting any faults that are experienced in the system but also for identifying and interpreting the root cause of that detected fault.

CHAPTER 9 CONCLUSION AND RECOMMENDATIONS

9.1 Summary of Works

In this thesis, the entity-relationship approach is used to develop a data model that characterizes, identifies, and stores the data in a pre-defined database structure. This database defines all of the required data and information required by stakeholders. Consequently, it is used to define key performance indices for the quality of the process data. The data stored in the database will be considered historical data, which is used for process quality monitoring in control charts.

Several limitations have appeared and have grown profusely because of an increase in the complexity of manufacturing processes. Companies strive towards integrating machine learning techniques with conventional quality tools according to Quality 4.0 aspects. A Logical Analysis of Data regression (*LADR*) technique has been developed based on generated patterns using a standard *LAD* methodology. It constructs a regression model that describes process performance in terms of the variables of the historical data that are stored and managed in the designed database.

Furthermore, the *LADR* technique is integrated with control charts to improve the sensitivity of anomaly detection. Since the *LADR* model is constructed based on interpretable patterns extracted from the original data, these patterns are used to perform root cause analysis. This determines the reason for an anomaly that is experienced in a process. Subsequently, corrective actions are taken to return the process to normal operation.

The research conclusions are as follows:

- The entity-relationship modeling produces data models that organize data in the system by identifying entities that represent the manufacturing process, and their relationships with each other and their attributes. The developed data model characterizes the relevant information, which is visually represented and easily and easily understood by the organization's users. It obtains well-structured and high-quality data required by each stakeholder, with no missing or redundant data. It provides the required key performance of indices and standardizes the communication between these stakeholders. Consequently, each stakeholder accesses, views, understands and tracks the data flow within the database. Subsequently, further analyses or reports take little time, unlike before.

- The *LADR* technique improves the performance of the regression models significantly based on three metrics: MSE, MAE, and R^2. The reason for that is the concept improvement by introducing new three proposed classification methods: *EW*, *KM*, and *% STD*. Moreover, strong patterns with different degrees, that are extracted from the original data, are used to construct the model. The *LADR* technique demonstrates better and more significant results compared to other machine learning techniques by reducing the MSE with an average percentage of 70%.

- The *LADR-EWMA* is a newly developed model-based control chart that is used to monitor the quality of the manufacturing process. This integration overcomes the limitations of the conventional control charts.

  - The *LADR* technique remedies the problems of autocorrelation and the curse of dimensionality that lead to high false alarms and missed detection rates. Data preprocessing are applied to remove any duplications, multicollinearity, and dependency between the new binary independent variables that construct the *LADR* model.
  - Moreover, the *LADR* technique has a better performance that strengthens to maintain accurate control chart limit. It improves the performance of anomaly detection. Consequently, the *LADR-EWMA* reduces false alarm and missed detection rates compared to well-known techniques-based control charts.
  - The *LADR-EWMA* is not only used for anomaly detection but also for the identification of the root cause of that anomaly. Unlike the other machine learning techniques, *LADR-EWMA* identifies and interprets the root cause of anomalies based on the extracted patterns without acquiring anomalous data or even generating sufficient data that describes different anomalies.

  Generally, it provides a reliable and robust approach for online monitoring of manufacturing processes.

- The integration of *LADR* with the residual control chart (*RCC*) is used to develop a new condition and warning mechanism to monitor the operating conditions of industrial systems. The mechanism is used to determine whether the system operates under normal or faulty conditions. The benefit of the mechanism is that it improves the performance of detecting faults experienced in the system, in addition to determining the reason for that fault. This reduces the downtime and allows to take corrective action to return the system operates under normal conditions.

9.2 Future Research

With regards to further research, several future research areas are considered to improve the quality of a process, as follows:

- Extend data modeling to develop a physical model that will be implemented using DBMS software in an industrial system. This model will represent a real database that gathers the data in real-time from different sources in different formats from the system to provide each stakeholder with relevant information. Moreover, the database will be considered a base for the creation of a digital twin.
- Adapt the *LADR* to make an appropriate decision that will return the process to an in-control condition. Once the *LADR-EWMA* detects and identifies the root cause of an anomaly, it will be adapted to specify and adjust the exact values of the independent variables using patterns that construct the *LADR* model.
- Investigate the *LADR* with adaptive control limits of the control charts. Sometimes, the fixed threshold increases *FAR* and *MDR*, which affects the accuracy of anomaly detection. Therefore, adaptive control limits can be a solution, as their value changes based on historical statistics. The main challenge is to indicate the required parameters for the adaptive control limit based on normal data to ensure that there will be no presence of any false alarm when training the data.
- Implement the *LADR* based control chart in other applications such as profile monitoring and monitoring fraction non-conforming products in manufacturing to investigate the performance of the *LADR* compared to well-known existing approaches.

REFERENCES

[1] D. C. Montgomery, *Statistical quality control.* John Wiley & Son, 2020.

[2] A. Fountoulaki, N. Karacapilidis, and M. Manatakis, "Augmenting statistical quality control with machine learning techniques: an overview," *International Journal of Business and Systems Research*, vol. 5, no. 6, pp. 610–626, 2011.

[3] D. Garvin, "Competing on the eight dimensions of quality," *Harv. Bus. Rev.*, pp. 101–109, 1987.

[4] A. Amiri and S. Allahyari, "Change point estimation methods for control chart postsignal diagnostics: a literature review," *Quality and Reliability Engineering International*, vol. 28, no. 7, pp. 673–685, 2012.

[5] N. Abbas, M. Riaz, and R. J. Does, "An ewma-type control chart for monitoring the process mean using auxiliary information," *Communications in Statistics-Theory and Methods*, vol. 43, no. 16, pp. 3485–3498, 2014.

[6] Y. Dou and P. Sa, "One-sided control charts for the mean of positively skewed distributions," *Total Quality Management*, vol. 13, no. 7, pp. 1021–1033, 2002.

[7] N. Abbas, M. Riaz, and R. J. Does, "Enhancing the performance of ewma charts," *Quality and Reliability Engineering International*, vol. 27, no. 6, pp. 821–833, 2011.

[8] S. Cuentas, R. Peñabaena-Niebles, and E. Garcia, "Support vector machine in statistical process monitoring: a methodological and analytical review," *The International Journal of Advanced Manufacturing Technology*, vol. 91, no. 1-4, pp. 485–500, 2017.

[9] M. Kharbach, Y. Cherrah, Y. Vander Heyden, and A. Bouklouze, "Multivariate statistical process control in product quality review assessment–a case study," in *Annales Pharmaceutiques Françaises.* Elsevier, 75 (6) (2017) 446-454.

[10] D. Jacob, *Quality 4.0 Impact and Strategy Handbook, Cambrige: LNS Research*, 2017.

[11] J. Wang, Y.-s. Li, W. Song, and A.-h. Li, "Research on the theory and method of grid data asset management," *Procedia computer science*, vol. 139, pp. 440–447, 2018.

[12] O. J. Reichman, M. B. Jones, and M. P. Schildhauer, "Challenges and opportunities of open data in ecology," *Science*, vol. 331, no. 6018, pp. 703–705, 2011.

[13] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *science*, vol. 332, no. 6025, pp. 60–65, 2011.

[14] M. Williams, J. Bagwell, and M. N. Zozus, "Data management plans: the missing perspective," *Journal of Biomedical Informatics*, vol. 71, pp. 130–142, 2017.

[15] S. Ahmad, M. Riaz, S. Hussain, and S. A. Abbasi, "On auxiliary information-based control charts for autocorrelated processes with application in manufacturing industry," *The International Journal of Advanced Manufacturing Technology*, vol. 100, no. 5, pp. 1965–1980, 2019.

[16] W. Jiang, K. Wang, and F. Tsung, "A variable-selection-based multivariate ewma chart for process monitoring and diagnosis," *Journal of Quality Technology*, vol. 44, no. 3, pp. 209–230, 2012.

[17] Z. Yan and Y. Yao, "Variable selection method for fault isolation using least absolute shrinkage and selection operator (lasso)," *Chemometrics and Intelligent Laboratory Systems*, vol. 146, pp. 136–146, 2015.

[18] T.-H. Kuang, Z. Yan, and Y. Yao, "Multivariate fault isolation via variable selection in discriminant analysis," *Journal of Process Control*, vol. 35, pp. 30–40, 2015.

[19] M. Turkoz, S. Kim, Y.-S. Jeong, K. N. Al-Khalifa, and A. M. Hamouda, "Distribution-free adaptive step-down procedure for fault identification," *Quality and Reliability Engineering International*, vol. 32, no. 8, pp. 2701–2716, 2016.

[20] S. T. A. Niaki and B. Abbasi, "Fault diagnosis in multivariate control charts using artificial neural networks," *Quality and reliability engineering international*, vol. 21, no. 8, pp. 825–840, 2005.

[21] T.-f. Li, S. Hu, Z.-y. Wei, and Z.-q. Liao, "A framework for diagnosing the out-of-control signals in multivariate process using optimized support vector machines," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[22] C.-S. Cheng and H.-T. Lee, "Identifying the out-of-control variables of multivariate control chart using ensemble svm classifiers," *Journal of the Chinese Institute of Industrial Engineers*, vol. 29, no. 5, pp. 314–323, 2012.

[23] F. A. P. Peres and F. S. Fogliatto, "Variable selection methods in multivariate statistical process control: A systematic literature review," *Computers & Industrial Engineering*, vol. 115, pp. 603–619, 2018.

[24] S. Yacout, D. Salamanca, and M.-A. Mortada, "Tool and method for fault detection of devices by condition based maintenance," 2017, Google Patents, US Patent 9,824,060.

[25] M. Sony, J. Antony, J. A. Douglas, and O. McDermott, "Motivations, barriers and readiness factors for quality 4.0 implementation: an exploratory study," *The TQM Journal*, 2021.

[26] N. M. Radziwill, "Quality 4.0: Let's get digital-the many ways the fourth industrial revolution is reshaping the way we think about quality," *arXiv preprint arXiv:1810.07829*, 2018.

[27] M. Sony, J. Antony, and J. A. Douglas, "Essential ingredients for the implementation of quality 4.0: a narrative review of literature and future directions for research," *The TQM Journal*, 2020.

[28] B. Diène, J. J. Rodrigues, O. Diallo, E. H. M. Ndoye, and V. V. Korotaev, "Data management techniques for internet of things," *Mechanical Systems and Signal Processing*, vol. 138, p. 106564, 2020.

[29] G. Pujolle, "An autonomic-oriented architecture for the internet of things," in *IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06)*. IEEE, 2006, pp. 163–168.

[30] Y. Arora and D. Goyal, "Review of data analysis framework for variety of big data," in *Emerging Trends in Expert Applications and Security*. Springer, 2019, pp. 55–62.

[31] K. Sambrekar, V. S. Rajpurohit, and J. Joshi, "A proposed technique for conversion of unstructured agro-data to semi-structured or structured data," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–5.

[32] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.

[33] K. Shin, C. Hwang, and H. Jung, "Nosql database design using uml conceptual data model based on peter chen's framework," *International Journal of Applied Engineering Research*, vol. 12, no. 5, pp. 632–636, 2017.

[34] D. L. Moody, "Metrics for evaluating the quality of entity relationship models," in *International Conference on Conceptual Modeling*. Springer, 1998, pp. 211–225.

[35] R. Ramakrishnan, J. Gehrke, and J. Gehrke, *Database management systems.* McGraw-Hill New York, Vol. 3, 2003.

[36] T. M. Connolly and C. E. Begg, *Database systems: a practical approach to design, implementation, and management.* Pearson Education, 2005.

[37] W. Lemahieu, S. vanden Broucke, and B. Baesens, *Principles of database management: the practical guide to storing, managing and Analyzing big and small Data.* Cambridge University Press, 2018.

[38] C. Tupper, *Data architecture: from zen to reality.* Elsevier, 2011.

[39] P. V. Openko, S. Y. Hohoniants, O. V. Starkova, K. V. Herasymenko, M. I. Yastrebov, and A. O. Prudchenko, "Problem of choosing a dbms in modern information system," in *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT).* IEEE, 2019, pp. 171–174.

[40] R. T. Yarlagadda, "Data models in information technology," *INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]*, 2016.

[41] H. Vyawahare, P. P. Karde, and V. M. Thakare, "A hybrid database approach using graph and relational database," in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE).* IEEE, 2018, pp. 1–4.

[42] O. Bucovetchi, C. P. Simion, and R. D. Stanciu, "Object-oriented modelling applied to electricity critical infrastructures," *Procedia Technology*, vol. 19, pp. 651–656, 2015.

[43] Z. Ma and L. Yan, "Data modeling and querying with fuzzy sets: A systematic survey," *Fuzzy Sets and Systems*, 2022.

[44] Y. Suansook and T. Taweesri, "Modeling language for systems engineering in defense industry," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 39, no. 1, 2019, p. 6.

[45] L. Jacobson and J. R. G. Booch, "The unified modeling language reference manual," 2021.

[46] M. T. Alasaady, M. G. Saeed, and K. H. Faraj, "Evaluation and comparison framework for data modeling languages," in *2019 2nd International Conference on Electrical, Communication, Computer, Power and Control Engineering (ICECCPCE).* IEEE, 2019, pp. 68–73.

[47] S. Wolny, A. Mazak, C. Carpella, V. Geist, and M. Wimmer, "Thirteen years of sysml: a systematic mapping study," *Software and Systems Modeling*, vol. 19, no. 1, pp. 111–169, 2020.

[48] Y. Liu, X. Zeng, K. Zhang, and Y. Zou, "Transforming entity-relationship diagrams to relational schemas using a graph grammar formalism," in *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 2018, pp. 327–331.

[49] P. P.-S. Chen, "The entity-relationship model: Toward a unified view of data," *ACM Transactions on Database Systems*, vol. 1, no. 1, pp. 9–36, 1976.

[50] N. Guarino and G. Guizzardi, ""we need to discuss the relationship": revisiting relationships as modeling constructs," in *International Conference on Advanced Information Systems Engineering*. Springer, 2015, pp. 279–294.

[51] S. Al-Fedaghi, "Conceptual data modeling: Entity-relationship models as thinging machines," *arXiv preprint arXiv:2109.14717*, 2021.

[52] A. Ribeiro, A. Silva, A. R. da Silva *et al.*, "Data modeling and data analytics: a survey from a big data perspective," *Journal of Software Engineering and Applications*, vol. 8, no. 12, p. 617, 2015.

[53] S. Riaz, M. Riaz, Z. Hussain, and T. Abbas, "Monitoring the performance of bayesian ewma control chart using loss functions," *Computers & Industrial Engineering*, vol. 112, pp. 426–436, 2017.

[54] P. T. Theodossiou, "Predicting shifts in the mean of a multivariate time series process: an application in predicting business failures," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 441–449, 1993.

[55] E. Mergen, D. Grant, and S. M. Widrick, "Quality management applied to higher education," *Total Quality Management*, vol. 11, no. 3, pp. 345–352, 2000.

[56] J. A. Sellick, "The use of statistical process control charts in hospital epidemiology," *Infection Control & Hospital Epidemiology*, vol. 14, no. 11, pp. 649–656, 1993.

[57] M. J. Anderson and A. A. Thompson, "Multivariate control charts for ecological and environmental monitoring," *Ecological Applications*, vol. 14, no. 6, pp. 1921–1935, 2004.

[58] T. Clark and A. Clark, "Continuous improvement on the free-throw line," *Quality Progress*, vol. 30, no. 10, pp. 78–80, 1997.

[59] H. Pham, *Springer handbook of engineering statistics*. Vol. 49. London: Springer, 2006.

[60] M. A. Mahmoud and P. E. Maravelakis, "The performance of the mewma control chart when parameters are estimated," *Communications in Statistics—Simulation and Computation®*, vol. 39, no. 9, pp. 1803–1817, 2010.

[61] L. S. Nelson, "The shewhart control chart—tests for special causes," *Journal of quality technology*, vol. 16, no. 4, pp. 237–239, 1984.

[62] S. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 42, no. 1, pp. 97–101, 2000.

[63] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[64] M. B. Khoo and S. Quah, "Multivariate control chart for process dispersion based on individual observations," *Quality Engineering*, vol. 15, no. 4, pp. 639–642, 2003.

[65] S. W. Cheng and K. Thaga, "Single variables control charts: an overview," *Quality and Reliability Engineering International*, vol. 22, no. 7, pp. 811–820, 2006.

[66] W. H. Woodall and D. C. Montgomery, "Research issues and ideas in statistical process control," *Journal of Quality Technology*, vol. 31, no. 4, pp. 376–386, 1999.

[67] O. Y. Esparza Albarracin, A. P. Alencar, and L. L. Ho, "Effect of neglecting autocorrelation in regression ewma charts for monitoring count time series," *Quality and Reliability Engineering International*, vol. 34, no. 8, pp. 1752–1762, 2018.

[68] X. Ma, L. Zhang, J. Hu, and A. Palazoglu, "A model-free approach to reduce the effect of autocorrelation on statistical process control charts," *Journal of Chemometrics, Wiley Online Library*, vol. 32, no. 12, p. e3070, 2018.

[69] S. S. Prabhu and G. C. Runger, "Designing a multivariate ewma control chart," *Journal of Quality Technology*, vol. 29, no. 1, pp. 8–15, 1997.

[70] M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo *et al.*, "Intrusion detection system using multivariate control chart hotelling's t2 based on pca," *Int. J. Adv. Sci. Eng. Inf. Technol*, vol. 8, no. 5, pp. 1905–1911, 2018.

[71] C. Wu, Y. He, and K. Mi, "A low false alarm rates oriented design scheme of multivariate control chart," in *The Proceedings of 2011 9th International Conference on Reliability, Maintainability and Safety*. IEEE, 2011, pp. 1107–1111.

[72] Y. Zhao, S. Wang, and F. Xiao, "A statistical fault detection and diagnosis method for centrifugal chillers based on exponentially-weighted moving average control charts and support vector regression," *Applied Thermal Engineering*, vol. 51, no. 1-2, pp. 560–572, 2013.

[73] L. A. Jones, "The statistical design of ewma control charts with estimated parameters," *Journal of Quality Technology*, vol. 34, no. 3, pp. 277–288, 2002.

[74] N. A. Saleh, M. A. Mahmoud, L. A. Jones-Farmer, I. Zwetsloot, and W. H. Woodall, "Another look at the ewma control chart with estimated parameters," *Journal of Quality Technology*, vol. 47, no. 4, pp. 363–382, 2015.

[75] C. Escobar, J. Arinez, and R. Morales-Menendez, "Process-monitoring-for-quality-a step forward in the zero defects vision," SAE technical paper, Tech. Rep., 2020.

[76] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, 2016.

[77] K. P. Murphy, *Machine learning: A probabilistic perspective (adaptive computation and machine learning series)*. The MIT Press: London, UK, 2018.

[78] G. Köksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," *Expert systems with Applications*, vol. 38, no. 10, pp. 13 448–13 467, 2011.

[79] P. H. Tran, A. Ahmadi Nadi, T. H. Nguyen, K. D. Tran, and K. P. Tran, "Application of machine learning in statistical process control charts: A survey and perspective," in *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*. Springer, 2022, pp. 7–42.

[80] F. Harrou, M. N. Nounou, H. N. Nounou, and M. Madakyaru, "Pls-based ewma fault detection strategy for process monitoring," *Journal of Loss Prevention in the Process Industries*, vol. 36, pp. 108–119, 2015.

[81] A. Bakdi, A. Kouadri, and A. Bensmail, "Fault detection and diagnosis in a cement rotary kiln using pca with ewma-based adaptive threshold monitoring scheme," *Control Engineering Practice*, vol. 66, pp. 64–75, 2017.

[82] J. Liu, G. Li, H. Chen, J. Wang, Y. Guo, and J. Li, "A robust online refrigerant charge fault diagnosis strategy for vrf systems based on virtual sensor technique and pca-ewma method," *Applied Thermal Engineering*, vol. 119, pp. 233–243, 2017.

[83] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[84] T. Bouwmans and E. H. Zahzah, "Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance," *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, 2014.

[85] W. Liu, H. Zhang, D. Tao, Y. Wang, and K. Lu, "Large-scale paralleled sparse principal component analysis," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1481–1493, 2016.

[86] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[87] D. Marcondes Filho and A. M. O. Sant'Anna, "Principal component regression-based control charts for monitoring count data," *The International Journal of Advanced Manufacturing Technology*, vol. 85, no. 5, pp. 1565–1574, 2016.

[88] H. Wen, L. Liu, and X. Yan, "Regression-adjusted poisson ewma control chart," *Quality and Reliability Engineering International*, vol. 37, no. 5, pp. 1956–1964, 2021.

[89] M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo, and H. Khusna, "Multivariate control chart based on pca mix for variable and attribute quality characteristics," *Production & Manufacturing Research*, vol. 6, no. 1, pp. 364–384, 2018.

[90] W. J. Lee, G. P. Mendis, M. J. Triebe, and J. W. Sutherland, "Monitoring of a machining process using kernel principal component analysis and kernel density estimation," *Journal of Intelligent Manufacturing*, vol. 31, no. 5, pp. 1175–1189, 2020.

[91] M. Ahsan, M. Mashuri, H. Khusna, M. H. Lee *et al.*, "Multivariate control chart based on kernel pca for monitoring mixed variable and attribute quality characteristics," *Symmetry*, vol. 12, no. 11, p. 1838, 2020.

[92] M. Mashuri, M. Ahsan, H. Kuswanto, D. Prastyo, H. Khusna *et al.*, "Comparing the performance of t 2 chart based on pca mix, kernel pca mix, and mixed kernel pca for network anomaly detection," in *Journal of Physics: Conference Series*, vol. 1752, no. 1. IOP Publishing, 2021, p. 012008.

[93] M. Ahsan, M. Mashuri, H. Khusna *et al.*, "Kernel principal component analysis (pca) control chart for monitoring mixed non-linear variable and attribute quality characteristics," *Heliyon*, p. e09590, 2022.

[94] M. Farokhnia and S. T. A. Niaki, "Principal component analysis-based control charts using support vector machines for multivariate non-normal distributions," *Communications in Statistics-Simulation and Computation*, vol. 49, no. 7, pp. 1815–1838, 2020.

[95] W. Jiang, K.-L. Tsui, and W. H. Woodall, "A new spc monitoring method: The arma chart," *Technometrics*, vol. 42, no. 4, pp. 399–410, 2000.

[96] B. N. de Oliveira, M. Valk, and D. Marcondes Filho, "Fault detection and diagnosis of batch process dynamics using arma-based control charts," *Journal of Process Control*, vol. 111, pp. 46–58, 2022.

[97] J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, "Forecasting of demand using arima model," *International Journal of Engineering Business Management*, vol. 10, p. 1847979018808673, 2018.

[98] B. John and V. Agarwal, "A regression spline control chart for monitoring characteristics exhibiting nonlinear profile over time," *The TQM Journal*, 2019.

[99] R.-S. Guh and Y.-R. Shiue, "An effective application of decision tree learning for online detection of mean shifts in multivariate control charts," *Computers & Industrial Engineering*, vol. 55, no. 2, pp. 475–493, 2008.

[100] S. He, G. A. Wang, M. Zhang, and D. F. Cook, "Multivariate process monitoring and fault identification using multiple decision tree classifiers," *International Journal of Production Research*, vol. 51, no. 11, pp. 3355–3371, 2013.

[101] S. Du, J. Lv, and L. Xi, "On-line classifying process mean shifts in multivariate control charts based on multiclass support vector machines," *International Journal of Production Research*, vol. 50, no. 22, pp. 6288–6310, 2012.

[102] F.-K. Wang, B. Bizuneh, and X.-B. Cheng, "One-sided control chart based on support vector machines with differential evolution algorithm," *Quality and Reliability Engineering International*, vol. 35, no. 6, pp. 1634–1645, 2019.

[103] Z. Jian, B. Xia, C. Wang, and Z. Li, "Diagnosis of out-of-control signals in multivariate manufacturing processes with random forests," in *International Workshop of Advanced Manufacturing and Automation*. Springer, 2018, pp. 262–267.

[104] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE transactions on semiconductor manufacturing*, vol. 20, no. 4, pp. 345–354, 2007.

[105] A. Apsemidis, S. Psarakis, and J. M. Moguerza, "A review of machine learning kernel methods in statistical process monitoring," *Computers & Industrial Engineering*, vol. 142, p. 106376, 2020.

[106] K. Atashgar and R. Noorossana, "An integrating approach to root cause analysis of a bivariate mean vector with a linear trend disturbance," *The International Journal of Advanced Manufacturing Technology*, vol. 52, no. 1, pp. 407–420, 2011.

[107] D. D. Diren, S. Boran, I. H. Selvi, and T. Hatipoglu, "Root cause detection with an ensemble machine learning approach in the multivariate manufacturing process," in *Industrial Engineering in the Big Data Era*. Springer, 2019, pp. 163–174.

[108] C.-S. Cheng and H.-P. Cheng, "Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 198–206, 2008.

[109] M. Salehi, A. Bahreininejad, and I. Nakhai, "On-line analysis of out-of-control signals in multivariate manufacturing processes using a hybrid learning-based model," *Neurocomputing*, vol. 74, no. 12-13, pp. 2083–2095, 2011.

[110] S. Hosseini and B. M. H. Zade, "New hybrid method for attack detection using combination of evolutionary algorithms, svm, and ann," *Computer Networks*, vol. 173, p. 107168, 2020.

[111] G. Büchi, M. Cugno, and R. Castagnoli, "Smart factory performance and industry 4.0," *Technological Forecasting and Social Change*, vol. 150, p. 119790, 2020.

[112] M. C. Lucas-Estañ, M. Sepulcre, T. P. Raptis, A. Passarella, and M. Conti, "Emerging trends in hybrid wireless communication and data management for the industry 4.0," *Electronics*, vol. 7, no. 12, p. 400, 2018.

[113] J. E. See, "Visual inspection: a review of the literature." 2012.

[114] M. Rice, L. Li, G. Ying, M. Wan, E. T. Lim, G. Feng, J. Ng, M. Jin Li, and V. Bab, "Automating the visual inspection of aircraft," in *Aerospace technology and engineering conference*, 2018.

[115] T. L. Johnson, S. R. Fletcher, W. Baker, and R. Charles, "How and why we need to capture tacit knowledge in manufacturing: Case studies of visual inspection," *Applied ergonomics*, vol. 74, pp. 1–9, 2019.

[116] A. Tiwari, K. Vergidis, R. Lloyd, and J. Cushen, "Automated inspection using database technology within the aerospace industry," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 222, no. 2, pp. 175–183, 2008.

[117] J. Aust, S. Shankland, D. Pons, R. Mukundan, and A. Mitrovic, "Automated defect detection and decision-support in gas turbine blade inspection," *Aerospace*, vol. 8, no. 2, p. 30, 2021.

[118] W. J. Verhagen, R. Curran *et al.*, "An ontology-based approach for aircraft maintenance task support." in *ISPE CE*, 2013, pp. 494–506.

[119] C. Okoh, R. Roy, and J. Mehnen, "Maintenance informatics dashboard design for through-life engineering services," *Procedia CIRP*, vol. 59, pp. 166–171, 2017.

[120] J. A. Rodger and P. Pankaj, "Enterprise architecture ontology for supply chain maintenance and restoration of the sikorsky's uh-60 helicopter," *Designing Enterprise Architecture Frameworks*, p. 255, 2016.

[121] H. K. Al-Masree, "Extracting entity relationship diagram (erd) from relational database schema," *International Journal of Database Theory and Application*, vol. 8, no. 3, pp. 15–26, 2015.

[122] P. Kashmira and S. Sumathipala, "Generating entity relationship diagram from requirement specification based on nlp," in *2018 3rd International Conference on Information Technology Research (ICITR)*. IEEE, 2018, pp. 1–4.

[123] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, *Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry; final report of the Industrie 4.0 Working Group*. Forschungsunion, 2013.

[124] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[125] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," *Expert systems with applications*, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[126] I. H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques/Ian H. Witten, Eibe Frank, Mark A. Hall*. Morgan Kaufmann, 2016.

[127] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, "An implementation of logical analysis of data," *IEEE Transactions on knowledge and Data Engineering*, vol. 12, no. 2, pp. 292–306, 2000.

[128] A. Bennane and S. Yacout, "Lad-cbm; new data processing tool for diagnosis and prognosis in condition-based maintenance," *Journal of Intelligent Manufacturing*, vol. 23, no. 2, pp. 265–275, 2012.

[129] S. Yacout, "Fault detection and diagnosis for condition based maintenance using the logical analysis of data," *The 40th International Conference on Computers & Industrial Engineering*, pp. 1–6, 2010.

[130] C.-A. Chou, T. O. Bonates, C. Lee, and W. A. Chaovalitwongse, "Multi-pattern generation framework for logical analysis of data," *Annals of Operations Research*, vol. 249, no. 1-2, pp. 329–349, 2017.

[131] C. Guo and H. S. Ryoo, "Compact milp models for optimal and pareto-optimal lad patterns," *Discrete Applied Mathematics*, vol. 160, no. 16-17, pp. 2339–2348, 2012.

[132] M.-A. Mortada, S. Yacout, and A. Lakis, "Fault diagnosis in power transformers using multi-class logical analysis of data," *Journal of Intelligent Manufacturing*, vol. 25, no. 6, pp. 1429–1439, 2014.

[133] P. L. Hammer, A. Kogan, and M. A. Lejeune, "Reverse-engineering banks' financial strength ratings using logical analysis of data," 2007.

[134] A. Ragab, M. El-Koujok, B. Poulin, M. Amazouz, and S. Yacout, "Fault diagnosis in industrial chemical processes using interpretable patterns based on logical analysis of data," *Expert Systems with Applications*, vol. 95, pp. 368–383, 2018.

[135] M.-A. Mortada, T. Carroll, S. Yacout, and A. Lakis, "Rogue components: their effect and control using logical analysis of data," *Journal of Intelligent Manufacturing*, vol. 23, no. 2, pp. 289–302, 2012.

[136] P. L. Hammer and T. O. Bonates, "Logical analysis of data—an overview: From combinatorial optimization to medical applications," *Annals of Operations Research*, vol. 148, no. 1, pp. 203–225, 2006.

[137] T. O. Bonates and P. L. Hammer, "Pseudo-boolean regression," Rutgers Center for OR, Tech. Rep., 2007.

[138] P. Lemaire, "Extensions of logical analysis of data for growth hormone deficiency diagnoses," *Annals of Operations Research*, vol. 186, no. 1, pp. 199–211, 2011.

[139] M.-A. Mortada, S. Yacout, and A. Lakis, "Diagnosis of rotor bearings using logical analysis of data," *Journal of Quality in Maintenance Engineering*, vol. 17, no. 4, pp. 371–397, 2011.

[140] M. Lejeune, V. Lozin, I. Lozina, A. Ragab, and S. Yacout, "Recent advances in the theory and practice of logical analysis of data," *European Journal of Operational Research*, vol. 275, no. 1, pp. 1–15, 2019.

[141] A. Ragab, S. Yacout, and M. Ouali, "Interpretable pattern-based machine learning for condition based maintenance," in *Conference RAMS2015, Florida, USA*, 2015.

[142] L. Torgo and J. Gama, "Regression by classification," *Brazilian symposium on artificial intelligence, Springer*, pp. 51–60, 1996.

[143] ——, "Regression using classification algorithms," *Intelligent Data Analysis*, vol. 1, no. 4, pp. 275–292, 1997.

[144] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, p. e1300, 2019.

[145] H. Xie, L. Zhang, C. P. Lim, Y. Yu, C. Liu, H. Liu, and J. Walters, "Improving k-means clustering with enhanced firefly algorithms," *Applied Soft Computing*, vol. 84, p. 105763, 2019.

[146] D. Conway and J. White, *Machine learning for hackers*. " O'Reilly Media, Inc.", 2012.

[147] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer (112), 2013.

[148] R. A. P. Dias, J. Petrini, J. B. S. Ferraz, J. P. Eler, R. S. Bueno, A. L. L. da Costa, and G. B. Mourão, "Multicollinearity in genetic effects for weaning weight in a beef cattle composite population," *Livestock Science*, vol. 142, no. 1-3, pp. 188–194, 2011.

[149] C. García, J. García, M. López Martín, and R. Salmerón, "Collinearity: Revisiting the variance inflation factor in ridge regression," *Journal of Applied Statistics*, vol. 42, no. 3, pp. 648–661, 2015.

[150] R. Salmerón Gómez, J. García Pérez, M. D. M. López Martín, and C. G. García, "Collinearity diagnostic applied in ridge estimation through the variance inflation factor," *Journal of Applied Statistics*, vol. 43, no. 10, pp. 1831–1849, 2016.

[151] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[152] D. C. Corrales, J. C. Corrales, and A. Ledezma, "How to address the data quality issues in regression models: a guided process for data cleaning," *Symmetry*, vol. 10, no. 4, p. 99, 2018.

[153] R. A. Johnson, D. W. Wichern *et al.*, *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 5 (8) 2002.

[154] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA, 2015. [Online]. Available: http://www.rstudio.com/

[155] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[156] D. G. Pereira, A. Afonso, and F. M. Medeiros, "Overview of friedman's test and post-hoc analysis," *Communications in Statistics-Simulation and Computation*, vol. 44, no. 10, pp. 2636–2653, 2015.

[157] A. Ragab, S. Yacout, M.-S. Ouali, and H. Osman, "Pattern-based prognostic methodology for condition-based maintenance using selected and weighted survival curves," *Quality and Reliability Engineering International*, vol. 33, no. 8, pp. 1753–1772, 2017.

[158] D. M. Hawkins, "Multivariate quality control based on regression-adiusted variables," *Technometrics*, vol. 33, no. 1, pp. 61–75, 1991.

[159] H. S. Ryoo and I.-Y. Jang, "Milp approach to pattern generation in logical analysis of data," *Discrete Applied Mathematics*, vol. 157, no. 4, pp. 749–761, 2009.

[160] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual reviews in control*, vol. 36, no. 2, pp. 220–234, 2012.

[161] A. Haq, R. Gulzar, and M. B. Khoo, "An efficient adaptive ewma control chart for monitoring the process mean," *Quality and Reliability Engineering International*, vol. 34, no. 4, pp. 563–571, 2018.

[162] S. B. Kim, W. Jitpitaklert, S.-K. Park, and S.-J. Hwang, "Data mining model-based control charts for multivariate and autocorrelated processes," *Expert Systems with Applications*, vol. 39, no. 2, pp. 2073–2081, 2012.

[163] N. Subramanyam and A. A. Houshmand, "Simultaneous representation of multivariate and corresponding univariate x charts using line-graph," *Quality Engineering*, vol. 7, no. 4, pp. 681–692, 1995.

[164] O. O. Atienza, L. C. Tang, and B. W. Ang, "Quality notes: Simultaneous monitoring of univariate and multivariate spc information using boxplots," *International Journal of Quality Science*, 1998.

[165] L. W. Blazek, B. Novic, and D. M. Scott, "Displaying multivariate data using polyplots," *Journal of Quality Technology*, vol. 19, no. 2, pp. 69–74, 1987.

[166] C. Fuchs and Y. Benjamini, "Multivariate profile charts for statistical process control," *Technometrics*, vol. 36, no. 2, pp. 182–195, 1994.

[167] B. Murphy, "Selecting out of control variables with the t2 multivariate quality control procedure," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 36, no. 5, pp. 571–581, 1987.

[168] M.-K. Chua and D. C. Montgomery, "Investigation and characterization of a control scheme for multivariate quality control," *Quality and Reliability Engineering International*, vol. 8, no. 1, pp. 37–44, 1992.

[169] R. L. Mason, N. D. Tracy, and J. C. Young, "Decomposition of t 2 for multivariate control chart interpretation," *Journal of quality technology*, vol. 27, no. 2, pp. 99–108, 1995.

[170] J. Kim, M. K. Jeong, E. A. Elsayed, K. Al-Khalifa, and A. Hamouda, "An adaptive step-down procedure for fault variable identification," *International Journal of Production Research*, vol. 54, no. 11, pp. 3187–3200, 2016.

[171] C. A. Escobar, M. E. McGovern, and R. Morales-Menendez, "Quality 4.0: a review of big data challenges in manufacturing," *Journal of Intelligent Manufacturing*, vol. 32, no. 8, pp. 2319–2334, 2021.

[172] S. Connell, "As industry 4.0 continues to evolve, what can quality professionals do to ensure they will be an integral asset throughout this industrial revolution?-quality in mind," *Quality in Mind*, 2017.

[173] A. Chiarini, "Industry 4.0, quality management and tqm world. a systematic literature review and a proposed agenda for further research," *The TQM Journal*, 2020.

[174] B. Mandel, "The regression control chart," *Journal of Quality Technology*, vol. 1, no. 1, pp. 1–9, 1969.

[175] W. Gani, H. Taleb, and M. Limam, "Support vector regression based residual control charts," *Journal of Applied Statistics*, vol. 37, no. 2, pp. 309–324, 2010.

[176] S. B. Kim, W. Jitpitaklert, V. C. Chen, J. Lee, and S.-K. Park, "Data mining model adjustment control charts for cascade processes," *European Journal of Industrial Engineering*, vol. 7, no. 4, pp. 442–455, 2013.

[177] C. F. Alcala and S. J. Qin, "Reconstruction-based contribution for process monitoring," *Automatica*, vol. 45, no. 7, pp. 1593–1600, 2009.

[178] D. Bhamare and P. Suryawanshi, "Review on reliable pattern recognition with machine learning techniques," *Fuzzy Information and Engineering*, vol. 10, no. 3, pp. 362–377, 2018.

[179] R. M. Khalifa, S. Yacout, and S. Bassetto, "Developing machine-learning regression model with logical analysis of data (lad)," *Computers & Industrial Engineering*, vol. 151, p. 106947, 2021.

[180] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–12, 1990.

[181] S. H. Steiner, "Ewma control charts with time-varying control limits and fast initial response," *Journal of Quality Technology*, vol. 31, no. 1, pp. 75–86, 1999.

[182] P. Mehta and P. Monteiro, *Concrete : Microstructure, Properties, and Materials: Microstructure, Properties, and Materials*, ser. McGraw Hill professional. Mcgraw-hill, 2005.

[183] I.-C. Yeh, "Modeling slump flow of concrete using second-order regressions and artificial neural networks," *Cement and concrete composites*, vol. 29, no. 6, pp. 474–480, 2007.

[184] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[185] A. Bakdi and A. Kouadri, "A new adaptive pca based thresholding scheme for fault detection in complex systems," *Chemometrics and Intelligent Laboratory Systems*, vol. 162, pp. 83–93, 2017.

[186] H. Zhu, W. Zhu, and W. Fan, "Dynamic modeling, simulation and experiment of power transmission belt drives: A systematic review," *Journal of Sound and Vibration*, vol. 491, p. 115759, 2021.

[187] S. Chowdhury and R. K. Yedavalli, "Dynamics of belt-pulley-shaft systems," *Mechanism and Machine Theory*, vol. 98, pp. 199–215, 2016.

[188] R. S. Beikmann, N. C. Perkins, and A. G. Ulsoy, "Free Vibration of Serpentine Belt Drive Systems," *Journal of Vibration and Acoustics*, vol. 118, no. 3, pp. 406–413, 1996.

[189] H. Zhu, Y. Hu, W. Zhu, and Y. Pi, "Optimal design of an autotensioner in an automotive belt drive system via a dynamic adaptive pso-ga," *Journal of Mechanical Design*, vol. 139, no. 9, p. 093302, 2017.

[190] H. Zhu, Y. Hu, W. Zhu, and H. Long, "Dynamic responses of an engine front-end accessory belt drive system with pulley eccentricities via two spatial discretization methods," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 232, no. 4, pp. 482–498, 2018.

[191] "GUNT PT500.14," https://www.gunt.de/en/products/mechatronics/machinery-diagnosis/belt-drive-kit/052.50014/pt500-14/glct-1:pa-148:ca-77:pr-1032, (accessed on 30 December 2021).

[192] "GUNT PT500," https://www.gunt.de/en/products/machinery-diagnostic-system-base-unit/052.50000/pt500/glct-1:pa-148:pr-1022, (accessed on 30 December 2021).

[193] R. L. Mott, *Machine Elements in Mechanical Design Edition: 6th edition*. Pearson, 2017.

[194] R. Martínez-Guerra, R. Garrido, R. Palacios, and J. Mendoza-Camargo, "Fault detection in a belt-drive system using a proportional reduced order observer," in *Proceedings of the 2004 American Control Conference*, vol. 4. IEEE, 2004, pp. 3106–3110.

[195] A. G. Piersol and T. L. Paez, *Harris' shock and vibration handbook*. McGraw-Hill Education, 2010.

[196] S. Ojha, D. Sarangi, B. Pal, and B. Biswal, "Performance monitoring of vibration in belt conveyor system," *Journal of Engineering Research and Applications*, vol. 4, no. 7, pp. 22–31, 2014.

[197] S. Kumar, M. Lokesha, K. Kumar, and K. Srinivas, "Vibration based fault diagnosis techniques for rotating mechanical components," in *IOP Conference Series: Materials Science and Engineering*, vol. 376, no. 1. IOP Publishing, 2018, p. 012109.

[198] H. Yang, J. Mathew, and L. Ma, "Vibration feature extraction techniques for fault diagnosis of rotating machinery: a literature survey," in *Asia-Pacific Vibration Conference*, no. 42460, 2003, pp. 801–807.

[199] A. Bulushi, G. Rameshkumar, and M. Lokesha, "Fault diagnosis in belts using time and frequency based signal processing techniques," *International Journal of Multidisciplinary Sciences and Engineering*, vol. 6, no. 11, pp. 12–20, 2015.

[200] A. Nabhan, M. R. El-Sharkawy, and A. Rashed, "Monitoring of belt-drive defects using the vibration signals and simulation models," *International Journal of Aerospace and Mechanical Engineering*, vol. 13, no. 5, pp. 332–339, 2019.

[201] A. Ameer, A. Nabhan, R. Mohamed, and A. Rashed, "Dynamic model analysis for unsteady operating of double v-belt drive system," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 2950–2963, 2021.

[202] A. R. Hassan and K. M. Ali, "Dignosis of pulley-belt system faults using vibration analysis technique," *Journal of University of Babylon for Engineering Sciences*, vol. 26, no. 2, pp. 167–180, 2018.

[203] M. Ahmed, F. Gu, and A. Ball, "Fault detection of reciprocating compressors using a model from principles component analysis of vibrations," in *Journal of Physics: Conference Series*, vol. 364, no. 1. IOP Publishing, 2012, p. 012133.

[204] W. Li, Z. Wang, Z. Zhu, G. Zhou, and G. Chen, "Design of online monitoring and fault diagnosis system for belt conveyors based on wavelet packet decomposition and support vector machine," *Advances in Mechanical Engineering*, vol. 5, p. 797183, 2013.

[205] M. Khazaee, A. Banakar, B. Ghobadian, M. Mirsalim, S. Minaei, M. Jafari, and P. Sharghi, "Fault detection of engine timing belt based on vibration signals using data-mining techniques and a novel data fusion procedure," *Structural Health Monitoring*, vol. 15, no. 5, pp. 583–598, 2016.

[206] A. A. Jaber and K. M. Ali, "Artificial neural network based fault diagnosis of a pulley-belt rotating system," *Int J Adv Sci Eng Inform Technol*, vol. 9, pp. 544–551, 2019.

[207] National Instruments, http://www.ni.com/labview/, National Instruments Website.

[208] Z. Peng, W. Zhang, Z. Lang, G. Meng, and F. Chu, "Time–frequency data fusion technique with application to vibration signal analysis," *Mechanical systems and signal processing*, vol. 29, pp. 164–173, 2012.

[209] R. M. Khalifa, S. Yacout, and S. Bassetto, "Root cause analysis of an out-of-control process using a logical analysis of data regression model and exponential weighted moving average," *Manuscript submitted for publication, (2022).*

[210] Y. Shaban, M. Meshreki, S. Yacout, M. Balazinski, and H. Attia, "Process control based on pattern recognition for routing carbon fiber reinforced polymer," *Journal of Intelligent Manufacturing*, vol. 28, no. 1, pp. 165–179, 2017.

[211] Y. Shaban, S. Yacout, and M. Balazinski, "Tool wear monitoring and alarm system based on pattern recognition with logical analysis of data," *Journal of manufacturing science and engineering*, vol. 137, no. 4, 2015.

[212] Y. Shaban, S. Yacout, M. Balazinski, and K. Jemielniak, "Cutting tool wear detection using multiclass logical analysis of data," *Machining Science and Technology*, vol. 21, no. 4, pp. 526–541, 2017.

[213] R. Osei-Aning, S. Abbasi, and M. Riaz, "Monitoring of serially correlated processes using residual control charts," *Scientia Iranica*, vol. 24, no. 3, pp. 1603–1614, 2017.

[214] J. S. Utley and J. G. May, "Monitoring service quality with residuals control charts," *Managing Service Quality: An International Journal*, 2009.

[215] A. A. Jaber and R. Bicker, "Industrial robot fault detection based on statistical control chart," *Am. J. Eng. Applied Sci*, vol. 9, pp. 251–263, 2016.

[216] J. P. Guarnieri, A. M. Souza, L. F. Jacobi, B. Reichert, and C. P. da Veiga, "Control chart based on residues: Is a good methodology to detect outliers?" *Journal of Industrial Engineering International*, vol. 15, no. 1, pp. 119–130, 2019.

[217] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R.* Spinger, 2013.

[218] G. A. Seber and A. J. Lee, *Linear regression analysis.* John Wiley & Sons, 2012, vol. 329.

[219] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis.* John Wiley & Sons, 2021.

[220] A. J. Smola and B. Schölkopf, "On a kernel-based method for pattern recognition, regression, approximation, and operator inversion," *Algorithmica*, vol. 22, no. 1, pp. 211–231, 1998.

[221] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[222] C. Bratsas, K. Koupidis, J.-M. Salanova, K. Giannakopoulos, A. Kaloudis, and G. Aifadopoulou, "A comparison of machine learning methods for the prediction of traffic speed in urban places," *Sustainability*, vol. 12, no. 1, p. 142, 2020.

[223] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[224] P. Ein-Dor and J. Feldmesser, "Attributes of the performance of central processing units: A relative performance prediction model," *Communications of the ACM*, vol. 30, no. 4, pp. 308–317, 1987.

APPENDIX A *UCI* DATASETS

The appendix provides a descriptions for the six datasets and identify the independent and dependent variables that are are used to build the models.
Boston housing is a dataset that was obtained by the U.S Census Service. It concerns on prediction of the median value of a house for different areas in Boston. It contains 13 independent variables and single dependent variable as shown in A.1.

Table A.1 Boston housing dataset

Abbrev.	Description	Variable type	
crim	Per capita crime rate by town.	Continuous	Independent
zn	Proportion of residential land zoned for lots over 25,000 sq.ft.	Continuous	Independent
indus	Proportion of non-retail business acres per town.	Continuous	Independent
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).	Discrete	Independent
nox	Nitrogen oxides concentration (parts per 10 million).	Continuous	Independent
rm	Average number of rooms per dwelling.	Continuous	Independent
age	Proportion of owner-occupied units built prior to 1940.	Continuous	Independent
dis	Weighted mean of distances to five Boston employment centres.	Continuous	Independent
rad	Index of accessibility to radial highways.	Continuous	Independent
tax	Full-value property-tax rate per $10,000.	Continuous	Independent
ptratio	Pupil-teacher ratio by town.	Continuous	Independent
b	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.	Continuous	Independent
lstat	Lower status of the population (percent).	Continuous	Independent
medv	Median value of owner-occupied homes in $1000s.	Continuous	Dependent

Computer Hardware dataset gathered the performance of 209 CPUs on the market from 1981 to 1984 . It contains six independent variables represent the specifications of the CPU in terms of memory size, cycle time,...etc in A.2. The models estimate the relative performance of these CPUs with respect to a base machine-the IBM 370/158 [224].

Auto-MPG dataset was maintained at Carnegie Mellon University. It estimates the fuel con-

Table A.2 Computer Hardware dataset

Abbrev.	Description	Variable type	
MYCT	Machine cycle time in nanoseconds	Continuous	Independent
MMIN	Minimum main memory in kilobytes	Continuous	Independent
MMAX	Maximum main memory in kilobytes	Continuous	Independent
CACH	Cache memory in kilobytes	Continuous	Independent
CHMIN	Minimum channels in units	Continuous	Independent
CHMAX	Maximum channels in units	Continuous	Independent
PRP	Published relative performance	Continuous	Dependent

sumption in miles per gallon in terms of 5 continuous and 3 multi-valued discrete independent variables as depicted in A.3.

Table A.3 Auto-MPG dataset

Abbrev.	Description	Variable type	
CYL	Cylinders	Multi-valued discrete	Independent
DISP	Displacement	Continuous	Independent
HP	Horsepower	Continuous	Independent
W	Weight	Continuous	Independent
ACCEL	Acceleration	Continuous	Independent
M_Y	Model_year	Multi-valued discrete	Independent
MPG	Miles per gallon	Continuous	Dependent

Servo dataset was a servo system simulation that was done at MIT in 1986. The system contains a motor, a servo amplifier, sliding carriage and a lead screw wit its nut. The dependent variable is respond time that the system requires to change the position set point A.4.

Table A.4 Servo dataset

Abbrev.	Description	Variable type	
M	Motor	Multi-valued discrete	Independent
S	Screw	Multi-valued discrete	Independent
PG	Pgain	Multi-valued discrete	Independent
VG	Vgain	Multi-valued discrete	Independent
C	Class	Continuous	Dependent

Airfoil Self-Noise dataset is experimental tests that was carried out by NASA. These tests

were applied on different airfoils at various wind tunnel in terms of speeds and attack angles. The dataset contains five independent variables and one dependent variable as shown in A.5.

Table A.5 Airfoil Self-Noise dataset

Abbrev.	Description	Variable type	
X1	Frequency (Hz)	Continuous	Independent
X2	Angle of attack (degrees)	Multi-valued discrete	Independent
X3	Chord length (m)	Continuous	Independent
X4	Free-stream velocity (m/s)	Continuous	Independent
X5	Suction side displacement thickness (m)	Continuous	Independent
Y	Scaled sound pressure level (db)	Continuous	Dependent

Concrete Compressive Strength dataset represents various mixes of concrete. The aim is the prediction of compressive strength in MPa of the high performance concrete (HPC). The eight independent variables in this dataset are continuous A.6.

Table A.6 Concrete Compressive Strength dataset

Abbrev.	Description	Variable type	
X1	Cement (kg in a m3 mixture)	Continuous	Independent
X2	Blast Furnace Slag (kg in a m3 mixture)	Continuous	Independent
X3	Fly Ash (kg in a m3 mixture)	Continuous	Independent
X4	Water (kg in a m3 mixture)	Continuous	Independent
X5	Superplasticizer (kg in a m3 mixture)	Continuous	Independent
X6	Coarse Aggregate (kg in a m3 mixture)	Continuous	Independent
X7	Fine Aggregate (kg in a m3 mixture)	Continuous	Independent
X8	Age (Day)	Continuous	Independent
Y	Concrete compressive strength (Mpa)	Continuous	Dependent

APPENDIX B *LADR* MODELS

The best *LADR* model for *Boston Housing* dataset:

$$
\begin{aligned}
Y_{[KM]} =& 1.2703 - 0.1164X_{P_3} + 0.0487X_{P_4} - 0.1058X_{P_5} - 0.1221X_{P_7} - 0.1884X_{P_8} + \\
& 0.0164X_{P_{11}} + 0.0238X_{P_{15}} + 0.0198X_{P_{16}} - 0.0605X_{P_{18}} - 0.0689X_{P_{19}} - \\
& 0.0203X_{P_{20}} - 0.0367X_{P_{23}} - 0.025X_{P_{27}} - 0.0345X_{P_{28}} - 0.0738X_{P_{32}} - \\
& 0.0252X_{P_{35}} - 0.0809X_{P_{36}} + 0.0131X_{P_{38}} - 0.0101X_{P_{43}} + 0.0937X_{P_{46}} + \\
& 0.0284X_{P_{47}} - 0.023X_{P_{48}} + 0.0167X_{P_{49}} + 0.0628X_{P_{51}} + 0.0113X_{P_{56}} + \\
& 0.0329X_{P_{57}} - 0.0254X_{P_{64}} - 0.0311X_{P_{65}} - 0.0167X_{P_{67}} - 0.0328X_{P_{68}} - \\
& 0.0261X_{P_{70}} - 0.0742X_{P_{71}} + 0.0204X_{P_{73}} + 0.015X_{P_{74}} - 0.0265X_{P_{76}} - \\
& 0.0129X_{P_{77}} + 0.0137X_{P_{79}} - 0.0267X_{P_{80}} + 0.0248X_{P_{83}} + 0.0237X_{P_{84}} + \\
& 0.0201X_{P_{92}} + 0.0212X_{P_{93}} - 0.0216X_{P_{94}} - 0.0139X_{P_{105}} + 0.0124X_{P_{107}} - \\
& 0.0116X_{P_{113}} + 0.0249X_{P_{114}} + 0.0122X_{P_{116}} + 0.1454X_{P_{117}} + 0.0264X_{P_{119}} - \\
& 0.0198X_{P_{124}} + 0.0225X_{P_{125}} + 0.0235X_{P_{129}} + 0.0359X_{P_{131}} + 0.0379X_{P_{132}} - \\
& 0.0161X_{P_{133}} + 0.0273X_{P_{134}} + 0.0334X_{P_{140}} - 0.0108X_{P_{143}} - 0.0288X_{P_{149}} - \\
& 0.0163X_{P_{150}} - 0.0268X_{P_{151}} + 0.0116X_{P_{153}} + 0.0125X_{P_{155}} + 0.0185X_{P_{157}} - \\
& 0.0161X_{P_{158}} + 0.0289X_{P_{159}} + 0.036X_{P_{163}} + 0.0247X_{P_{164}} + 0.0566X_{P_{166}} + \\
& 0.042X_{P_{167}}
\end{aligned} \tag{B.1}
$$

The best *LADR* model for *Computer Hardware* dataset:

$$
\begin{aligned}
Y_{[KM]} =& 11.2587 - 2.0575X_{P_1} - 1.3239X_{P_2} + 1.2485X_{P_4} - 1.4162X_{P_5} - 1.3892X_{P_6} - \\
& 1.3585X_{P_7} + 0.7451X_{P_{22}} - 1.2595X_{P_{25}} - 0.4814X_{P_{27}} - 1.0078X_{P_{30}} - \\
& 0.9581X_{P_{32}} - 0.8165X_{P_{33}} + 0.9431X_{P_{39}} + 0.6209X_{P_{45}} - 0.5994X_{P_{51}} - \\
& 0.8359X_{P_{54}} - 0.7095X_{P_{60}} - 0.9653X_{P_{67}} + 0.8524X_{P_{78}} + 1.1236X_{P_{80}} + \\
& 2.6497X_{P_{110}} - 1.5187X_{P_{111}} + 1.0201X_{P_{112}} + 0.8517X_{P_{113}} + 4.49X_{P_{115}} + \\
& 3.867X_{P_{117}} - 1.87X_{P_{129}} + 1.105X_{P_{132}} + 1.6002X_{P_{135}} + 0.8697X_{P_{137}} - \\
& 2.2783X_{P_{141}} + 2.0073X_{P_{152}} + 1.2658X_{P_{155}} - 0.9564X_{P_{157}} - 1.001X_{P_{160}} + \\
& 1.431X_{P_{162}} - 1.0867X_{P_{166}} + 1.8233X_{P_{168}} + 3.4175X_{P_{170}} + 9.7904X_{P_{172}}
\end{aligned} \tag{B.2}
$$

The best *LADR* model for *Auto-mpg* dataset:

$$
\begin{aligned}
Y_{[EW]} =& 1.3714 - 0.0294X_{P_1} - 0.0617X_{P_2} - 0.0292X_{P_4} - 0.0464X_{P_5} - 0.047X_{P_7} - \\
& 0.0489X_{P_8} + 0.029X_{P_9} - 0.0128X_{P_{10}} - 0.0296X_{P_{13}} + 0.0135X_{P_{19}} - \\
& 0.0314X_{P_{28}} - 0.0444X_{P_{32}} - 0.0301X_{P_{33}} - 0.0543X_{P_{34}} - 0.096X_{P_{35}} - \\
& 0.0452X_{P_{37}} - 0.0257X_{P_{42}} - 0.044X_{P_{43}} + 0.0236X_{P_{49}} + 0.0147X_{P_{50}} - \\
& 0.0294X_{P_{52}} + 0.0234X_{P_{55}} + 0.0134X_{P_{56}} + 0.0245X_{P_{60}} + 0.018X_{P_{61}} + \\
& 0.0152X_{P_{64}} + 0.02X_{P_{65}} + 0.0674X_{P_{66}} + 0.0739X_{P_{67}} + 0.0362X_{P_{68}} + \\
& 0.0128X_{P_{70}} + 0.0139X_{P_{71}} + 0.0234X_{P_{72}} - 0.0251X_{P_{75}} - 0.0107X_{P_{77}} - \\
& 0.0186X_{P_{78}} - 0.0282X_{P_{79}} - 0.0182X_{P_{90}} - 0.0257X_{P_{91}} - 0.0323X_{P_{92}} - \\
& 0.0301 * X_{P_{93}} - 0.0801X_{P_{95}} + 0.0179X_{P_{96}} - 0.0712X_{P_{98}} - 0.0635X_{P_{100}} + \\
& 0.0674X_{P_{102}} + 0.1285X_{P_{103}} + 0.0623X_{P_{104}} + 0.1128X_{P_{106}} + 0.024X_{P_{107}} + \\
& 0.1938X_{P_{108}} + 0.0283X_{P_{109}} + 0.0441X_{P_{110}}
\end{aligned}
\tag{B.3}
$$

The best *LADR* model for *Servo* dataset:

$$
\begin{aligned}
Y_{[KM]} =& 0.3816 - 0.1031X_{P_1} - 0.0696X_{P_2} - 0.1315X_{P_3} - 0.1079X_{P_4} - 0.0562X_{P_5} - \\
& 0.0662X_{P_6} - 0.046X_{P_8} - 0.0596X_{P_{10}} + 0.054X_{P_{13}} + 0.0405X_{P_{14}} + 0.0395X_{P_{15}} + \\
& 0.0522X_{P_{20}} + 0.0593X_{P_{21}} + 0.0842X_{P_{22}} + 0.044X_{P_{23}} + 0.1542X_{P_{24}} - 0.0275X_{P_{26}} \\
& - 0.0218X_{P_{28}} - 0.0637X_{P_{30}} + 0.0486X_{P_{32}} + 0.0412X_{P_{33}}
\end{aligned}
\tag{B.4}
$$

The best *LADR* model for *Airfoil Self-Noise* dataset:

$$
\begin{aligned}
Y_{[EW]} =& 125.1 + 1.5X_{P_4} - 1.5X_{P_5} + 2.1X_{P_6} + 3.2X_{P_8} + 2.2X_{P_9} - 2.2X_{P_{10}} + 1.3 \\
& X_{P_{28}} + X_{P_{30}} + 1.3X_{P_{33}} + 3.1X_{P_{36}} - 0.7X_{P_{37}} + 2X_{P_{38}} + 2X_{P_{43}} + 0.9X_{P_{44}} \\
& + 1.2X_{P_{46}} - 0.6X_{P_{48}} - 0.9X_{P_{50}} - 0.8X_{P_{51}} - 0.5X_{P_{52}} + 1.3X_{P_{54}} - 6.3X_{P_{55}} \\
& + 5.3X_{P_{57}} - 0.9X_{P_{58}} - 3.9X_{P_{59}} - 2.7X_{P_{63}} - 2.6X_{P_{64}} + 1.8X_{P_{66}} - 1.4X_{P_{68}} \\
& - 2.1X_{P_{71}} + 8.4X_{P_{72}} - 3.6X_{P_{73}} - 2.2X_{P_{74}} + 2.2X_{P_{77}} - 0.8X_{P_{78}} + 1.2X_{P_{79}} \\
& + 0.3X_{P_{82}} + 1.1X_{P_{86}} - 0.2X_{P_{90}} + 0.4X_{P_{91}} + 0.5X_{P_{96}} - 0.6 * X_{P_{97}} + 0.5 \\
& X_{P_{106}} + 1.1 * X_{P_{110}} + 0.3X_{P_{116}} + 0.2X_{P_{117}} - 1.6X_{P_{118}} - 1.4X_{P_{120}} + 1.3X_{P_{121}} \\
& - 4.9X_{P_{122}} + 1.5X_{P_{125}} - 2.3X_{P_{126}} - 1.2X_{P_{127}} - 2.1X_{P_{128}} - 1.3X_{P_{132}} - 3.2 \\
& X_{P_{135}} - 1.1X_{P_{138}} - 2.7X_{P_{139}} - 6X_{P_{140}} + 4X_{P_{141}} - 3.2X_{P_{142}} - 1.3X_{P_{148}} - 2.3 \\
& X_{P_{149}} - 1.9X_{P_{151}} - 1.6X_{P_{152}} + 1.5X_{P_{153}} + 1.1X_{P_{155}} + 1.6X_{P_{157}} - 0.6X_{P_{159}} - \\
& 2.7X_{P_{161}} + 0.6X_{P_{168}} + 0.5X_{P_{169}} + 0.5X_{P_{172}} - 6X_{P_{174}} - 5X_{P_{175}} + 2.1X_{P_{177}} - \\
& 1.4X_{P_{180}} - 1.1X_{P_{181}} - 2.3X_{P_{182}} - 0.4X_{P_{183}} - 2.2X_{P_{184}} + 0.4X_{P_{185}} - 1.6X_{P_{186}} \\
& - 3.3X_{P_{188}} - 0.7X_{P_{189}} - 3X_{P_{191}} - 0.8X_{P_{195}} - 0.8X_{P_{196}} - 1.4X_{P_{197}} - 0.8X_{P_{198}} \\
& + X_{P_{200}} + 1.9X_{P_{203}} + 0.9X_{P_{205}} - 1.4X_{P_{207}} + 1.4X_{P_{209}} - X_{P_{210}} + 0.6X_{P_{216}} - \\
& 0.5X_{P_{220}} + 0.5X_{P_{221}} - 0.3X_{P_{223}} + 0.8X_{P_{224}} - 1.6X_{P_{227}} + 0.8X_{P_{230}} + 2.3X_{P_{234}} \\
& - 0.8X_{P_{236}} + 0.6X_{P_{237}} + 1.2X_{P_{239}} - 0.5X_{P_{245}} + 0.4X_{P_{246}} - 0.2X_{P_{247}} - 1.4 \\
& X_{P_{248}} + 0.5X_{P_{249}} + 1.7X_{P_{252}} + 0.5X_{P_{253}} + 0.3X_{P_{255}} + 0.7X_{P_{257}} - 0.6X_{P_{258}} + \\
& 1.4X_{P_{259}} - 0.9X_{P_{260}} + X_{P_{261}} + 0.6X_{P_{262}} + X_{P_{265}} + 1.2X_{P_{267}} + 2X_{P_{268}} + 0.9 \\
& X_{P_{271}} + 1.1X_{P_{272}} + 2.1X_{P_{279}} - 0.3X_{P_{280}} + 2.7X_{P_{281}} + 0.5X_{P_{282}} + 1.9X_{P_{283}} + \\
& 0.4X_{P_{284}} + 1.1X_{P_{288}} + 1.2X_{P_{289}} + 2.7X_{P_{290}} + 1.4X_{P_{291}} + 2.2X_{P_{292}} + 0.7X_{P_{293}} \\
& - 1.1X_{P_{294}} + 2.8X_{P_{295}} - 0.7X_{P_{299}} - 3.4X_{P_{308}} - 0.4X_{P_{310}} + X_{P_{313}} - 2.45X_{P_{314}} \\
& - X_{P_{319}} - 2.5X_{P_{322}} - 1.7X_{P_{324}} - 0.4X_{P_{326}} - 1.9X_{P_{327}} - 1.1X_{P_{330}} - 4.2X_{P_{331}} \\
& + 0.3X_{P_{332}} - 3.4X_{P_{334}} - 3.1X_{P_{339}} - 1.1X_{P_{340}} - 0.9X_{P_{341}} - 2.7X_{P_{343}} - 1.9 \\
& X_{P_{345}} - 1.6X_{P_{347}} - 0.7X_{P_{349}} - 4X_{P_{350}} - 1.2X_{P_{351}} - 0.2X_{P_{352}} - 1.4X_{P_{355}} - \\
& 1.9X_{P_{358}} - 1.3X_{P_{360}} - 1.1X_{P_{362}} - 1.2X_{P_{363}} - 2X_{P_{365}} + 0.4X_{P_{366}} - 0.8X_{P_{367}} - \\
& 1.1X_{P_{368}} - 3.8X_{P_{372}} - 2.2X_{P_{376}} - 1.6X_{P_{377}} - 2.1X_{P_{378}} - 0.4X_{P_{379}} - 0.7X_{P_{380}} \\
& - 0.6X_{P_{382}} - X_{P_{383}} + 0.9X_{P_{384}} - 1.5X_{P_{388}} - 0.8X_{P_{389}} - 1.8X_{P_{390}} + 1.9X_{P_{392}} \\
& - 1.7X_{P_{393}} + 1.1X_{P_{399}} - 1.7X_{P_{400}} - 0.4X_{P_{404}} - 0.3X_{P_{405}} - 0.4X_{P_{406}} - 0.8X_{P_{409}} \\
& - 0.9X_{P_{412}} - 0.9X_{P_{414}} - 1.4X_{P_{415}} - X_{P_{417}} - 0.8X_{P_{419}} + 2X_{P_{420}} - 0.5X_{P_{424}} + \\
& 0.6X_{P_{425}} + 1.6X_{P_{426}} + 0.7X_{P_{429}} - 1.4X_{P_{432}} - 0.6X_{P_{433}} - 1.7X_{P_{434}} + 0.9X_{P_{436}} \\
& + X_{P_{437}} + 1.1X_{P_{438}} - 1.2X_{P_{439}} + 2X_{P_{441}} - 2X_{P_{444}} + 1.3X_{P_{447}} + 0.7X_{P_{451}} + \\
& 1.8X_{P_{454}} + X_{P_{456}} - 4X_{P_{457}} + 2.4X_{P_{458}} + 0.6X_{P_{459}} + 1.9X_{P_{461}} + 1.2X_{P_{462}} +
\end{aligned}
$$

$$
\begin{aligned}
&1.5X_{P_{465}} + 2.9X_{P_{466}} + 4.6X_{P_{467}} + 2.4X_{P_{468}} - 1.2X_{P_{470}} + 0.5X_{P_{471}} + 1.9 \\
&X_{P_{472}} + 1.3X_{P_{476}} + 2.2X_{P_{478}} + 2.2X_{P_{479}} + 0.8X_{P_{483}} - 2X_{P_{484}} + 1.2X_{P_{485}} + \\
&1.9X_{P_{486}} + 0.6X_{P_{488}} + 1.9X_{P_{489}} - 1.5X_{P_{491}} + 2X_{P_{493}} + 3.1X_{P_{494}} + 7.5X_{P_{496}} \\
&- 2.2X_{P_{498}} + 3.6X_{P_{499}} + 4X_{P_{501}} + 1.9X_{P_{502}} - 3X_{P_{504}} + 3.6X_{P_{505}} + 1.9X_{P_{506}} \\
&+ 2.1X_{P_{508}} + 1.5X_{P_{509}} + 2.7X_{P_{510}} + 2.6X_{P_{511}} + 1.5X_{P_{512}} + 2.7X_{P_{513}} + 1.9 \\
&X_{P_{514}} + 3.5X_{P_{516}} + 1.1X_{P_{517}} + X_{P_{518}} + 0.6X_{P_{520}} + 2.2X_{P_{522}} + 2.8X_{P_{524}} + \\
&1.1X_{P_{525}} + 5.4X_{P_{526}} + 3.2X_{P_{528}} + 2.1X_{P_{530}} + 8.8X_{P_{531}} + 3.6X_{P_{532}} + 0.9X_{P_{533}} \\
&- 0.3X_{P_{536}} - 1.4X_{P_{537}} - 1.6X_{P_{539}} + 0.7X_{P_{543}} + 1.1X_{P_{548}} - 0.3X_{P_{557}} - 0.4 \\
&X_{P_{559}} - 1.5X_{P_{560}} + 2X_{P_{562}} - 0.8X_{P_{574}} - 0.4X_{P_{579}} - 1.3X_{P_{582}} - 0.3X_{P_{583}} - \\
&2X_{P_{584}} - 0.5X_{P_{586}} - 1.8X_{P_{587}} - X_{P_{590}} - 1.6X_{P_{592}} + 0.5X_{P_{593}} - 1.2X_{P_{598}} \\
&- 2X_{P_{601}} - 1.9X_{P_{602}} + 3.1X_{P_{608}} + 1.7X_{P_{611}} + 6.2X_{P_{613}} + 1.9X_{P_{617}} + 3 \\
&X_{P_{620}} + 2.5X_{P_{621}} + 2.3X_{P_{623}} - 1.6X_{P_{624}} + 7.9X_{P_{625}} + 5.1X_{P_{626}} - 1.8X_{P_{629}} \\
&+ 4.7X_{P_{630}} + 4.6X_{P_{631}} + 14.5X_{P_{632}}
\end{aligned}
\tag{B.5}
$$

The best *LADR* model for *Concrete Compressive Strength* dataset:

$$
\begin{aligned}
Y_{[EW]} =& 36.5 + 4.4X_{P_1} + 3.3X_{P_3} - 4.7X_{P_6} - 6.6X_{P_{12}} - 4.8X_{P_{13}} - 3.6X_{P_{17}} + 1.1X_{P_{20}} \\
&+ 3.7X_{P_{21}} + 2X_{P_{23}} - 2.4X_{P_{24}} + 0.7X_{P_{32}} + 1.8X_{P_{34}} - 2.3X_{P_{39}} - 6.7X_{P_{41}} + \\
&2X_{P_{46}} + 4.4X_{P_{47}} + 2.3X_{P_{48}} - 12.6X_{P_{49}} + 2.1X_{P_{56}} + 3.8X_{P_{57}} - 6.8X_{P_{62}} - \\
&2.6X_{P_{63}} + 1.8X_{P_{64}} + 4X_{P_{68}} + 1.3X_{P_{71}} - 0.9X_{P_{73}} - 1.2X_{P_{81}} + 2.5X_{P_{83}} - \\
&2.5X_{P_{90}} - 1.7X_{P_{94}} + 2X_{P_{95}} - 1.2X_{P_{97}} + 1.3X_{P_{100}} + 1.5X_{P_{101}} + 2.5X_{P_{110}} + \\
&7.7X_{P_{120}} + 1.5X_{P_1}121 - 2.6X_{P_{122}} - 3.8X_{P_{124}} - 14.7X_{P_{131}} - 1.7X_{P_{132}} - \\
&1.5X_{P_{135}} - 2X_{P_{137}} - 5.6X_{P_{140}} - 2.8X_{P_{142}} - 2.3X_{P_{147}} - 8.2X_{P_{150}} + 1.7X_{P_{151}} \\
&+ 0.9X_{P_{153}} + 1.5X_{P_{158}} - 3.6X_{P_{160}} + 1.3X_{P_{161}} - 0.8X_{P_{165}} + 2.1X_{P_{166}} + 2.6 \\
&X_{P_{167}} + 2X_{P_{170}} + 1.6X_{P_{173}} + 2.7X_{P_{175}} + 1.5X_{P_{176}} + 4.2X_{P_{183}} + 3X_{P_{186}} + \\
&3.3X_{P_{187}} - 3X_{P_{188}} - 6.1X_{P_{208}} - 1.6X_{P_{210}} + 3.4X_{P_{211}} - 6.7X_{P_{213}} - 2X_{P_{217}} \\
&+ 3.8X_{P_{218}} - 2.7X_{P_{220}} + 2.5X_{P_{222}} - 3.8X_{P_{224}} + 1.4X_{P_{225}} - 1.5X_{P_{229}} - 3.7 \\
&X_{P_{234}} + 1.1X_{P_{235}} + 5.5X_{P_{236}} - 5.4X_{P_{237}} - 4.9X_{P_{239}} + X_{P_{240}} - 1.6X_{P_{243}} + \\
&3.7X_{P_{245}} - 4.3X_{P_{248}}\breve{}2.9X_{P_{249}} + 1.2X_{P_{251}} - 3.2X_{P_{252}} + 1.7X_{P_{255}} - 1.5X_{P_{258}} \\
&- 5.4X_{P_{259}} - 2.7X_{P_{260}} + 2.6X_{P_{263}} - 2.3X_{P_{266}} - 4.1X_{P_{267}} + 2.4X_{P_{269}} + 10.3 \\
&X_{P_{270}} - 4X_{P_{273}} + 2.7X_{P_{279}} + 1.6X_{P_{281}} + 5.6X_{P_{284}} + X_{P_{289}} + 7.8X_{P_{291}} + 1.9 \\
&X_{P_{292}} + 2.7X_{P_{293}} + 4X_{P_{294}} + 6X_{P_{295}} + 5.1X_{P_{297}} + 3.2X_{P_{298}} - 4.4X_{P_{304}}1.5X_{P_{306}} \\
&- 3.2X_{P_{308}} + 1.9X_{P_{309}} + 4.4X_{P_{313}} + 1.1X_{P_{315}} - 3.6X_{P_{316}} + 2.3X_{P_{317}} + 4.4X_{P_{320}} - \\
&2X_{P_{326}} - 3.6X_{P_{334}} - 0.7X_{P_{340}} - 2.3X_{P_{346}} - 1.5X_{P_{351}} - 1.3X_{P_{353}} - 1.5X_{P_{354}} -
\end{aligned}
$$

$$
\begin{aligned}
&3.2X_{P_{361}} - 3X_{P_{362}} - 2.2X_{P_{363}} - 1.7X_{P_{365}} - 4X_{P_{366}} - 0.9X_{P_{369}} - 1.6X_{P_{371}} - \\
&3.5X_{P_{372}} - 6.2X_{P_{373}} - 2.6X_{P_{374}} - 4.5X_{P_{380}} - 0.9X_{P_{381}} + 1.2X_{P_{384}} - 2.5 \\
&X_{P_{385}} - 4.3X_{P_{386}} - 2.3X_{P_{387}} + 3.6X_{P_{390}} - 2.8X_{P_{391}} - 1.7X_{P_{394}} + 3.7X_{P_{403}} + \\
&4.3X_{P_{404}} + 2.4X_{P_{406}} + 1.9X_{P_{408}} + 2X_{P_{411}} - 2.7X_{P_{412}} + 4.2X_{P_{419}} - 1.5X_{P_{420}} \\
&+ 5.3X_{P_{421}} + 2.3X_{P_{424}} + 5.5X_{P_{427}} + 3.2X_{P_{429}} + 3.2X_{P_{430}} + 4.6X_{P_{431}} - 5.3 \\
&X_{P_{433}} + 3.4X_{P_{437}} + 1.2X_{P_{438}} + 2.6X_{P_{444}} + 0.9X_{P_{474}} - 1.6X_{P_{488}} - 2X_{P_{492}} - \\
&3.3X_{P_{498}} - 2.4X_{P_{500}} - 1.6X_{P_{502}} + 8.1X_{P_{504}} - 1.4X_{P_{510}} - 3.4X_{P_{517}} - 7.5X_{P_{522}} \\
&+ 4.1X_{P_{526}} + 2.8X_{P_{527}} - 1.8X_{P_{533}} - 12.8X_{P_{537}} + 3X_{P_{538}} + 3.4X_{P_{544}} + 3.2 \\
&X_{P_{547}} + 0.9X_{P_{557}} - 0.8X_{P_{563}} - 1.8X_{P_{565}} - 3.2X_{P_{568}} + 0.7X_{P_{571}} - 5.6X_{P_{572}} - \\
&3.8X_{P_{581}} - 9.5X_{P_{585}} + 10.3X_{P_{592}} + 1.9X_{P_{613}} - 3.4X_{P_{614}} + 2.3X_{P_{616}} - 2.6 \\
&X_{P_{617}} + 3.1X_{P_{622}} + 6.7X_{P_{624}} + 6.1X_{P_{626}} + 21.1X_{P_{627}} + 6.4X_{P_{628}} + 3.9X_{P_{630}} - \\
&2.2X_{P_{640}} - 2.3X_{P_{641}} + 1.9X_{P_{643}} - 1.1X_{P_{645}} + 5.4X_{P_{648}} + 10.8X_{P_{649}} + 15.6 \\
&X_{P_{651}} + 23X_{P_{653}} + 11X_{P_{654}} + 31.6X_{P_{655}}
\end{aligned}
\tag{B.6}
$$

APPENDIX C ARTICLE 5: EXPERIMENTAL VIBRATION DATA COLLECTED FOR A BELT DRIVE SYSTEM UNDER DIFFERENT OPERATING CONDITIONS

Ramy M. Khalifa, Soumaya Yacout, Samuel Bassetto, Yasser Shaban

Submitted to:
Data in Brief, 2022

C.1 Abstract

Vibration analysis is the cornerstone of vibration-based condition monitoring that analyzes a vibration signal, detects faults or anomalies, and diagnoses the operating conditions of a belt drive system. This data article contains experiments that collect vibration signals of a belt drive system at different levels of speed and pretension of the belt under varying operating conditions. The collected dataset includes low, medium, and high operating speeds at three levels of the belt's pretensioned values. This article covers three operating conditions: normal or healthy operation using a healthy belt, unbalanced operation by adding unbalanced weight to the system, and abnormal operation using a faulty belt. The collected data provides an understanding of the performance of the belt drive system during the operation to identify the root cause of an anomaly when detected.

Keyword: Vibration signal, Belt drive system, Condition monitoring, Belt conditions

C.2 Specifications Table

Subject	Mechanical Engineering, Industrial Engineering
Specific subject area	Vibration-based condition monitoring and vibration analysis of a belt drive system under different operating conditions
Type of data	Tables in .txt files and figures in .JPG files
How data were acquired	The vibration data was collected by data acquisition system (Accelerometers, Amplifier, and USB data acquisition card) during system operation. The operating speed of the system are controlled and maintained by a speed controller. The pretension value of the belt is adjusted by using a pretension gauge.
Data format	Raw
Parameters for data collection	The data was acquired based on three experimental settings: healthy operation of the system, presence of unbalanced weight, and faulty operation. All of these settings have three parameters in which the experiments are carried out at different operating speeds and pretension values of the belt, in addition to the value of adding weights in case the unbalanced weight settles.

Description of data collection	The data is of vibration signals that were collected by using data acquisition system through the two accelerometers mounted on the test rig. Then, the data was transmitted to a laptop for analysis.
Data source location	Data was obtained from the test rig shown in Figure C.1 in: Institution: Faculty of Engineering - Department of Mechanical Design - Helwan university City: Cairo Country: Egypt
Data accessibility	The data is available in the Mendeley repository at: `https://data.mendeley.com/datasets/jf8v2ndydr/1`
Related research article	R. M. Khalifa, S. Yacout, S. Bassetto, Y. Shaban, Condition monitoring and warning of the belt drive system based on *LADR* based residual control chart, Mechanical Systems and Signal Processing. In Press.

C.3 Value of the Data

- The data represents the vibration signals collected from the belt drive system under healthy and faulty conditions. This data is for a commonly used system in various industrial applications.
- The data is useful to the researcher and practitioners in mechanical and industrial engineering, to analyze the vibration signals of the belt drive system and to determine its characteristics under healthy and faulty conditions.
- The data can be used for online condition process monitoring in order to detect and diagnose any anomaly or faulty condition in the system. It can be used to evaluate developed machine learning approaches that distinguish the conditions of the belt drive system.

C.4 Data Description

The belt drive system is widely used in different industrial applications for power transmission such as conveyors, machine tools, and motors [186]. It consists of a motor, shaft(s), bearings,

belt(s), and driver and driven pulleys [187]. The system operates at different speeds and transmits power using a pretension of the belt. The system is used to produce different types of anomalies that are developed due to several abnormal sources of vibrations in the system, such as cut or damage of the belt, unbalance problems, and misalignment. [188–190].

This article comprises the experiments that collect vibration signals of the belt drive system at different levels of speed and pretension of the belt under different conditions of healthy belt, faulty belt, and the presence of unbalanced weight. The vibration signals are collected from accelerometers attached to the driver and driven pulleys. The collected data includes 17 levels of speed; 400 to 2000 RPM by step of 100; three levels of pretension values, 70, 110, and 150 N, and two levels, according to whether there was the presence or absence of unbalanced weight. The data is from three different operating conditions: normal operation using a healthy belt, unbalanced operation by adding weights that cause imbalance in the system, and anomalous operation from using a faulty belt. Each experiment has been repeated three times, which resulted in 459 runs.

This article is accompanied by nine folders and its name is given as "T-BC-W" where T, BC, W denote the belt pretension in newton, identification of the belt condition, and absence (W=0) or presence (W=U) of unbalanced weight, respectively. Thus, the data contains the following folders:

- Data 70-H-0: The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 70 N and in healthy condition in addition to the absence of unbalanced weight at all levels of speed.

- Data 110-H-0 : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 110 N and in healthy condition in addition to the absence of unbalanced weight at all levels of speed.

- Data 150-H-0 : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 150 N and in healthy condition in addition to the absence of unbalanced weight at all levels of speed.

- Data 70-F-0 : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 70 N and in the faulty condition in addition to the absence of unbalanced weight at all levels of speed.

- Data 110-F-0 : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 110 N and in the faulty condition in addition to the absence of unbalanced weight.

- Data 150-F-0: The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 150 N and in the faulty condition in addition to the absence of unbalanced weight at all levels of speed.
- Data 70-H-U : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 70 N and in healthy condition in addition to the presence of unbalanced weight at all levels of speed.
- Data 110-H-U : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 110 N and in healthy condition in addition to the presence of unbalanced weight at all levels of speed.
- Data 150-H-U : The vibration signals are collected from the belt drive system when the belt is pre-tensioned by 150 N and in healthy condition in addition to the presence of unbalanced weight at all levels of speed.

Each folder contains 51 TXT files and 51 figures in JPG format that describe each operating condition at different speeds. Both TXT and JPG have the same names. Each TXT file contains 10000 samples. Since each experiment run has been repeated three times, every three files represent the vibration signals that are collected from the operation of the system at the same speed as in table C.1.

C.5 Experimental Design, Materials and Methods

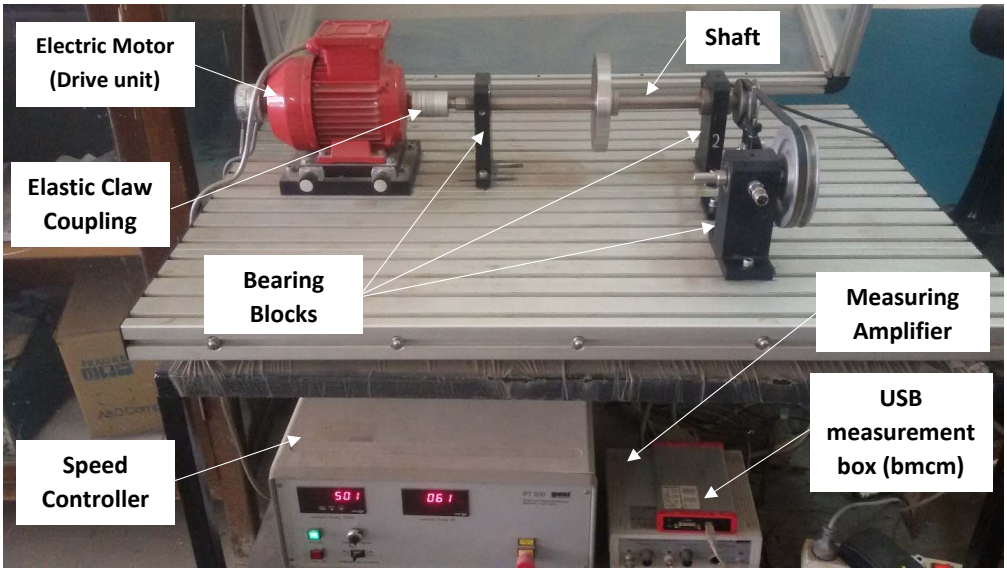
The experiments are performed using the belt drive kit of the G.U.N.T machinery diagnostic system (PT 500.14) [191] as depicted in figure C.1. The key component of the experiment is a base, unit G.U.N.T (PT 500), [192] that consists of an electric motor as the rotating equipment, a shaft, and two bearing blocks. The speed of the electric motor "N" is adjusted by a speed controller. An elastic coupling is considered as the connection between the motor and the shaft where it is used to avoid misalignment and increase the flexibility of the shaft. The two bearing blocks have a ball bearing type that supports the shaft. On the other hand, the G.U.N.T (PT 500.14) consists of a per-tensioned V-belt that connects small driver and large driven pulleys. The diameter of the small driver pulley is 63 mm which is connected to the shaft of the G.U.N.T (PT 500). The V-belt is SPZ type with a length of 912mm and width of 10 mm. It is the machine element that transmits power to the large driven pulley , which has a diameter of 125 mm. The pretension of the belt "T" is adjusted using tensioning rollers. A pretension gauge is used to measure the value of T as shown in figure C.2.

The Vibration signals are collected using a Data Acquisition system. This has two piezoelec-

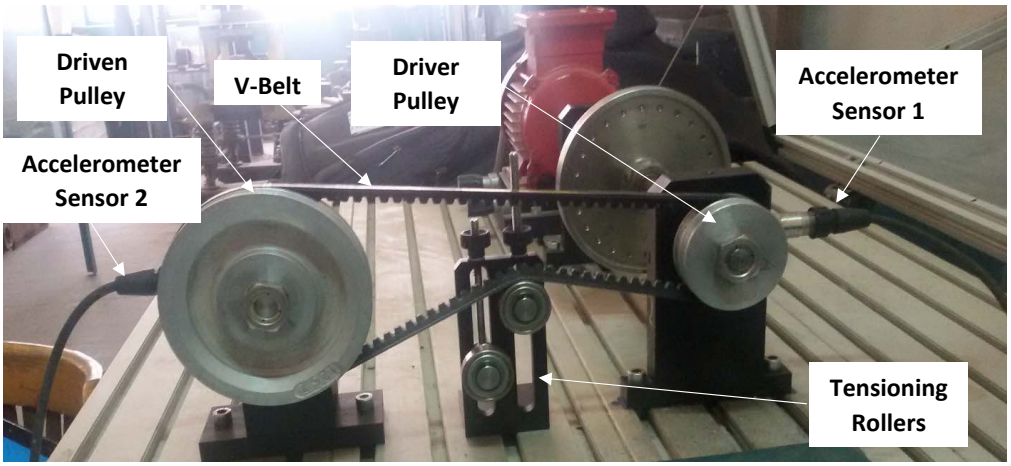
Table C.1 Files and figures description

TXT/JPG files	Speed, RPM
1 to 3	400
4 to 6	500
7 to 9	600
10 to 12	700
13 to 15	800
16 to 18	900
19 to 21	1000
22 to 24	1100
25 to 27	1200
28 to 30	1300
31 to 33	1400
34 to 36	1500
37 to 39	1600
40 to 42	1700
43 to 45	1800
46 to 48	1900
49 to 51	2000

tric accelerometers (IMI 603C01); accelerometers 1 and 2 are attached to the bearing block of the driver and driven pulley, respectively, in a horizontal direction using studs. They are used to measure the signals during the experimental run. An amplifier is used to amplify these signals. The output signals from the amplifier are digitalized using a USB data acquisition card (bmcm) and the collected signals are transferred to the *LabVIEW* script installed on a laptop for further analysis. The experiments are conducted to investigate the characteristics of the vibration signals for the belt drive system under three levels of T (70, 110, and 150 N), absence and presence of unbalanced weight, using healthy and faulty belts at 17 levels of speed N that ranges from 400 to 2000 RPM in step 100. Figures C.3 and C.4 show the presence of unbalanced weight, and the healthy and faulty belts.



(a)



(b)

Figure C.1 G.U.N.T machinery diagnostic system (PT 500.14) Description

Figure C.2 The pretension gauge

(a) (b)

Figure C.3 G.U.N.T (PT 500.14) - the healthy and faulty belts

Figure C.4 G.U.N.T (PT 500.14) - presence of unbalanced weights