AUDIT OPINION DECISION USING ARTIFICIAL INTELLIGENCE TECHNIQUES: EMPIRICAL STUDY OF UK AND IRELAND

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

by

Aram Nawaiseh

Department of Electronic and Electrical Engineering
College of Engineering, Design and Physical Sciences
Brunel University London

December 2021

Abstract

The reliability and quality of a final decision regarding auditing opinion is a significant issue for auditors. The new field of data mining in auditing remains in its infancy and is increasingly explored through creating reliable and effective auditing opinion classification models. Previous studies have called for more exploration that is needed of the individual classifier models and committee combiner methods in the auditing field. Particularly, previous studies have not yet investigated or applied data mining dynamic modelling and even they have not encouraged future studies to do search in dynamic modelling in auditing and accounting area. Thus, this thesis study investigates the ability of a classification tool to classify correct audit opinion and explores dynamic modelling. To the best of this researcher's knowledge, this is the first research that involves dynamic modelling research in auditing opinion. Two evaluation measurement parameters that have not been used in any previous auditing studies or any related area are used to evaluate performance accurately: Brier score and area under reliability diagram (AURD).

This thesis aims to develop the ability performance of nine classifiers (support vector machines, artificial neural networks, K-nearest neighbour, decision trees, naïve Bayes network, logistic regression, linear discriminant analysis, boosting ensemble and a novel deep learning model), offering a classification tool for correct audit opinion. The empirical evaluation results indicate that for the four tested datasets, the deep learning model revealed superior ability in classifying the audit opinion accurately, outperforming all other models by obtaining highest values at all nine evaluation parameters. Subsequently, significance statistical testing revealed that the deep learning model has best ability to classify audit opinion correctly.

Thereafter, the audit opinion model was enhanced by combining all nine individual classifiers to improve the accuracy of the audit opinion modelling according to six traditional committee modelling rules (Average, Weighted average, Median, Min; Max and Majority voting). Moreover, the Consensus combiner and Fuzzy logic combiner models were added to the committee modelling. The performance of each committee modelling technique was assessed individually, and subsequently their abilities to classify audit opinion correctly were compared to determine whether committee modelling can improve upon the accuracy performance of individual classifiers. Consensus model showed superior ability to classify audit opinion correctly, and enhanced accuracy in audit opinion modelling compared with individual classifiers, which delivered the best evaluation measurement results over the four datasets, and the best statistical test results.

The final contribution was developing of traditional dynamic modelling methods (nonlinear autoregressive exogenous and nonlinear autoregressive) and novel dynamic model (deep learning-LSTM), utilised to predict audit opinion in advance. These models were tested, and individual performance results being compared with the benchmark model result, which was a deep learning classifier that tested actual audit opinion data for the advanced year. Lastly, all dynamic modelling performances were compared using the benchmark classifier. Deep learning-LSTM had better performance in predicting audit opinion in advance compared with the other models, in terms of the best evaluation results.

Publications Based on this Research

Conferences

Nawaiseh, A.K. and Abbod M.F. (2021) ‘Financial statement audit utilising naive Bayes networks, decision trees, linear discriminant analysis and logistic regression’. *The importance of new technologies and entrepreneurship in business development: in the context of economic diversity in developing countries*. ICBT 2020, Lecture Notes in Networks and Systems, vol. 194. Cham: Springer, pp. 1305-1320. doi: 10.1007/978-3-030-69221-6_97.

Nawaiseh, A.K., Abbod, M.F. and Itagaki, T. (2020) ‘Financial statement audit using support vector machines, artificial neural networks and K-nearest neighbor: an empirical study of UK and Ireland’, *International Journal of Simulation Systems, Science & Technology Special Issue: Conference Proceedings UKSim2020*, 21(2), pp. 7.1-7.6. doi: 10.5013/IJSSST.a.21.02.07.

Journals

Nawaiseh, A.K., Abbod, M.F. Consensus Combiner Model Approach for Audit Opinion Classifier. Expert Systems with Applications (Submitted 28 Oct. 2021, Under 1st review).

Declaration

I declare that the research in this thesis is the author’s work and submitted for the first time to the Post Graduate Research Office at Brunel University London. The study was originated, composed and reviewed by the mentioned author in the Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, UK. All information derived from other works has been referenced and acknowledged.

Aram Nawaiseh

December 2021

London, UK

Acknowledgements

In the name of Allah, the Most Beneficent, the Most Merciful.

Ultimately, I am thankful and indebted to God, who has given me patience and power to finalise this thesis.

I dedicate this thesis to my parents, sisters, and brother for their continuous love, encouragement, and support throughout my life’s journey. I would like to extend my deepest gratitude to my parents: I owe them everything. Their sacrifices have made it possible for me to reach where I am now. They believed and incessantly reminded me that no dream is impossible, with them standing by my side, and with God as my centre. This lifelong dream of obtaining a PhD has finally come true.

I am tremendously grateful to my supervisor, Dr Maysam Abbod, for his motivation and valuable guidance throughout this thesis. He steadfastly encouraged me to work on this thesis, which was essential to achieving the purposes of this research. I very much appreciate his patience with me, particularly when introducing me to new areas. His input has helped me understand machine learning and he has always been there whenever I have needed him. Given that my background is in accounting, and the highly technical field of machine learning was thus new to me, his advice and knowledge fundamentally enabled me to understand the subject, and he contributed immensely to my personal and academic development.

Finally, I extend my gratitude to my PhD colleagues, friends, and officemates, who have assisted me and shared their genuine support during this important episode of my life.

Table of Contents

Abstract i

Publications Based on this Research iii

Conferences iii

Journals iii

Declaration iv

Acknowledgements v

Table of Contents vi

List of Tables xi

List of Figures xiii

List of Equations xv

List of Abbreviations xvii

Chapter 1 Introduction 1

1.1. Background 1

1.2. Research Motivation 2

1.3. Aims and Objectives of the Research 3

1.4. Contributions to Knowledge 4

1.5. Research Methodology 5

1.6. Thesis Outline 5

Chapter 2 Theoretical Background and Literature Review 7

2.1. Background 7

2.2. Auditing Opinion Decision 7

2.2.1. Definition of Financial Statement Audit 7

2.2.2. Audit Phases 8

2.2.3. Audit Evidence 9

2.2.4. Analytical Review 9

2.3. Audit and Information Technology 10

2.3.1. Big Data and Audit Process 11

2.3.2. Data Mining Analysis in Auditing Decision 14

2.4. Audit Opinion Decision Studies 18

2.4.1. Literature Search 19

2.4.2. Literature Review Analysis and Decision 20

2.5. Summary 34

Chapter 3 Research Methodology for the Audit Opinion Model 36

3.1. Introduction 36

3.2. Prediction Model Approaches 36

3.2.1. Individual Classifiers 36

3.2.1.1. Logistic Regression 37

3.2.1.2. K-Nearest Neighbour 38

3.2.1.3. Naïve Bayes Network 39

3.2.1.4. Linear Discriminant Analysis 40

3.2.1.5. Decision Tree 41

3.2.1.6. Artificial Neural Network 42

3.2.1.7. Support Vector Machine 44

3.2.1.8. Boosting Ensemble Classifier 46

3.2.1.9. Deep Learning Model 47

3.2.2. Committee Combiner Modelling 49

3.2.3. Dynamic Modelling 50

3.3. Research Strategy 51

3.3.1. Data Collection 51

3.3.2. Data Pre-Processing 53

3.3.2.1. Data Imputation 53

3.3.2.2. Data Normalisation 54

3.3.2.3. Feature Selection 55

3.3.3. Data Splitting Method 56

3.3.3.1. Holdout Technique 57

3.3.3.2. K-Fold Cross-Validation 57

3.4. Model Evaluation Techniques 58

3.4.1. Confusion Matrix 58

3.4.1.1. Sensitivity and Specificity Ratios 59

3.4.1.2. Type I and Type II Error 59

3.4.1.3. Average Accuracy Ratio 59

3.4.1.4. F-Measure 60

3.4.2. Area Under Curve 60

3.4.3. Brier Score 61

3.4.4. Area Under Reliability Diagram 61

3.5. Statistical Significance Testing 62

3.6. Proposed Framework Research Design 65

3.7. Summary 65

Chapter 4 Audit Opinion Using Single Classifier Modelling 67

4.1. Introduction 67

4.2. Single Classifiers Development and Experimental Test Results in Classification Audit Opinion 67

4.2.1. Logistic Regression Model 67

4.2.2. K-Nearest Neighbour Model 70

4.2.3. Naïve Bayes Networks Model 72

4.2.4. Linear Discriminant Analysis Model 75

4.2.5. Decision Trees Model 77

4.2.6. Artificial Neural Network Model 80

4.2.7. Support Vector Machine Model 83

4.2.8. Boosting Ensemble Classifier Model 85

4.2.9. Deep Learning Model 88

4.3. Comparative Analysis and Discussion 91

4.4. Statistical Significance Testing 95

4.5. Summary 99

Chapter 5 Committee Machine Classifiers Models Combiner 101

5.1. Introduction 101

5.2. Combination Method Development 101

5.2.1. Consensus Combiner Model 102

5.2.1.1. Measuring Classifier Rankings and Constructing Decision Profiles 102

5.2.1.2. Measuring Uncertainty Assessment Matrix of Classifiers 102

5.2.1.3. Measuring Classifier Weights 104

5.2.1.4. Update Calculations and Final Decision 104

5.2.2. Fuzzy Logic Combiner Method 106

5.2.2.1. Building Fuzzy Logic Combiner Method 106

5.2.2.2. Fuzzy Logic Membership Functions and Rules Processing Inference... 107
5.2.2.3. Fuzzy Logic Combiner Operation ... 108
5.2.2.4. Defuzzification ... 108
5.2.3. Average Method ... 110
5.2.4. Weighted Average Method ... 111
5.2.5. Median Method ... 112
5.2.6. Majority Voting Method (MajVot) ... 113
5.2.7. MIN Method ... 114
5.2.8. MAX Method ... 116
5.3. Combination Method Results ... 117
5.3.1. Consensus Combiner Model ... 117
5.3.2. Fuzzy Logic Combiner Model ... 120
5.3.3. Average Method ... 122
5.3.4. Weighted Average Method ... 124
5.3.5. Median Method ... 127
5.3.6. Majority Voting Method ... 129
5.3.7. MIN Method ... 131
5.3.8. MAX Method ... 133
5.4. Comparative Analysis and Discussion ... 135
5.5. Statistical Significance Testing ... 142
5.6. Classification Model Training Time ... 145
5.7. Summary ... 146
Chapter 6 Dynamic Modelling ... 148
6.1. Introduction ... 148
6.2. Dataset Preparation ... 148
6.3. Dynamic Modelling Development and Experimental Results ... 149
6.3.1. Deep Learning LSTM (DPL-LSTM) ... 149
6.3.2. Nonlinear Autoregressive Exogenous (NARX) Neural Network ... 152
6.3.3. Nonlinear autoregressive (NAR) neural network ... 156
6.4. Comparative Analysis and Discussion ... 160
6.5. Dynamic Modelling Training Time ... 163
6.6. Summary ... 164

Chapter 7 Conclusion and Future Work 166
7.1. Main Study Outcomes 166
7.2. Research Framework 166
7.3. Empirical Contributions 167
7.3.1. Improved Simulation and Computing Platform 167
7.3.2. Improved Performance Accuracy 167
7.3.3. Applying Three Dynamic Modelling Techniques 168
7.4. Limitations 168
7.4.1. Literature 168
7.4.2. Methodology 169
7.5. Directions for Future Work 169
References 170

List of Tables

Table 2.1: Related studies 21

Table 3.1: Summary of all datasets 53

Table 3.2: Independent variables (n = 34) after feature selection 56

Table 3.3: K-fold cross-validation process 57

Table 3.4: Confusion matrix 58

Table 3.5: Transforming table of model outputs to rankings during Friedman test 64

Table 4.1: Evaluation test results of the LR model 68

Table 4.2: Evaluation test results of the K-NN model 71

Table 4.3: Evaluation test results of the NBN model 73

Table 4.4: Evaluation test results of the LDA model 76

Table 4.5: Evaluation test results of the DT model 78

Table 4.6: Evaluation test results of the ANN model 81

Table 4.7: Evaluation test results of the SVM model 84

Table 4.8: Evaluation test results of the BEC model 86

Table 4.9: Evaluation test results of the DPL model 89

Table 4.10: Comparing evaluation test results of all models with 2019 dataset 91

Table 4.11: Comparing evaluation test results of all models with 2018 dataset 91

Table 4.12: Comparing evaluation test results of all models with 2017 dataset 92

Table 4.13: Comparing evaluation test results of all models with All-Data dataset 92

Table 4.14: Friedman test results 96

Table 4.15: Pairwise comparison results for 2019 dataset 97

Table 4.16: Pairwise comparison results for 2018 dataset 97

Table 4.17: Pairwise comparison results for 2017 dataset 97

Table 4.18: Pairwise comparison results for All-Data dataset 98

Table 5.1: Evaluation test results of CON 118

Table 5.2: Evaluation test results of FC 120

Table 5.3: Evaluation test results of AVG 122

Table 5.4: Evaluation test results of WAVG 125

Table 5.5: Weighted average coefficients for nine classifiers across four datasets 125

Table 5.6: Evaluation test results of MED 127

Table 5.7: Evaluation test results of MajVot 129

Table 5.8: Evaluation test results of MIN 131

Table 5.9: Evaluation test results of MAX 133

Table 5.10: Comparing evaluation test results of all models with 2019 dataset 136

Table 5.11: Comparing evaluation test results of all models with 2018 dataset 137
Table 5.12: Comparing evaluation test results of all models with 2017 dataset 138
Table 5.13: Comparing evaluation test results of all models with All-Data 139
Table 5.14: Friedman test results 142
Table 5.15: Pairwise comparison results for 2019 dataset 143
Table 5.16: Pairwise comparison results for 2018 dataset 143
Table 5.17: Pairwise comparison results for 2017 dataset 143
Table 5.18: Pairwise comparison results for All-Data dataset 144
Table 5.19: Training time for single models in seconds 146
Table 5.20: Training time for committee combiner model in seconds 146
Table 6.1: Evaluation test results of DPL-LSTM model with benchmark model 151
Table 6.2: Evaluation test results of NARX model with benchmark model 155
Table 6.3: Evaluation test results of NAR model with benchmark model 159
Table 6.4: Comparing evaluation test results of dynamic models with benchmark model .. 160
Table 6.5: Dynamic and benchmark model training times 164

List of Figures

Figure 2.1: Audit phases. 8
Figure 3.1: DT structure 42
Figure 3.2: Topology of a feed-forward back propagation ANN 43
Figure 3.3: SVM model 45
Figure 3.4: Ensemble classifier framework 47
Figure 3.5: Topology for the DPL model 49
Figure 3.6: Deep learning framework for time series modelling 51
Figure 3.7: Proposed framework research design 66
Figure 4.1: ROC curves for the LR model over four datasets 69
Figure 4.2: LR reliability diagrams over four datasets 69
Figure 4.3: ROC curves for the K-NN model over four datasets 71
Figure 4.4: K-NN reliability diagrams over four datasets 72
Figure 4.5: ROC curves for the NBN model over four datasets 74
Figure 4.6: NBN reliability diagrams over four datasets 74
Figure 4.7: ROC curves for the LDA model over four datasets 76
Figure 4.8: LDA reliability diagrams over four datasets 77
Figure 4.9: ROC curves for DT model over four datasets 79
Figure 4.10: DT reliability diagrams over four datasets 79
Figure 4.11: ANN structure 80
Figure 4.12: ROC curves for ANN model over four datasets 82
Figure 4.13: ANN reliability diagrams over four datasets 82
Figure 4.14: ROC curves for the SVM model over four datasets 84
Figure 4.15: SVM reliability diagrams over four datasets 85
Figure 4.16: ROC curves for the BEC model over four datasets 87
Figure 4.17: BEC reliability diagrams model over four datasets 87
Figure 4.18: ROC curves for DPL model over four datasets 90
Figure 4.19: DPL reliability diagrams over four datasets 90
Figure 4.20: Bonferroni-Dunn correction for Individual classifier models, with significance levels 99
Figure 5.1: Two fuzzy sets with different widths and centre positions 108
Figure 5.2: AVG method mechanism 111
Figure 5.3: WAVG method mechanism 112
Figure 5.4: MED method mechanism 113
Figure 5.5: MajVot method mechanism 114

Figure 5.6: MIN method mechanism 115
Figure 5.7: MAX method mechanism 117
Figure 5.8: CON ROC curves over four datasets 119
Figure 5.9: CON reliability diagrams over four datasets 119
Figure 5.10: FC ROC curves over four datasets... 121
Figure 5.11: FC Reliability diagrams over four datasets 121
Figure 5.12: AVG ROC curves over four datasets.. 123
Figure 5.13: AVG reliability diagrams over four datasets.. 123
Figure 5.14: WAVG ROC curves over four datasets 126
Figure 5.15: WAVG reliability diagrams over four datasets 126
Figure 5.16: MED ROC curves over four datasets 128
Figure 5.17: MED reliability diagrams over four datasets 128
Figure 5.18: MajVot ROC curves over four datasets 130
Figure 5.19: MajVot reliability diagrams over four datasets 130
Figure 5.20: MIN ROC curves over four datasets... 132
Figure 5.21: MIN reliability diagrams over four datasets... 132
Figure 5.22: MAX ROC curves over four datasets.. 134
Figure 5.23: MAX reliability diagrams over four datasets.. 134
Figure 5.24: Bonferroni-Dunn correction for Committee combiner models, with significance levels. 145
Figure 6.1: ROC curves for DPL-LSTM model with benchmark model 151
Figure 6.2: Reliability diagrams for DPL-LSTM model and benchmark model 152
Figure 6.3: NARX view command 153
Figure 6.4: NARX closed-loop.. 154
Figure 6.5: NARX predicted one step ahead 154
Figure 6.6: ROC curves for NARX model with benchmark model................................. 155
Figure 6.7: Reliability diagram for NARX model and benchmark model 156
Figure 6.8: NAR topology.. 157
Figure 6.9: NAR topology with closed-loop... 158
Figure 6.10: NAR topology with one step ahead prediction 158
Figure 6.11: ROC curves for NAR model with benchmark model 159
Figure 6.12: Reliability diagrams for NAR model with benchmark model....................... 160
Figure 6.13: ROC curves for three models with benchmark model............................... 161
Figure 6.14: Reliability diagrams for three models with benchmark model 161

List of Equations

3.1……48

3.2……48

3.3……48

3.4……48

3.5……48

3.6……48

3.7……42

3.8……42

3.9……43

3.10……43

3.11……38

3.12……38

3.13……39

3.14……40

3.15……45

3.16……45

3.17……54

3.18……55

3.19……59

3.20……59

3.21……59

3.22……59

3.23……60

3.24……60

3.25……61

3.26……62

3.27……63

3.28……64

4.1……80

5.1……102

5.2……102

5.3……103

5.4……103

5.5……103

5.6……………………………………………………………………………………...103
5.7……………………………………………………………………………………...104
5.8……………………………………………………………………………………...104
5.9……………………………………………………………………………………...104
5.10………………………………………………………………………………………104
5.11………………………………………………………………………………………105
5.12………………………………………………………………………………………105
5.13………………………………………………………………………………………105
5.14………………………………………………………………………………………105
5.15………………………………………………………………………………………107
5.16………………………………………………………………………………………107
5.17………………………………………………………………………………………107
5.18………………………………………………………………………………………108
5.19………………………………………………………………………………………108
5.20………………………………………………………………………………………109
6.1……………………………………………………………………………………...152
6.2……………………………………………………………………………………...156

List of Abbreviations

AC	Average Accuracy
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
AURD	Area Under Reliability Diagram
AVG	Average
BD	Big Data
BDA	Big Data Analytics
BEC	Boosting Ensemble Classifier
CD	Critical Difference
CFS	Combined Fuzzy Set
CL	Confidence level
CM	Committee combiner modelling
CoG	Centre of gravity
CON	Consensus Combiner Model
CP	Classifier predictions
CSD	Confidence standard deviation
DM	Data Mining
DP	Decision Profile
DPL	Deep Learning
DT	Decision Trees
FAME	Financial Analysis Made Easy
FC	Fuzzy Logic Combiner Method
FN	False Negative
FP	False Positive
GAAP	Generally Accepted Accounting Principles

IFRS	International Financial Reporting Standards
IoT	Internet of Things
IT	Information technology
K-NN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LoM	Largest of maximum
LR	Logistic Regression
LSTM	Long Short-Term Memory
MajVot	Majority Voting
MAX	Max
MCC	Mathew’s Correlation Coefficient
MED	Median
MIN	Min
MLP	Multilayer Perceptron
MoM	Middle of maximum
MSR	Mean square root error
NAR	Nonlinear Autoregressive
NARX	Nonlinear Autoregressive Exogenous
NBN	Naïve Bayes Network
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
OP	Optimal mean
SoM	Smallest of maximum
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
P-value	Critical value
TN	True Negative
TP	True Positive

WAVG	Weighted Average
X_F^2	Friedman test

Chapter 1
Introduction

1.1. Background

Financial statement auditing necessitates evidence related to non-financial and financial data. These pieces of evidence are evaluated and tested to produce and achieve a final audit opinion about the credibility of a financial statement (Arens *et al.*, 2020). Currently, information technology advances are providing new techniques, such as data mining (DM) models, to help auditors analyse the new data types like unstructured data, which is non-traditional data, with traditional data, in order to obtain greater confidence in the reliability of audit opinion.

DM modelling is one of the emerging new technologies that is fast becoming the most significant method for providing decision support. The DM process is a data analysis method that uses a set of mathematical, machine learning and artificial intelligence (AI) processes. DM is used to elicit valuable information from input data recorded in operational information systems that can then be utilised in enhancing the decision-making process (Sharma and Panigrahi, 2012; Özdağoğlu *et al.*, 2017). Conventional auditing data analysis tools cannot manage Big Data (BD). However, the DM models can provide solutions for auditors to deal with it in the form of data analysis tools that can provide classification of different types of data or allow for prediction of information one year in advance (Khemakhem and Boujelbene, 2018). DM has been receiving increasing attention in auditing decision for enhancing the competence of managing the reliability and quality of audited opinion report's decision through the creation of efficient classification models.

In recent decades, researchers have contributed to the development of predictive models for creating audit opinions, and they have relied on using several different statistical modelling approaches, such as linear discriminate analysis and logistic regression (Dopuch *et al.*, 1987; Krishnan and Krishnan, 1996; Francis and Krishnan, 1999). Recently, researchers that have started receiving attention, have demonstrated and illustrated the raised preference for modern AI techniques, such as decision trees (DTs), neural networks (NNs) and support vector machines (SVMs) (Kirkos *et al.*, 2007; Gaganis *et al.*, 2007; Tsai, 2008; Ravisankar *et al.*, 2011; Saif *et al.*, 2012). The employment of such techniques for structuring audit opinion models has evolved over time, with most of the studies initially tending to utilise them individually, then later, to overcome the shortcomings of each model being used, with researchers tending to customise a design audit opinion model. In recent years, researchers have tended to increase the complexity in their model design by utilising, such arrangements as committee combiner modelling combined with several individual classifiers. This combination the committee method presented better performance compared with the single model. Nevertheless, previous studies supposed that complexity could lead to the best and inclusive classification models for auditing opinion model, which is one of the prime aims of this research.

Generally, previous studies' results have revealed no superior classification model that can be used in the audit opinion decision and decision sport. Choosing a classifier depends upon the variables used, availability in the environment and market, data structure and nature of the issue of interest (e.g. Tsai, 2008, 2014; Sivasankar *et al.*, 2020). DM classification modelling, specifically the committee combiner method, remains under researched in the auditing field. Hence, further investigation is warranted to develop classification models that can be used to identify appropriate auditing opinion.

1.2. Research Motivation

In recent years, DM methods have emerged to play the primary role in financial and accounting areas. The classification power provided via these tools has proven their efficiency when tested in the financial and accounting fields. To date, this association between auditing and DM model is in its early stages and initial applications of it in auditing remain in their infancy (Cao *et al.*, 2015). Previous studies have been actively moving towards building and using single classification models. Now, researchers have started to call for using these multiple single classifier models with committee combiner methods. The idea behind this is that a combination of several individual classifiers, and a committee combiner method outperforms individual classifiers being used separately, as the final classifier can minimise the impact of the utilised single classifiers' errors.

Most previous studies in the auditing field have been focused on using homogenous classifier ensembles (combining classifiers with the same base learning algorithm). However, a few studies have used heterogeneous committee classifiers (combining classifiers with several base learning algorithms) through the traditional combination of rules and basic fusion techniques, such as average, weighted voting, weighted average, majority voting, random forest and stacking (Song *et al.*, 2014; Tsai, 2014; Fernández *et al.*, 2018; Stanišić *et al.*, 2019; Hooda *et al.*, 2020; Sivasankar *et al.*, 2020; Kiziloz, 2021). These related studies have illustrated that committee combiner methods obtain better results than the individual classifier model in the auditing field and related areas. In addition, the majority of researchers in the auditing field have used one committee combiner method and only a few have used two, whilst Kiziloz (2021) employed six. In the committee combiner method, all individual models are trained independently to process their decisions, which are then merged through a heuristic algorithm to arrive at a final outcome.

However, to date, researchers in the auditing field working on developing new classification models are doing so without investigating the application of DM of the dynamic modelling to use it as an auditing model to produce audit opinion in advance. Consequently, no related study has discussed or developed dynamic modelling or warranted further research to introduce dynamic modelling in the auditing field.

1.3. Aims and Objectives of the Research

To this researcher's knowledge, there have been few attempts to utilise the power of DM algorithms in the auditing literature, which thus, represents a gap that needs to be filled. Hence, the first aim of this thesis to explore and develop several individual classifiers and compare novel classification deep learning model with traditional classifiers. The secondary aim is the improvement of traditional committee combiner methods and then, introducing two new combiner methods – consensus model and fuzzy logic combiner - together as an individual classifier. This new combiner classifier is to be compared with the traditional committee combiner models in anticipation that the former performs better at classifying audit opinion. The third aim is to develop a dynamic modelling in the auditing field that can be used as an auditing opinion model to predict audit opinion in advance. Such dynamic modelling has not been applied and tested by any of the related previous studies in the auditing field.

The fourth aim is probing the question as to whether intricacy in modelling auditing opinion model is worth investigating involving sundry phases for achieving the primary aim of this thesis. The proposed approach includes phases, beginning with collecting datasets, pre-processing, splitting datasets, clustering datasets, developing models (simple classifiers, dynamic modelling, and a complex classifier model using committee combiner models), in sequence. The models are then evaluated using nine parameters and statistical testing, after which comparisons of the experimental results are made to obtain the best outcomes in terms of efficiency and effectiveness.

The objectives of this thesis are as follows.

1. Apply two new evaluation measurement parameters, Brier score and area under reliability diagram, to evaluate model performance.
2. Develop and test deep learning classifier model with other eight single classifiers: support vector machines; artificial neural networks; K-nearest neighbour; decision trees; naïve Bayes network; logistic regression; linear discriminant analysis; and

boosting ensemble. Then, there is comparison of the evaluation results achieved from each classifier’s performance to help auditors select a suitable DM classifier to classify correct audit opinion and to determine whether the novel deep learning classifier has better performance compared with other classifiers.

3. Improve upon six traditional committee combiner methods (average; weighted average; median; minimum; maximum; and majority voting) and introducing two new committee combiner methods (consensus model and fuzzy logic combiner model) to the auditing field. Additionally, to present the best method to increase the accuracy of the classification auditing opinion modelling in comparison with other committee combiner models.
4. Develop three dynamic models (nonlinear autoregressive exogenous; nonlinear autoregressive and deep learning-LSTM) to be utilised to predict audit opinion one year in advance. Then, two processes are carried out to evaluate the model results: first, each model results are compared with results of the benchmark classifier model (deep learning classification model tested on the original dataset for one year in advance). Second, there is comparison of the results achieved from each dynamic model’s performance, which will allow for the identification of the model that can predict auditing opinion in advance most correctly.

1.4. Contributions to Knowledge

In this thesis, sundry algorithms are enhanced and developed to enhance the performance of individual, committee combiner models and dynamic models. The main contributions of this thesis are as follows.

1. A critique of related studies on various DM models as committee combiner techniques and individual classifiers, by considering several aspects of their modelling approaches for the period 2007–2021.
2. Introducing two new evaluation parameters (Brier score and area under reliability diagram) to evaluate the model performance, which has not been used in any previous auditing studies or in any related area.
3. Applying a new single deep learning classifier to a classification audit opinion model and comparing its performance to other classifier models. Whilst a deep learning model has not been applied in any prior auditing opinion model, a few researchers investigated this model in the auditing field.
4. The application of committee combiner models in the auditing field and related areas remains in its infancy and previous studies have only used majority voting as a

combining method. In this thesis, several traditional combiner methods are applied with a new one as a consensus model and fuzzy logic combiner model.

5. Dynamic modelling is explored and improved upon in terms of its utilisation as an auditing opinion model to predict audit opinion in advance.

1.5. Research Methodology

To achieve the aims and objectives of this thesis, the proposed research framework has included seven sequential stages. Four datasets drawn upon to structure and validate the individual classifiers and committee combiner model and another four datasets to structure and validate the dynamic modelling. These datasets were collected from companies across all industries in Ireland and UK using the Financial Analysis Made Easy Database (FAME) software. The datasets were applied to three pre-processing methods, namely, data imputation, data normalisation and features selection processing. After the data pre-processing, each of the four datasets was divided into a testing and training set, which were used to structure and assess the individual classifiers, and committee combiner. Clustering another four datasets which were used to structure and assess the model dynamic modelling. Thereafter, several evaluation measurement techniques (average accuracy; AUC; specificity; sensitivity; Type II Error; Type I Error; F-measures; Brier score and area under reliability diagrams) were applied to evaluate the performance of each model. Last, significance testing was conducted to valid model performance statistically.

1.6. Thesis Outline

Chapter 2 presents the background and a literature review of audit opinion. The theoretical background section is presented in two parts: Firstly, financial statement audit is covered in terms of definition, audit phase, audit evidence and analytical review. Secondly, the effects of the development of information technology as BD and DM models are explored along with the challenges faced by auditors when applying BD and DM models to financial statement auditing. The next section presents a review of the literature related to previous studies on auditing models of the proposed modelling approach in this thesis. Then, critical analysis of these selected related works is provided tracked by drawings and findings. Finally, highlights the gaps in the reviewed related works are identified.

Chapter 3 presents the research methodology design adopted. The process is explained in phases, with each discussing the different problems that need addressing to obtain a better modelling approach, thereby delivering a more accurate model.

In Chapter 4, illustrates the developments in the performance of nine classifiers (deep learning; support vector machines; artificial neural networks; K-nearest Neighbour; decision trees; naïve Bayes network; logistic regression; linear discriminant analysis; and boosting ensemble). The results obtained from each classifier are presented, followed by a comparison of their performance in terms of the statistical test results and nine evaluation parameter results by each to determine which can classify audit opinion most correctly.

Chapter 5 introduces each of the six traditional committee combiner methods (average; weighted average; median; min; max and majority voting) and two new committee combiner models (consensus model and fuzzy logic combiner model) for use in combination for the predictions for each single classifier used in Chapter 4. Then, the experimental results obtained from each committee combiner modelling method after being tested individually are presented and discussed. The final step is to determine which one has the best ability to enhance the accuracy of single classifiers in terms of providing the correct audit opinion. This involves two processes:1) comparison of the experimental results for the nine evaluation parameters for the committee methods; and 2) comparison of the statistical results for the committee combiner models.

Chapter 6 presents the developments in dynamic modelling performance, including nonlinear autoregressive exogenous (NARX), nonlinear autoregressive (NAR) and deep learning-LSTM for predicting audit opinion in advance for year 5 (2019). Then, the evaluation results of each model are presented and compared with the benchmark model of this chapter, which is the deep learning classification that tests actual audit opinion for the year 2019. In the final step, comparison is made between nine evaluation parameters' experimental results for the three models and benchmark model.

Chapter 7 provides a summary of the thesis, including the main conclusions drawn from the experimental results in terms of the novel contributions to the field. Moreover, the limitations are discussed and proposals for future potentially beneficial research directions are made.

Chapter 2
Theoretical Background and Literature Review

2.1. Background

This chapter provides the theoretical background in relation to audit opinion decision in terms of financial statement auditing definitions, auditing phases and evidence, and analytical review. It explores the impact of BD and DM models on the process of audit opinion decision, in terms of the challenges that auditors face in application, according to extant literature. Recently, some studies have investigated how the processing of BD through DM models can enhance the audit process in relation to auditor prediction and identifying classifiers that can deliver correct audit opinion. However, the application of these technologies to the auditing process is still in its infancy. Most previous studies have focused on investigating the ability of DM techniques to classify fraud, and there have been a small number that have contributed to developing classification models for decision making audit opinion. In order to achieve the aims of this study, this chapter develops an appropriate theoretical framework based on the review of related literature pertaining to the use of DM models in auditing and related areas. It analyses and discusses such studies to present the conceptual framework of this study, tailored to achieving the study aims.

2.2. Auditing Opinion Decision

2.2.1. Definition of Financial Statement Audit

During the early 20th century, auditing reporting was increasingly subject to institution obligations and outcomes, and a national and international process of standardisation in independent reporting. In recent years, auditing has evolved through the development of practices to ensure financial accountability and to detect fraud or error, and providing consultancy services as feedback about institutional financial data (Baharud-din *et al.*, 2014; Chan and Vasarhelyi, 2018).

Financial statement auditing is the process of searching for evidence relating to management assertions (i.e., reports) in financial statements and non-financial data (Arens *et al.*, 2020; BPP Learning Media, 2020). This evidence is then objectively evaluated in order to come up with an opinion about the credibility and the extent to which these assertions match with the accounting criteria. Subsequently, the auditor publishes a report about the level of the quality and reliability assurance of the financial statement that is free of material misstatement, whether caused by fraud or error.

The audit report delivers the auditor's opinion about the fairness and credibility of the financial statement. There are two main types of audit opinion, qualified and unqualified. Auditors' published unqualified opinion is also called the clean report. In this report, the auditor states that the financial statement for the firm is credible, presented correctly and free from material misstatements (Arens *et al.*, 2020). However, when there is sufficient audit evidence that have been financial misstatements and/or the company has not presented financial records in line with accounting standards, then a qualified audit opinion is given, and no clean report is provided (Millichamp and Taylor, 2018).

2.2.2. Audit Phases

The auditor is required to issue an audit opinion about the fairness and credibility of financial statements, based on analysis via several phases, in order to reach a reasonable opinion regarding the fairness of statements (Arens *et al.*, 2020). Figure 2.1 shows a chart of the audit phase process.

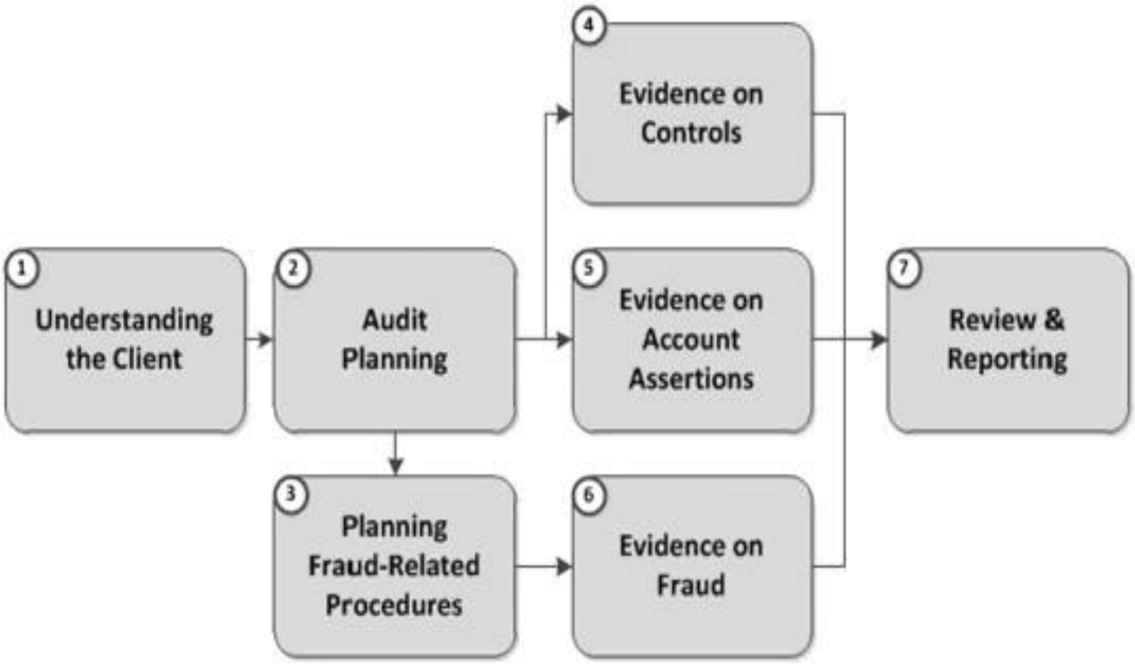


Figure 2.1: Audit phases.

Source: Gray and Debreceny (2015, p. 364) and Debreceny and Gray (2011, p. 207)

Auditing has been enhanced with improved analysis tools, sizes of sample audits, and increasing types of data and sources used during the audit processes (Kotsiantis *et al.*, 2006; Appelbaum, 2016; Ala'raj and Abbod, 2016; Issa *et al.*, 2016; Mentz *et al.*, 2018). For example, Issa *et al.* (2016) stated that in the substantive test step process, auditors have the ability to test the whole of the data of evidence, rather than a test sample; and in the pre-planning phase, the auditor is now able to obtain a large volume of client-related evidence to assist in making correct decisions.

2.2.3. Audit Evidence

Audit evidence refers to all the information utilised by the auditor from internal and external sources, which may be oral or written. This enables the auditor to make a professional judgment (audit opinion) as to whether the financial statements give a fair view (Rashid, 2017). The internal sources are obtained from the client (i.e., institution), such as accounting data relating to financial statements, including accounting records and information relating to the company, like management and investment data. External sources pertain to the data that the auditor needs from outside of the institution to help in checking the credibility of the financial statement, including those from customers and suppliers (Hayes *et al.*, 2014). The Association of Chartered Certified Accountants asserts that audit evidence types include analytical, documentary, electronic, and physical (Arens *et al.*, 2020; BPP Learning Media, 2020), all of which needs to be characterised in terms of sufficiency, reliability, and relevance (Bhattacharjee *et al.*, 2012; Zuca, 2015; Efiong *et al.*, 2017; Abdul Rahim *et al.*, 2017; Mentz *et al.*, 2018; Millichamp and Taylor, 2018; Gospel *et al.*, 2019). The availability and successful analysis of such evidence leads to a high degree of quality and levels of assurance and credibility in the audit process; conversely, inappropriate evidence or analysis leads to inappropriate conclusions, and wastes in costs and time, undermining the efficiency and effectiveness of the audit process quality.

2.2.4. Analytical Review

The analytical review is a compulsory phase of audit process planning, pertaining to analyse of the relationship between the items of the financial statement and non-financial ones for the same period, which are then compared for both types of information for previous periods. This is to determine the degree of homogeneity between the information and to identify unexpected relationships. The objective of the analytical review at any stage of the audit process is to provide guidance regarding the required evidence, which enables accurate final results for the audit process, thus informing a correct audit opinion concerning financial statement data (Eilifsen, 2010; Millichamp and Taylor, 2018).

The analytical review facilitates the auditor in understanding client activity, thus allowing for assessing the risks and identifying any misstatement in the financial statement (Mentz *et al.*, 2018). Lina *et al.* (2003) and Rose *et al.* (2020) contended that producing more analytical procedures will enhance audit quality, such as more accurate evaluation of risk and relative evidence in assessment operations, which affects the results of analytical review. When the level risk and the importance of the evidence are higher, then the auditor will have more reliance on the analytical review results, focused on detailed audit procedures. Likewise, the results of the analytical review results can be compatible with the expected results of the review.

2.3. Audit and Information Technology

In the previous decade, information technology (IT) has dramatically affected businesses due to its increased capacity to store, capture, analyse, and process BD forms of information. It is widely used in different fields, including accounting, auditing, and other finance-related business operations, including a range of activities such as planning, documentation, and operations, aimed at enhancing performance and efficiency (Mustapha and Lai, 2017). In the context of this thesis, IT has dramatically affected the audit profession by affecting the size of possible evidence selected for analysis, augmenting auditor skills, improving the quality of the audit opinion process, and increased knowledge participation (Ashraf *et al.*, 2020).

With the increased deployment of IT developments in the auditing profession, standards were issued to guide auditors (Curtis *et al.*, 2009). Regarding the extent of the impact of IT on the audit profession, even small auditing companies have begun to utilise advanced technologies to automate their work, thereby simplifying documentation processes (Li *et al.*, 2018). Previous studies have discussed the importance of the utilising advanced IT tools in auditing, but adoption in practice remains relatively limited, notably: (e.g. Bierstaker *et al.*, 2014; Dutta *et al.*, 2017; Mustapha and Lai, 2017; Al-Hiyari *et al.*, 2019; Ashraf *et al.*, 2020). Jahani and Soofi (2013) and Janvrin *et al.* (2008) stated that empirical evidence demonstrates that auditors can utilise IT effectively and it can significantly affect the audit process, particularly by rendering more beneficial and clean data, which is germane to improved final decisions (audit opinion) arising from the audit process. The findings suggested that while auditors use IT widely for analytical procedures, it is not employed extensively for other aspects of auditing, such as digital analytics.

IT auditing analysis has to some extent been spearheaded by client-led IT adoption, as the wholesale use of IT systems in firms' commercial operations, integrated with their own internal accounting (and supply chain management etc.), produces IT-based and automated accounting data that is used by auditors. The Public Company Accounting Oversight Board's auditing guidelines and standards request auditors to increasing utilisation of IT tools, like DM model and BD, to improve the effectiveness and quality of the auditing process (Bradford *et al.*, 2020). Previous studies suggested that auditors need to increase their use of new data analysis tools, such as DM and BD analytics (BDA), in their audit process, as these tools can help increase the size of evidence deployed, thus informing better quality and more accurate audit opinions (Titera, 2013; Mustapha and Lai, 2017; Ashraf *et al.*, 2020).

According to this suggestion, this section presents auditing-related IT techniques (concerning BD and DM) and their implications for auditing financial statements, noting their advantages and limitations, and challenges for the auditing process.

2.3.1. Big Data and Audit Process

BD is an inherently new and multidisciplinary field, which is reflected in varying definitions of the concept, based on its properties known as the seven Vs: variability, variety, volume, velocity, veracity, visualisation, and value (George *et al.*, 2016; Seddon and Currie, 2017; Aryal *et al.*, 2018; Mikalef *et al.*, 2018). Gandomi and Haider (2015) defined BD as having high velocity, high variety, and massive size. As such, it requires cost-effective innovative treatment of assets to improve vision and decision making. Wu *et al.* (2014) stated that BD is essentially a method of data analysis made possible by recent technological advances, enabling the capturing of variable and complex (semi-structured, unstructured, and structured) data with a high velocity, thus facilitating its analysis, administration, distribution, and storage.

Evolution in communication and IT has led to dramatic increase in the volume of data sources available to accounting and financial analysis, in which regard BD can have major impacts on the efficiency and effectiveness of decision making (Earley, 2015). Auditors can utilise the advantages of BD in auditing, whereby it can assist auditors to achieve immediate results in a timely manner by improved digitalisation, storage, recuperation, and analysis of evidence data. On the other hand, the continual growth of BD features and capabilities poses challenges to traditional auditing, which is required to adapt to this new milieu, namely: (Tang and Karim, 2017). Zhang *et al.* (2015) and Richins *et al.* (2017) pointed out that BD provide unstructured data, which is non-traditional data, which can be utilised with traditional data by auditors to inform audit opinion about the credibility of financial statements. For instance, the validity of auditing evidence is improved by combining unstructured data with traditional data, because collecting several data types can display a more comprehensive overview of a firm’s actual profile than traditional (conventional) information. Likewise, external auditors can enhance fraudulent risk assessments through utilising BD financial fraud models, which progressively employ information from previous scams, helping provide valuable data and red flags for further analysis by auditors (Dechow *et al.*, 2011; Humpherys *et al.*, 2011).

In addition, BD can disseminate documentation as unstructured and unformatted data to auditors through using modern technologies such as smartphone and sensor devices. A PwC (2015) company report said that the development of technology has provided auditors with new tools to extract and visualise data, which allows the use of non-traditional data, which is not necessarily in the form of numbers. Data in different forms, including images or words obtained from various sources, can enable more complex data analysis. For example, the Internet of Things (IoT) can provide BD from different sources to describe the behaviour of the firm's environment (Bandyopadhyay and Sen, 2011). This kind of BD collection can improve decision making through completing data from IoT applications, so that information derived from analysis of such data can be used to improve audit processes phases (Risteska and Trivodaliev, 2017). Auditors can employ the infrastructure of the IoT to collected and different data in real time (Brown-Liburd *et al.*, 2015).

In the light of the high speed and the large volume of BD features, the auditing process will increasingly have the ability to implement and analyse diverse forms of data more quickly and accurately compared to analysis of traditional sources alone. BD can be fairly reliable for auditors, because BD itself is often generated from external sources, and obtained by the auditor directly. On the other hand, the high speed and massive volumes of data can create a gap between the requirements of BDA and the audit analytics process (Vasarhelyi *et al.*, 2015). In addition, massive amounts of data create challenges for auditors in the form of a lot of noise and messy data, which need to be addressed so as not to affect the integrity and quality of the audit evidence (Earley, 2015). Likewise, the veracity of automatic BD red flags can be suspect, as indicated by large rates of false positives, which reduce reliability in audit decisions based on such data (Ramlukan, 2015; Yoon *et al.*, 2015). Modern companies can easily assemble vast amounts of data, but the greater the volume, the greater the difficulty they face in formatting and analysing it (Earley, 2015).

Many previous studies highlighted the need to conduct research on the development of prediction auditing models in order to predict complicated and sophisticated fraud, due to BD models having the ability to process diverse information beyond the scope of traditional regression models (Hogan *et al.*, 2008). There have been several studies demonstrating BD-based fraud detection methods utilising analytics (e.g., Ramona *et al.*, 2014; Chen and Wu, 2017; Sathyapriya and Thiagarasu, 2017; Carcillo *et al.*, 2018). Chen and Wu (2017) who examined the attributes of the diversity and value of BD in the economy and investment area using fraud detection models for the financial statements of commercial firms. They found that BD attributes led to increased ability of models to classify fraud and decreased risk or investment losses, thereby improving decision making for creditors and investors.

However, the application of BDA in the auditing profession is still in its infancy, as a result of which auditors continue to treat it as a new phenomenon (Salijeni *et al.*, 2019). The lack of wholesale practical application increases auditing challenges and limits understanding of how final auditing decision quality is impacted by BDA. The implication of BDA is that auditing firms will be able to improve their practice, by using it to "identify the data, assess their suitability for the task at hand, and decide whether the analyses should be outsourced" (Warren *et al.*, 2015, p. 44). Auditors may be vulnerable to a blame culture in BDA adoption and application in the event that they fail to identify fraud or errors; under the traditional paradigm of audit analysis, auditors analyse data samples, which can inherently lead to the possibility of erratic data not being detected, and this is beyond the control of the auditor; however, BDA allows auditors to analyse all company data comprehensively, thus they may incur unprecedented liability in the case of controversial or erroneous decisions (Cao *et al.*, 2015).

In addition to using BDA with conventional audits, its data management capacity could enable the analysis of multiple audits across the whole portfolio to identify trends, as well quality issues, and outliers (Ramlukan, 2015). The most recent auditing standards place increased responsibility on auditors to detect whether there has been any fraud in financial statements (Wang, 2010). BDA can make it easier for auditors to identify fraudulent activity or errors that would have slipped through the net under traditional sampling methods (Yeo and Carter, 2017). Appelbaum (2016), Tang and Karim (2018), and Kaplan *et al.* (2012) suggested using BDA in the process of detecting fraud, because this can provide the ability to find data in quick and clever ways, which allow for auditors to interpret evidence related to the risk of material misstatement and fraud in the financial statement efficiently. BDA allows auditors to analyse a whole population testing data, which can lead to decreased risks related to data. Likewise, BDA can decrease challenges faced when auditor may conclude value from BD, and guarantee auditors' decisions based on data having high credibility and relevance (Brown-Liburd and Vasarhelyi, 2015).

While there is general consensus on the potential applications of BD in auditing, Warren *et al.* (2015) noted that the new information value derived from BD is wholly dependent on the analysis undertaken on it by auditors; if this is not effective, the value of BD in itself is zero. Responding to the need to improve audit methodology and practice, recent papers on auditing have investigated the significance BDA in the areas of financial statement performance and risk assessment, namely: (Brown-Liburd *et al.*, 2015; Yoon *et al.*, 2015; Cao *et al.*, 2015; Appelbaum, 2016; Alles and Gray, 2016; Yeo and Carter, 2017; Gepp *et al.*, 2018; Salijeni *et al.*, 2019; Balios *et al.*, 2020; Kend and Nguyen, 2020). Cao *et al.* (2015) demonstrated how BDA can be applied in auditing by discussing the attributes of

BDA that distinguish it from conventional auditing. Salijeni *et al.* (2019) explored the integration of BDA into auditing methodology through 22 interviews with people who had considerable experience in evolving, assessing, or implementing the effect of BDA in auditing, using documentation on BDA published in an audit domain and general review of BDA-related variations in audit formwork. They elicited that the auditing process must be developed based on BDA method enhancement, and they stated that auditors need more understanding of the essential relevance and important of BDA in the auditing process. Alles and Gray (2016) stated that different data analytical tools like DM techniques have the ability to analyse BD in a highly beneficial way for auditors.

2.3.2. Data Mining Analysis in Auditing Decision

New auditing information from BD is worthless for auditors without effective analysis, such as that undertaken through DM (Balios *et al.*, 2020). DM uses a set of mathematical, machine learning, statistical, and AI processes to elicit valuable information to identify interesting patterns in databases that can be utilised in the decision-making process. DM analysis tools that have made the considerable contribution to a decision in various science disciplines through clustering, optimisation, prediction, visualisation, and classification analyses (Sharma and Panigrahi, 2012; Özdağoğlu *et al.*, 2017). DM consists of six phases: collecting data, data planning, model development, model estimation, post processing and publication (Cho *et al.*, 2020).

In addition, Amani and Fadlalla (2017) and Cho *et al.* (2020) contended that DM models can potentially benefit all phases of the audit procedure, from clean data to audit reporting. DM has been increasingly utilised in advanced auditing throughout auditing phases, predicting audit opinion outcomes. In the pre-planning audit stage, DM can help ensure more reliable and unbiased data collection, though used evidence data are related to a customer company structures like commercial operations, and how these data are related to financial statements. In the risk assessment stage, DM algorithms enable auditors to determine data type, timing, operation, operational techniques, and financial and accounting systems used. Likewise, through utilising further and good quality data, auditors are able to specify outliers with the assistance of DM derivate pattern recognition, discovering irregularities via comparison between forecasting data created via DM and original data.

The main difference between classic data analysis and DM is that the former assumes that suppositions have already been made and their credibility checked against the data, whilst with the latter, suppositions and patterns are automatically extracted from data. Lin *et al.* (2015) and Zerbino *et al.* (2018) reviewed conventional audit analysis, and concluded that it fundamentally involves ratio analysis, which has achieved rather limited success in recognising fraud, because one of the issues with utilising ratio analysis is the associated subjectivity involved in the recognition of the consistency of ratios between various statement values as being probable to detect fraud in financial statements. DMs deploy an iterative procedure to detect different correlations, by either manual or automatic techniques. DM models help resolve non-linear issues by modelling massive data sets in several real-world implementations with a design relying on the instance of correlation between input and output data, in dramatically decreased time, namely: (Pumsirirat and Yan, 2018). Jahani and Soofi (2013) stated that relationships between financial statement items are non-linear, thus conventional analytical frameworks with linear models cannot accurately determine non-linear correlations between items. They contended that ANNs (ANNs) are effective models for uncovering correlations among different types of data, which is why they have played a significant role in enhancing the quality of financial decisions. This has further prompted researchers to identify the salient auditing factors pertaining to the application of ANNs.

Traditional data analytic models are not considered as useful to predict fraud due to making it difficult to predict or classify companies as fraudulent or not, due to proportional infrequency of fraudulent compared to non-fraudulent companies (Mohammadi *et al.*, 2020). Likewise, Mohammadi *et al.* (2020) and Albashrawi (2016) stated that an infrequency of observed fraud is relative to the large number of descriptive variables specified by previous fraud studies, which can result in overfitted production techniques with poor performance for forecasting new observations. Furthermore, they observed that most previous fraud studies dealt with all fraud cases homogenously, which can make fraud prediction or classification harder, because forecasting techniques should design models that deal with various deception types.

One of the main limitations auditors currently confront when using Computer Assisted Audit Tools is their poor ability to analyse whole BD input data, because such tools lack the ability to deal with data outliers (Zerbino *et al.*, 2018). This is inadmissible with regard to audit regulation standards; thus, auditors need the find tools to do the auditing process in a timely manner, with the ability to treat outliers and detect a number of financial statement frauds (Lin *et al.*, 2015). Likewise, the on-going concern opinion topic has become more complex, leading to numerous improved classification techniques to predict ongoing-concern decision opinion more easily and accurately, notably: (Carson *et al.*, 2013). Kirkos *et al.* (2007) and Fernández-Gámez *et al.* (2016) explored qualified opinion report classification using numerous DM models (DT, Bayesian belief network, multilayer perceptron, and probabilistic NN), to examine performance ability to classify case auditor published qualified reports. They found that DM classification techniques significantly improve performance in forecasting reliability and explanatory strength, due to this they suggest that these DM classifier techniques should be used as decision support tools by auditors.

Sun and Sales (2018), Sun and Vasarhelyi (2018), Sun (2019) and Zhang *et al.* (2018) reported that deep learning model has more powerful performance to classify or predict audit decision than traditional DM models, due to the ability to analyse different types of data (such as traditionally structured, unstructured, and semi-structured data). This increases auditors' ability to analyse different types of input data utilised to make decisions. Additionally, deep learning model can detect data characteristics by automatic readability, which enhances auditors' understanding of customers and assists in efficient risk assessment. Secondly, if the auditor uses big input data, the deep learning model illustrates superior capability to forecast and classify these data through the deep learning algorithms' performance, increasing data coverage for more comprehensive data analysis, with deeper insights for making decisions. The application of DM in accounting is specifically relevant in auditing during the initial stages of development, and auditors are cautiously exploring its effectiveness in the following areas (Sun and Vasarhelyi, 2018; Sun, 2019; Cho *et al.*, 2020):

- Prejudice can arise during any phase of DM, such as representation, evaluation, deployment, and measurement bias. For example, auditors can have assessment bias if DM model performance is not evaluated by utilising suitable evaluation measurements, such as in the case of unbalanced datasets being used to classify financial fraud cases. Average accuracy measurement can be inefficient because accuracy rates can reach high values in relation to low specificity rates, and vice-versa. Measurement bias increases if selected data characteristics do not mirror correct values of decisions, because auditors can leave out significant factors, or not filter out noise in the data, resulting in differential performance.
- Auditors still lack expertise in handling BD and modelling DM effectively.
- Auditors face challenges when choosing suitable DM models because of its significance for identifying unusual and unique data. They need to take into consideration numerous factors when choosing a DM model, such as the average accuracy, dataset size, and ability to enhance outcomes. In auditing, the main considerations faced when applying DM are frequent and automatic alterations in the DM model, and choosing alterations that can be suitable for auditing data.

There are also technological limitations of using DM in auditing. Firstly, the traditional auditing process lacks sophisticated use of new evaluation technology infrastructure, for extracting and analysing information in differing forms and structures (Hunton and Rose, 2010). Secondly, the auditing field has been hesitant to apply DM and DPL models, because they are viewed as black-box in nature, which is related to the inherent complexity of understanding and interpreting machine learning algorithms. To use such models, auditors need to process analytical audit evidence based on black-box algorithms to transfer data and produce the report. Auditors need to figure out the appropriate balance between their own decisions, and approving the output of these data analytic models (Sun, 2019). For example, Khemakhem and Boujelbene (2018) stated that ANN is will always be a black-box nature, which makes it hard to indicate the relevant correlation between basic standards of expounded decisions and variables involved. Based on this issue, ANN is still not applied widely in the business domain for the assessment of credit risk.

Previous studies have employed DM techniques for financial restatement, auditing selection, risk assessment and evaluation, audit opinion, and financial statement fraud detection (e.g. Kirkos *et al.*, 2008; Fernaˊndez-Gaˊmez *et al.*, 2016; Dutta *et al.*, 2017; Sánchez-Serrano *et al.*, 2020; Hamal and Senvar, 2021).

Hamal and Senvar (2021) tested the effectiveness of machine learning performance to classify fraud in financial accounting, comparing a dataset consisting of 1384 non-fraudulent financial statements and 321 fraudulent ones for Turkish small- and medium-sized enterprises from 2013 to 2017. They applied two steps: data pre-processing, followed by an evaluation of the results of NBN, SVM, K-NN, LR, ANN, Bagging, and Random Forest classification performance models through specificity rate, sensitivity rate, precision, average accuracy rate, ROC curve, and G-measure, then they compared the evaluation results for the seven models. Bagging and Random Forest classifier models were found to have better performance to classify fraud compared to NBN, SVM, K-NN, LG, and ANN; NBN model and K-NN achieved the worst performance evaluation.

Sánchez-Serrano *et al.* (2020) tested capability of the deep networks with convolution layers, multilayer perceptron (MLP), and the radial base function network to classify audit opinion financial statements correctly. They employed a dataset from Spanish companies comprising 87 qualified opinions with 211 unqualified ones for the year 2017. They used financial ratios and qualitative variables to find that the MLP model outperformed the others in terms of achieving the best parameter evaluation results (average accuracy rate, sensitivity, precision, and F-measure), over the training, testing, and validation datasets.

Hooda *et al.* (2020) utilised single classifiers (DT, SVM, ANN, NBN, ensemble models, Probit linear model, and decision stump) and traditional combiner rule (majority voting rule) for a decision model for external auditors to classify fraudulent companies correctly. The researchers used evaluation measurement parameters (TP, TN Type I and II Error, AC, F-score, AUC, and MCC) and statistical significance testing to validate model performance results. It was found that majority voting had the best ability to classify fraudulent companies correctly, compared to the other individual models' performances.

Dutta *et al.* (2017) enhanced NBN, Bayesian Belief Network, ANN, SVM, and DT models to classified financial restatements. The researchers used a dataset including audit analytical and financial data from 3,513 companies with financial restatement and 60,720 companies with non-restatement for the period 2001 to 2014. They tested the five models' performances over the whole dataset for the studied period in two phases: 2001 to 2008 (before the financial crisis) and 2009 to 2014. The experimental evaluation results of testing on the three datasets revealed that the ANN and DT predictive models outperformed the others. NBN achieved the poorest performance prediction regarding unintentional financial restatements over all parameters (confusion matrix, average accuracy rate, F-measures, and AUC).

Fernández-Gámez *et al.* (2016) investigated distinct categories of variables commonly neglected by studies in this field. They combined financial and corporate governance variables for 447 companies, examining MLP and PNN models to see which had the best ability to classify audit opinion correctly. MLP achieved the best average accuracy percentage, with above 98% for the training and testing datasets. This combination of variables was found to have a significant impact on classification ability model performance in terms of selecting the correct audit opinion.

Kirkos *et al.* (2008) tested the ability of ANN, SVM, and DT to classify the correct audit opinion with 338 Irish and UK companies, including 157 "non-big" audit firms and 181 big audit firms, listed for the years 2003 to 2005, with 39 financial ratios. The experimental results revealed that DT outperformed SVM and ANN classification models, achieving higher accuracy (84%), whilst ANN had the worst performance to detect correct auditing decisions.

2.4. Audit Opinion Decision Studies

Since the 2000s, auditing professionals have seen DM analytics models affecting their role in auditing opinion decision in various ways. Because of this, researchers have concentrated on illustrating and investigating the behaviour and performance of DM models for auditing opinion, but research about this area remains relatively limited. Nevertheless, significant papers have been selected on evaluating audit opinion prediction and auditing decision-making process for analysis in this thesis, concentrating on the domain of quantitative methods utilised in developing data analysis models for auditing opinion decision. DM and statistical methods are generally used in auditing in order to obtain efficient and certain outcomes. Related techniques include applying committee combiner techniques, individual classifier models, and dynamic modelling. Since auditors traditionally adopted conventional statistical methods to evaluate audit opinions, DM models are considered to address shortcomings of traditional mathematical statistical methods.

Real historical datasets are utilised in practice to evolve audit opinion models. Each dataset has different features and attributes, particularly in terms of data size. This thesis tested single classifiers' ability to classify different data attributes, and used committee combiner models to benefit from individual classifiers' features while compensating for their individual weaknesses when used in isolation, to demonstrate their capacity to learn data on several divisions of data and characteristic spaces. Most experimental results from previous studies presented that committee combiner models achieve better performance compared to single classifiers performance. This thesis tests the ability of dynamic modelling to predict audit opinions one year in advance, thus it concentrates on related works that used single classifier, committee combine, and dynamic models for auditing. The following subsections explain the process of searching for and selecting related studies with pertinent data, and discusses the salient findings of the existing literature.

2.4.1. Literature Search

In recent decades, initial applications of DM and machine learning tools in auditing are still in their infancy, and auditors are still tentatively exploring their effectiveness. Consequently, there are relatively few studies analysing DM tools to understand impacts on auditing decision making and the development of audit opinion models. Likewise, recent developments in the auditing field have led to renewed interest in DM techniques, investigating how the application of DM tools can affect the performance and efficiency of audit decision-making and if it can enhance classification and prediction audit opinion tools. The process of searching for relevant studies in these regards was undertaken as described below:

1. Searching began by keyword searching (financial statement audit, audit opinion; decision making, DM classification, machine learning classification, individual classifier, committee combiner models, ensemble model, dynamic models, and multi-classifier model) in relevant areas through academic databases and search engines, including Springer, IEEE Xplore, Google Scholar, and ScienceDirect. The searching process presented results centred on studies related to audit opinion, financial statement fraud and misstatement detection, financial restatements, going-concern opinion, auditing selection, and risk assessment prediction. Studies published during the period 2007-2021 were considered, to incorporate the most recent and advanced findings. Initial searching yielded a massive volume of works from conferences, journal articles, accounting textbooks, and theses.

2. The search was narrowed to peer-reviewed journal papers, which are deemed more acceptable in relation to new developments in the audit opinion domain, offering more in-depth and critical perspectives on techniques utilised in auditing.
3. The filtering phase removed unrelated papers (as determined from their titles and abstracts), which were not related to financial statement auditing, pertinent datasets to evaluate developed models, or model performance improvement. Consequently, all papers that aimed to utilise developed single DM modelling and committee machine to enhance the performance of classification models were retained, along with papers on dynamic modelling.
4. In the final phase, all previous related work was collated in chronological sequence, from 2007 to 2021, and these papers were read and analysed in depth to determine their research design (e.g. dataset size, number of datasets, data splitting, variable selection and pre-processing, DM classifier and committee tools utilised, technical criteria used to compare the predicted performance models, hypothesis testing used, and main findings).

2.4.2. Literature Review Analysis and Decision

Table 2.1 summarises the extracted characteristics and information from 39 papers handpicked from peer-reviewed scientific journals, as explained above. These papers include worthy findings and information that could lead to credible inferences regarding committee combiner modelling, individual classifiers, dynamic modelling, and the development of new methods for auditing opinion. Table 2.1 summarises the main features of related works, considering the different fundamental characteristics that need to be considered in the design and development procedure of any classification or prediction audit opinion model, such as dataset size and number of datasets employed to evaluate the models, data splitting methods used, and data pre-processing and cleaning to obtain best performance for the models. It explains the number of DM classifier tools, committee combiner models, and dynamic modelling techniques utilised in each paper, which is related to the dimensions by which models are compared by the stated evaluation measurement parameters. The statistical significance tests used to confirm the robustness and accuracy of the models are also presented. The table summarises the main findings for each study, showing which models were found to have the best performance compared to others tested.

Table 2.1: Related studies

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
Key: AC: Average accuracy; CI: confidence interval; CL: Clustering; ER: error ratio ; FP and FT: false positive and false negative; FS: feature selection; GM: G-measure; MCC: Mathew's Correlation Coefficient; MFFNN: multi-layer feed forward NN; PR: precision; PT: process times; RF: random Forest; RS: random subspace; SP: scatter plot; SR: stepwise regression; TP, and TN: true positive and true negative; WABEM: weight-adjusted boosting ensemble method; W&WOFS: with and without feature selection.							
1. (Jan, 2021)							
Taiwan (352)	Hold-out	W&WOFS	2	✗	AC, TN, TP, PT, Type I and II Error, F-score, PR	✗	CART-RNN best ability to classify going-concern opinion.
2. (Hamal and Senvar, 2021)							
Turkey (341)	Without sampling, over-under-sampling	W&WOFS	7	✗	AC, TP, TN, PR, GM, ROC curve	✗	RF better classification of financial fraud.
3. (Kiziloz, 2021)							
11 datasets from UCI Machine Learning repository and financial dataset from a previous study.	K-fold	FS	5	6 (Greedy, AVG, WAVG, WVOT, Blending, MajVot)	AC, PT, number of solutions, feature size	✓	Greedy, Blending and WVOT better correct classification of FS data.
4. (Craja *et al.*, 2020)							
US (7549)	Financial, linguistic, text	FS	3	✗	TN, TP, FN, FP, F-score, AC, AUC	✗	DPL best financial fraud classification.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
5. (Mohammadi *et al.*, 2020)							
Iran (330)	Hold-out	FS	5	✗	AC, TP, TN, Type I and II Error	✗	ANN best fraud detection performance.
6. (Sánchez-Serrano *et al.*, 2020)							
Spain (298)	K-fold	✗	3	✗	TP, TN, AC, F-measure, PR	✓	MLP best performance to detect right audit opinion.
7. (Papík and Lenka, 2020)							
One dataset (40)	Hold-out	FS	2	✗	TP, TN, FP, FN, AC	✓	LDA good ability to predict financial restatements.
8. (Priyanka *et al.*, 2020)							
India, 2 datasets (776 each)	K-fold	✗	3	✗	PT, TP, TN, FP, FN, ROC, AUC, SP, AC	✗	SVM good ability to classify suspicious companies.
9. (Hooda *et al.*, 2020)							
India (776)	K-fold	FS	10	MajVot	TP, TN, Type I and II Error, AC, F-score, AUC, MCC	✓	MajVot higher ability in auditing decision-making.
10. (Bertomeu *et al.*, 2020)							
One dataset (54,345)	K-fold	✗	9	✗	ROC, AUC, pseudo R2, F-score, catch right restatements, catch right AAERs	✓	RUSBoost best misstatement classification.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
11. (Hsu and Lee, 2020)							
US and Taiwan (17,716)	Before and after crisis period	FS	4	RF	AC, AUC, Type II Errors, kappa value, F-measure, PR, Recall	✗	RF best ability to classify going-concern.
12. (Lahmiri *et al.*, 2020)							
Credit scoring dataset (690 cases), two datasets on corporate bankruptcy prediction (7028, 250 cases)	K-fold	✗	5	✗	AC, ER, Number of weaker learners, PT	✗	AdaBoost good ability to classify financial data.
13. (Sivasankar *et al.*, 2020)							
Australia (700), Germany (1000), Japan (635)	K-fold	FS	5	2(RS and WABEM)	AC, AUC, TN, TP	✓	WABEM best performance.
14. (Stanišić *et al.*, 2019)							
Serbia (13,561)	K-fold	FS	12	2 (Stacked ensemble and combining feature-selection with mixed-effects LR)	AUC, Kappa with 95%CI, AC	✓	Stacked ensemble best ability to classify audit opinion correctly.
15. (Chen, 2019)							
Taiwan (196)	K-fold	FS	2	✗	AC, Type I and II Error	✗	SR-CART better capacity to classify going-concern.
16. (Omidi *et al.*, 2019)							
China (2659)	Hold-out	CL	5	✗	PR and Recall	✓	MFFNN best ability to detect financial fraud.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
17. (Holowczak *et al.*, 2019)							
1 dataset, the Item 4.01 text of 8-K documents, 3509 pairs of market response category	Hold-out	FS	5	✗	PT, TP, TN, FP, FN, AC	✓	BN best performance to classify information on changes in auditors.
18. (Dewiani *et al.*, 2019)							
1 dataset (777)	K-fold	FS	2	✗	TP, TN, FN, FP, AC	✗	Ensemble SVM best ability to detect fraud.
19. (Yao *et al.*, 2019)							
China (537)	✗	FS	6	✗	AC, AUC, F-score, PR, Recall	✗	SVM best fraud detection.
20. (Fernández *et al.*, 2018)							
Spain (252)	✗	FS	5	3 (RF, MajVot and Adaboost)	AC	✓	Adaboost best capacity to classify going-concern correctly.
21. (Hooda *et al.*, 2018)							
India (777)	K-fold	FS	10	✗	AC, Type I and II Error, ER, TP, TN, MCC, F-measure, AUC	✓	BBN and J48 best ability to detect financial fraud.
22. (Randhawa *et al.*, 2018)							
Malaysia (287,325)	K-fold	✗	12	2 (Adaboost, MajVot)	AC, TP, TN, MCC	✓	MajVot best financial fraud detection.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
23. (Jan, 2018)							
Taiwan (160)	✗	FS	2	✗	TP, TN, AC, Type I and II Error	✗	ANN-CART able to detected financial statement fraud.
24. (Dong *et al.*, 2018)							
US (128)	K-fold	FS	4	✗	AC, AUC, F1-Score, recall	✗	SVM best ability to classify financial fraud.
25. (Dutta *et al.*, 2017)							
US (3513)	Based on period	FS	5	✗	FP, AC, TP, AUC, PR	✗	DT best ability to detect financial restatement.
26. (Hajek and Roberto, 2017)							
US (622)	Hold-out	FS	13	✗	AC, TP, TN, AUC, F-measure and MCC	✓	BBN best performance to detect financial statement fraud.
27. (Ozdagoglu *et al.*, 2017)							
Turkey (224)	K-fold	FS	3	✗	AC, AUC, TP, precision, F-measure	✗	ANN higher rates for most results classifying correct audit opinion.
28. (Fernández-Gámez *et al.*, 2016)							
Spain (447)	Hold-out	FS	2	✗	TP, TN, FN, FP, AC	✗	MPL good capacity to detect correct audit opinion.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
29. (Yaşar *et al.*, 2015)							
Turkey (110)	Based on variables	FS	3	✗	TP, TN, AC	✗	DT best ability to classify the correct audit opinion.
30. (Tsai, 2014)							
Australia (690), Germany (1000), Japan (653), bankruptcy (241), UC competition (106,777)	K-fold	CL	5	WVOT	Type I and II Error, AC	✓	SOM-WVOT best ability to classify financial distress.
31. (Song *et al.*, 2014)							
China (550)	K-fold	FS	4	MajVot	AC, AUC, CI, Type I and II Error	✓	MajVot best ability to detect the risk assessment of financial fraud.
32. (Saif *et al.*, 2013)							
Iran (1018)	✗	✗	1	✗	TP, FN, FP, TN	✗	ANN good performance to correctly classify audit opinion.
33. (Pourheydari *et al.*, 2012)							
Iran (1018)	Hold-out	FS	4	✗	TP, TN, AC	✗	RBF best to classify correct audit opinion.
34. (Saif *et al.*, 2012)							
Iran (780)	Based on period	FS	1	SVM -DT	Type I and II Error, TP, TN, AC	✗	SVM -DT best performance to detected correct audit opinion.

Table 2.1: Related studies (cont.)

Sample, size, and no. datasets	Data splitting	Data pre-processing	N/ classifier tools used	N/ committee tools used	Performance measures	Sig. test	Main findings
35. (Ravisankar *et al.*, 2011)							
China (202)	K-fold	FS	6	✗	AUC, AC, TP, TN	✓	GP and PNN best ability to detected financial statement fraud.
36. (Tsai, 2008)							
Japan (690), Germany (1000), Australia (690), bankruptcy (240)	Hold-out	✗	2	✗	AC	✓	SVM good ability for financial decision support.
37. (Kirkos *et al.*, 2008)							
UK and Ireland (338)	K-fold	FS	3	✗	AC, TP, TN	✗	DT best ability to classify audit opinion correctly.
38. (Kirkos *et al.*, 2007)							
UK and Ireland (450)	K-fold	FS	3	✗	AC, TP, TN, Type I and II Error	✗	BBN best tool to classify correct audit opinion.
39. (Gaganis *et al.*, 2007)							
UK (264)	Hold-out	FS	3	✗	TP, TN, ROC, Gini index	✓	ANN best correct classification of audit opinion.

Source: Author

The following analysis is based on the data summarised in Table 2.1. It can be seen that the majority of related studies (30 out of 39) utilised one dataset to assess their classifiers. The remaining nine studies used 2-12 datasets:

- 2 (Kirkos *et al*., 2007; Kirkos *et al*., 2008; Hsu and Lee, 2020; Priyanka *et al*., 2020).
- 3 (Lahmiri *et al*., 2020; Sivasankar *et al*., 2020).
- 4 (Tsai, 2008).
- 5 (Tsai, 2014).
- 12 (Kiziloz, 2021).

The average of number of the datasets employed in all 39 studies equated to 1.43, which is relatively small. In-depth analysis of these datasets revealed that over half of the studies used less than 800 companies, but five studies utilised datasets size ranging between 1,000 to 18,000 firms (Tsai, 2008; Tsai, 2014; Stanišić *et al.*, 2019; Hsu and Lee, 2020; Sivasankar *et al.*, 2020); and three employed substantially larger datasets of 106,777 (Tsai, 2014), 54,345 (Bertomeu *et al.*, 2020), and 287,325 (Randhawa *et al.*, 2018).

Most studies concentrated on shared datasets from Asian countries, such as China, India, Iran, Japan, Malaysia, and Taiwan. Datasets from other countries were less numerous, although many used datasets from Australia, Germany, Ireland, Poland, Serbia, Spain, Turkey, the UK, and the US. Researchers depending on datasets collected from a homogenous cluster of countries when developing new models can lead to prejudice (Ala'raj and Abbod, 2016). On the other hand, datasets with highly divergent features can be prone to impartiality and skewed results and conclusions (Kiziloz, 2021).

Seven studies (Holowczak *et al.,* 2019; Dewiani *et al.,* 2019; Papík and Lenka*,* 2020; Priyanka *et al.,* 2020; Bertomeu *et al.,* 2020; Lahmiri *et al.,* 2020; Kiziloz*,* 2021) utilised private data prepared by firms to test accuracy and validate their models. The reason most previous studies did not use private datasets is because audit datasets are not publicly and easily obtainable due to confidentiality policies.

To train models and evaluate data, most studies used K-fold cross-validation (n = 17) and hold-out (n = 12) data splitting methods. Four studies (Fernández *et al*., 2018; Jan, 2018; Saif *et al*., 2013; Yao *et al*., 2019) did not use of any data splitting methods. Six split their datasets by other ways. Hamal and Senvar (2021) divided the dataset based on sampling; Yaşar, Yakut, & Gutnu, (2015) and Craja *et al*. (2020) based on data type; and Saif *et al*. (2012), Hsu and Lee, (2020) and Dutta *et al*. (2017) based on years. Hold-out method splits a sample randomly into two or three sections for training, testing, and validating model performance. K-fold method includes splitting the sample into K folds of equal size, whereby

K cannot override the dataset size, due to which model training is based on K1 subsets. A residual K subset in the model is conserved for testing and training, then the procedure is carried on until all K subsets are employed for assessment. The examined K forecasting is utilised to assess model accuracy, simply averaging all accuracy for K forecasting. Most researchers selected data splitting methods based on suitability to deal with problems like stratification of sample by class or size, but some studies did not use any data splitting because they preferred to train models utilising whole datasets.

The pre-processing of data is of paramount significance in building models, because every dataset consists of variables with various features and differing characteristics. Samples may include noisy data, outliers, and irrelevant features that can undermine model performance. A good quality dataset is essential to test and develop model performance, and it directly affects data qualification and quality. The most common data pre-processing methods are feature selection and clustering (Ravisankar *et al.*, 2011), as used in most of the analysed papers, but seven of them (Tsai, 2008; Saif *et al*., 2013; Randhawa *et al*., 2019; Sánchez-Serrano *et al*., 2020; Priyanka *et al*., 2020; Lahmiri *et al*., 2020; Bertomeu *et al*., 2020) did not apply pre-processing, which they did not regard as having any impact on their models' performance. 28 studies used feature selection in the pre-processing phase to be selecting most relevant and significant variables utilised in training models. Just two studies employed clustering in their pre-processing phase. Tsai (2014) utilised cluster method in the pre-processing phase in constructing hybrid classifier by clustering data in terms of differences. Omidi *et al.* (2019) employed clustering method to specify and filter noise and outliers from the dataset, after which unfiltered data was used to train classifier performance.

In addition, two studies tested classifier performance with and without pre-processing in order to determine whether feature selection affects classifier performance, and how significant this phase is to improve model performance (Hamal and Senvar, 2021; Jan, 2021). Hamal and Senvar (2021) and Jan (2021) concluded from the evaluation of experimental results that using feature selection has a positive impact on developed model performance. The data pre-processing phase is clearly essential for the improvement of audit decision opinion models.

After the pre-processing stage, classifier model development and utilisation are deployed. All related studies tested DM model classification, without considering prediction. Additionally, the number of classification models utilised in the studies varied from 1-36, with a mean of 4.73 classifier models. This can be attributed to each study testing a developed classification model and comparing it with others within the same work; or comparing developed models with other models from previous works (Kirkos *et al.*, 2008; Tsai, 2008; Fernández-Gámez *et al.*, 2016; Craja *et al.*, 2020; Hooda *et al.*, 2020; Papík and Lenka, 2020). In general, the number of classification models utilised in the work relies on a developed classifier and how many models are used for performance comparison. Analysis of these studies demonstrates that there is no universal parameter to compare classifiers, and adjusted comparison factors, like splitting methods, evaluation measures used, and datasets utilised, must be considered in order to demonstrate relative superiority. Papík and Lenka (2020) noted in comparing their experimental outcomes with other studies' outcomes that comparison can be credibility sufficient when researchers are well-versed in the particular models utilised.

A total of 29 related studies tested the performance of individual classifiers without using committee combiner methods. Most of these studies used the same models, such as ANN, DT, LR, LDA, and SVM. There were 24 out of 29 studies that just considered the test performance of the base single classifiers, without using ensemble or combiner models, in order to evaluate and compare classifier performance. This is related to the relatively novel nature of DM use in auditing. Five out of the 29 studies (Hamal and Senvar, 2021; Bertomeu *et al.*, 2020; Lahmiri *et al.*, 2020; Dewiani *et al.*, 2019; and Hooda et al., 2018) constructed and used ensemble classifiers to learn and exploit the powers and address weaknesses of base classifiers, such as Random Forest, J48, AdaBoost, and RUSBoost. Dewiani *et al.* (2019) used ensemble method to enhance SVM model performance. The main purposes of the 28 studies were to test and compare different numbers of single classifier models' performance with other models.

Ten works used followed committee combiner model, which is significant in building audit opinion models. All of them used parallel structure for a variety of data from base classifiers, trained in a combiner model. Combiner models improved base classifier performance, as indicated in better evolution results, with several types of classifier models being utilised for data training. After having the data from base models ready, combiner models were developed to train data subsets to reach final decisions. Five out of the ten studies (Hooda et al., 2020; Hsu & Lee, 2020; Tsai, 2014; Song et al., 2014; Saif et al., 2012) tested one committee combiner model, and three studies (Sivasankar et al., 2020; Stanišić et al., 2019; Randhawa et al., (2018) tested two combiner models, whilst Fernández et al., (2018) tested three combiner models, and Kiziloz, (2021) used and compared six combiner models. In addition, nine studies used a popular combiner model tested by previous studies. These signifies that committee combiner model has given low attention on usage, development, or comparison of different committee combiner performance together. Therefore, most related works concentrated on testing and comparing individual classifiers' performance rather than trying to enhance individual classifier performance by using combiner rule. Sivasankar *et al.*

(2020) and Stanišić *et al.* (2019) structured classifier models that popularise and perform data analysis effectively, suggesting new ideas or generating various modelling builds.

All related studies utilised traditional performance evaluation measures derived from the confusion matrix:

- 34 used average accuracy rate
- 22 used sensitivity
- 24 used specificity
- 16 used Type I and Type II Error
- 2 used error rate
- 10 used F-Measurement
- 5 used MCC
- 6 used precision
- 15 used AUC
- 3 used ROC
- 5 used time performance

New performance evaluation measurements not used previously in the audit opinion were developed by six studies (Gaganis *et al.*, 2007; Song *et al.*, 2014; Bertomeu *et al.*, 2020; Lahmiri *et al.*, 2020; Priyanka *et al.*, 2020; Hamal and Senvar, 2021): Kappa value, G-measure, Scatter plot, Pseudo R2, Number of weaker learners, Confidence interval, and Gini index. From the empirical perspective, all of these measurements illustrate diverse views of model performance. Each study used convenience evaluation measurements that are suitable for particular models, but most of the studies did not add any new parameter to generate new perspectives on model performance.

After evaluating model performance, it is necessary to determine if the results are credible and powerful, and not random. Over half of the studies did not utilise statistical significance testing, relying only on measurement parameters *per se* to evaluate the model performance. However, 17 studies did define the statistical significance test of their proposed models' performance outcomes. These studies stated that this was a significant phase in the model development procedure, and it helped to draw powerful conclusions about model performance. However, they used different statistical test methods available for validating and contrasting performance outcomes, like non-parametric tests (e.g., Friedman and chi-squared test) and parametric tests (such as analysis of variance (ANOVA) and t-test). Analysis of these 17 studies reveals that selecting a suitable test method for implementation relies on several considerations, including the number of classifier models employed and requiring to be compared, the number of variables used, and the number of datasets utilised.

All of the studies used statistical models with machine learning models, such as linear regression and linear discriminant analysis. The results showed that the statistical models had inferior performance compared to machine learning to classify data correctly. Studies tested and compared between individual classifiers, ensemble models, and committee models, finding that ensemble models and committee combiner models have better performance to classify and predict correctly compared with individual classifier performance, achieving the fundamental aim of committee and ensemble models (to enhance the accuracy achievable by individual classifiers used in isolation). Studies that used committee combiner methods illustrated that these techniques had superior classification accuracy, outperforming both ensemble and individual classifier models (Song *et al.*, 2014; Tsai, 2014; Fernández *et al.*, 2018; Randhawa *et al.*, 2018; Stanišić *et al.*, 2019; Hooda *et al.*, 2020; Sivasankar *et al.*, 2020; Kiziloz, 2021). However, among studies that only tested the performance of single classifiers, two studies compared the performance of deep learning model (DL) with NN; Jan (2021) stated that recurrent NN achieved better performance than DL, while Craja *et al.* (2020) found that the DL model performance outperformed NN. In addition, studies using more than one classifier showed that DT, SVM, and all types of NNs had better capacity performance to classify and product correctly.

From the reviewed studies, the major phases of building the general framework design for the audit opinion model include the following:

- Data collection was undertaken by most studies using public benchmark datasets. It is better to have a real industrial dataset in order to include varied views on various datasets, and data features, sizes, and class distributions.
- Selecting convenient splitting methods for the dataset is significant in regard to data size and classes distribution (such as minority and majority classes).
- Modelling approach is shaped by researcher perspective and interest, but efficient models with credible outcomes are necessary. Some researchers train individual classifier models with raw data, but others train after cleaning and analysing raw data. Other developers attempt to develop new ideas through ensemble and committee combiner models. The modelling approach is heavily influenced by the developer’s perspective.
- Evaluation measurement parameters of many kinds are available, and researchers must choose the most suitable to mirror all angles of machine learning model performance.
- Statistical significance testing proves the validity of model outcomes to indicate DM models’ inherent value for real application.

Several gaps in extant literature were identified from this analysis, including the following:

- Related studies did not undertake consistent pre-process phases for data cleaning, normalisation, splitting, and feature selection for raw datasets. Most merely used splitting *and* feature selection, and some just used one step (feature selection *or* splitting data).

- Most of studies tested base single classifiers without consideration of homogenous classifier ensembles and committee combiner models. Likewise, the majority of works tested performance based on single classifiers that have been tested by previous studies, without testing new single classifiers and comparing them with old single classifiers' performance.
- Only ten studies tested committee combiner model, of which five tested one committee combiner model, three tested two combiner models, one tested three combiner model, and one tested six combiner models. In addition, nine of the studies tested the same popular combiner model, which has been tested by other previous studies. Little attention has been given to developing, using, or comparing different committee combiner models' performance. Likewise, most works just concentrated on testing and comparing individual classifier performance with combiner models, rather than to trying to enhance individual classifier performance by using combiner rule.
- Ten works developed a committee combiner process, whereby every constituent single model was given independent rulings before combination into one individual outcome, without any coordination or collaboration among the base model by the learning process. There is a need to utilise a new committee combiner model with decision-making that includes single models working in the set to obtain consensus on the final outcome.
- All related studies concentrated on using and testing classification machine learning; no studies discussed, developed, or tested the ability of dynamic modelling performance to predict auditing opinion in advance, and no related studies called for research in dynamic modelling for auditing.
- A few studies used statistical significance testing to confirm validation of model performance outcomes, and this was infrequently utilised for developed models across the analysed studies as a whole. Studies which tested for statistical significance stated that this was a fundamental step to determine DM model validity.
- None of the related studies suggested model integration in relation to the tasks achieved in this thesis, including the use of:
    - Public datasets from different sectors.
    - Four pre-processing data techniques.
    - Testing and developing a novel individual classifier model (Deep learning model) and comparing its performance with other single classifiers.

- Comparing between more than three combiner rules.
- Novel committee combiner model.
- Developing and testing dynamic modelling.
- Sundry evaluation measurement parameters to comprehensively validate performance models.
- Adding a new parameter not used before in the field.
- Statistical significance testing (in this context).

Despite the *prima facie* complexity of this method and its multiple steps, this thesis examines the extent to which every phase of modelling can improve performance, and answer the question of whether intricacy is worth investigating to develop auditing opinion models, given that most auditing studies used only simple statistical models to build efficient audit opinion models.

2.5. Summary

This chapter explored the theoretical background of audit opinion decision and information technology, and reviewed directly related literature concerning this thesis topic. It first presented the background in terms of financial statement audit definition, audit phase, audit evidence, and analytical review, and explored the development of new information technology applications in relation to DM and BD in the auditing process, explaining the significance of these lines of inquiry. These applications significantly alter the size of audit evidence and the classification of audit opinions, and this chapter analysed the advantages, limitations, and challenges of the application of the BD and DM in audit opinion.

The chapter investigated how the attribute of BD has affected the size of audit evidence, and illustrated the impact of BDA in its positive and challenging implications for audit decision-making. DM techniques were explained in terms of how they affect audit phases, and the structuring of prediction and classification tools for auditor decision support. The variations between classic audit analysis and DM were explained, presenting the practical issues pertaining to auditors in initial applications of DM in the auditing process.

The latter part of the chapter reviewed related literature, explaining the mechanisms used to search for pertinent studies and the types of data sourced from them pertaining to DM use in auditing and related fields. The identified relevant works were systematically analysed, summarised, and discussed based on their experimental research design (e.g., dataset size, number of datasets, and data partitioning, pre-processing stage, individual classifier and committee combiner models used, evaluation measurements used to compare models' performance, statistical testing, and main findings).

From the analysis and discussion of the related works, various issues, conclusions, and findings were highlighted that merit further investigation, for instance the limited use of committee combiner model beyond testing relative performance against single classifiers, and concentrating on introducing novel committee combiner models. Likewise, previous studies did not explore using dynamic modelling to predict audit opinion, and did not even identify the need to research such DM models for auditing opinion.

The next chapter presents the research methodology for the development procedure of the framework design of audit opinion modelling, and the associated problems are identified and addressed.

Chapter 3
Research Methodology for the Audit Opinion Model

3.1. Introduction

This chapter explains how this study's proposed experimental framework for the audit opinion model was designed and executed, with description of every stage included in the framework and associated justifications. It first introduces the audit opinion modelling approaches used, comprising individual classifiers, dynamic modelling, and committee combiner modelling, with the topology for each model considered separately. It then explains the dataset collection process and pre-processing methods applied to select the final datasets. The two methods used to split these datasets are compared, in order to identify the optimal one. The third section explains the evaluation measurement techniques used to evaluate each proposed classifier's performance for subsequent comparison of these. In the next section an explanation of further evaluation of the model's statistical significance testing is performed. The final section summarises the proposed experimental framework.

3.2. Prediction Model Approaches

The main objective when structuring an audit opinion model is to devise a classification model that can best distinguish between audit opinion types. As seen in Table 2.1, many researchers have utilised a wide range of individual classification models to structure audit opinion models. In this section, the commonly utilised state of the art classification models related to the objectives of this thesis are summarised. The first subsection explains the topology of each of the nine individual classifiers. The second presents the development of committee combiner models used for the second contribution for this thesis, with the aim of enhancing the accuracy performance of individual classifiers. The third presents the theoretical framework for dynamic modelling for forecasting audit opinion in advance.

3.2.1. Individual Classifiers

In recent decades, studies have proposed single classifiers with the aim of delivering better performing audit opinion modelling. Table 2.1 shows that these studies were aimed at achieving better performance according to different features, including the variables utilised, size of the data, and the type of market. Early audit opinion classifiers employed statistical methods, with the most commonly deployed being logistic regression, linear discriminant analysis. These statistical models illustrated different shortcomings that led to their replacement through new and improved machine learning classifiers such as SVM and ANNs. Most related studies have investigated the utilisation of single models for such matters as financial statement audit opinion (Hooda *et al.*, 2018; Bertomeu *et al.*, 2020; Hamal and Senvar, 2021).

Utilising single models to construct audit opinion models can result in efficient classifiers. In this thesis, nine machine learning classifiers are trained and tested: deep learning; support vector machines; artificial neural networks; K-nearest neighbour; decision trees; naïve Bayes network; logistic regression; linear discriminant analysis; and boosting ensemble. These models' performance outcomes are compared to ascertain which is the most efficient in classifying audit opinion correctly. In this section, the framework and topology of each of these individual classifiers is presented.

3.2.1.1. Logistic Regression

The LR model has been used widely in the financial and accounting areas. The primary purpose for using it more than other statistical models or traditional classifiers is because it supplies a convenient balance for the dataset and a high level of efficiency, interpretability, and accuracy regarding the prediction outcomes (Nikolic *et al.*, 2013). LR classification method was developed and utilised in auditing and fraud studies to solve binary classification. LR constitutes appropriate regression analysis when the dependent variable is binary compared to liner regression. Because of the LR is used to explain the relationship between different independent variables and the binary classification the modelling target by fitting them to the s-shape logistic curve an where all of the output values are between 0 and 1 (Chintalapati and Jyotsna, 2013). This regression is used to determine the conditional likelihood of the observation relating to the class, which can help resolve the binary classification issue of the identified likelihood in the case where there is a binary output target variable, which consists of only two possible values (no/yes, 0/1, or false/true) for predicting the variable. It is possible to obtain multiple classes through performing this method for every class (Ozdagoglu *et al.*, 2017).

LR creates an s-shape logistic curve, where values are between 1 and 0, which describes the relationship among the independent variables and the probability of a binary output for the target variable, utilising the non-linear function of the independent ones. LR substitutes an original actual variable that cannot be approximated correctly by utilising the linear model through the log conversion of the odds rate, where $In\left(\frac{L_I}{1-L_I}\right)$ represents the log odds of having qualified or unqualified audit opinion. A logit function (equation 3.1) converts the straight-line curve into an approximately non-linear s-curve, and changes the range from 1 and 0 to positive and negative infinity. The coefficient of logit and the intercept are evaluated utilising a maximum likelihood function, which is a repeated operation aiming to reach the best value in terms of increasing the log probability. Having evaluated the odds of the logit function, their likelihood is known, the output of which will be between 1 and 0. These two likelihoods are compared, whereby the larger one represents the class label value that is most probably the target value (Kotsiantis *et al.*, 2006).

$$logit(L_I) = In\left(\frac{L_I}{1-L_I}\right) = \beta_0 + \beta_1 X_I + \cdots + \beta_M X_{IM} = In\left(\frac{Prob(Y_I = 1)}{Prob(Y_I = 0)}\right) \qquad 3.1$$

3.2.1.2. K-Nearest Neighbour

K-NN is a supervised model, where the outcome of a new class inquiry is categorised based K-NN classes. The model is determined according to three prime factors: the number of neighbours utilised to categorise new data, distance rule applied to obtain classification from the K-NN, and distance metric utilised to define nearest neighbours. The main phase of the K-NN model is measuring the distance between all prior and new data. The new data is added to the largest group of K (number of neighbours), which is chosen from the sample by the K-NN model based on distance-based learning (Entezari-Maleki *et al.*, 2009). Other distance-based classifier models begin with groups of distances as original training data, but in the K-NN model every new point is contrasted with a current data point through the utilisation of the distance metric, with Euclidean distance measuring the distance between the existing point (XN) and the new input data point (XM), as shown in equation 3.2. Isolating these distances in rising order and choosing K items with the lowest distance values to the input data point stratifies the voting rule, enabling the discovery of a plurality class between these k-samples. A nearest current point (T) is utilised to assign a class as new, with the model sometimes utilising more than one nearest neighbour, and the majority class of a nearest k-neighbours is appointed as the new one (Entezari-Maleki *et al.*, 2009; Zareapoor and Shamsolmoali, 2015).

$$dist(X_N, X_M) = \sqrt{\textstyle\sum_{T=1}^{I} (X_{NT} - X_{MT})^2} \;\; T = 1,2,3 \cdots I \qquad 3.2$$

K-NN model is called a lazy model, because it learns very slowly compared to other machine learning ones. On the other hand, it is one of the most robust models, because it does not have to make any presumptions about the data. Moreover, the distance estimate can be measured consistently between two classes and calculated according to several paths, because K-NN is a non-linear model without a function form. Furthermore, the K-NN classifier follows a learning operation producer to learn in which it keeps concentrating on storing the data until it is actually having the input data whose class is meant to be predicted. In the K-NN model, the optimal value of k is one which balances among bias and variance. Large k values ignore underlying trends in the data, thus will have contribute to smoother decision boundary of classifier which mean raise bias but lower variance. The larger K values can assist in decreasing the impact of a noise on classification and provide probabilistic information when the ration of data for every label (class) gives information regarding the ambiguity of the decision. Finally, K-NN has the ability to solve complex issues even with very small datasets (Zareapoor and Shamsolmoali, 2015; Kiran *et al.*, 2018).

3.2.1.3. Naïve Bayes Network

Naive Bayes Network (NBN) classifier is used to classify classes based on conditional likelihoods determined from training data. NBN is a simple probabilistic classifier based on Bayes theory, and it is very effective for the measurement of high-input attribute spaces. Bayes theory pertains to making decisions based on statistical probability regarding the predicting of an event, by drawing on learning from past proceedings. It is a very simple model for making categorising rules that are more precise than rules made via other models; and this model follows a powerful or weak statistical statehood hypothesis to predict the variable. NBN works based on the supposition that attribute likelihoods are independent of one another (Dutta *et al.*, 2017; Kiran *et al.*, 2018).

Bayes rule is measured as shown in equation 3.3. For instance, in an auditing opinion situation, take a training dataset A= [$B_{1,}$…., B_m], where every B is made up of m features [B_{11},…., B_{1m}], with the class label A (qualified or unqualified audit opinion), where n is number of the class. The NBN model concentrates on analysing these training dataset instances and defining the mapping function f: [B_{11},…., B_{1m}] > (A), which is used to set the label of the unknown B= [B_1,…., B_m].Bayesian classifiers choose the class that has highest posterior probability P (A_n| B_1,…., B_m) as the class label, based on the higher posterior likelihood criterion and lower error likelihood, which means when *P (A_n | B) = MAX P(A_n| B)*, B can be assigned to the particular class and A_n can be specified. Under the model, it is assumed that the variables of the data are not correlated and are independent; this is deemed a weakness, since dependence between variables can exist.

$$p(A_n \mid B_1 \cdots B_m) = \frac{p(B_1 \cdots B_m \mid A_n) * p(A_n)}{p(B_1 \cdots B_m)} \qquad 3.3$$

Where p(A_n) denotes the prior likelihood of class (qualified and unqualified audit) A_n before seeing data B, *P($B_{1,}$…., B_m)* is the probability of data x belonging classA_n, *p (A_n| B_1,…., B_m)* is a posterior likelihood of class A after viewing data B, and *p (B_1,…., B_{1m}|A_n)* is the probability of the data B belonging class A_n.

While previous studies have provided evidence that NBN has poor performance compared to other classifier models, it has some properties that can be helpful to predict better outcomes, such as (Dutta *et al.*, 2017; Kiran *et al.*, 2018):

- It utilises a class independence hypothesis which is simple to implement, understand, and explain.
- It is quick to train and classify datasets, with good ability to distinguish irrelevant features and outliers.
- It is a widely used supervised learning technique due to its accurate performance for many real-world datasets.
- It is highly efficient because it evaluates parameters by utilising very small data for training, which enables classification.

3.2.1.4. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) model distinguishes two or more groups based on some independent features. When there are two categories (two-group discriminant analysis), LDA is comparable to binary LR analysis. As seen in equation 3.4, the LDA function takes the form of the linear integration of the coefficients of variables and their respective variables in order to categorise classes or groups for observations (Uddin *et al.*, 2013; Santoso and Wibowo, 2018). In terms of auditing opinion classification, suppose there is dataset of A companies, where every company in dataset has variables H= [H_1, …, H_A], and these are utilised to classify companies into their classes (C), which are qualified (0) and unqualified (1) opinion. The aim of the LDA model is to assess the likelihood of the company *p(C|H)* given the vector of its variables or features. The LDA relies on a maximum probability of right classification, and it structures some linear integrations of coefficients of independent variables (β = [β_1, …, β_A]), known as discriminant functions, to divide observations into predetermined classes. In other words, this function suggests dividing the features through linear integration of the coefficients of independent variables, in order classify objects in stable classes (Santoso and Wibowo, 2018).

$$Z = \alpha + \beta_1 H_1 + \cdots + \beta_A H_A = \alpha + \sum_{A=1}^{I} \beta_A H_A \qquad 3.4$$

Where Z is the discriminant score, α denotes the constant term, β_A represents weights of variable or discriminant coefficient, H_A present the predictor or independent variable, and A is the number of predictor variables (*A* = 1, 2, …., I).

The cut-off point is measured according to the cost of error and the *a priori* likelihoods of classes (Rafiei *et al.*, 2011). Every company is categorised into unqualified or qualified classes, based on cut-off point and discriminant score. Companies are categorised into classes based on their discriminant scores being higher or lower than a cut-off point. Alternatively, companies may be categorised on the basis of the likelihood of belonging to one of the classes and cut-off probability point (Gaganis, 2009).

One of the benefits of the LDA model is its simplicity of application when classifying a linear dataset. However, it supposes that there are linear correlations between variables, and LDA has uncertain performance when treating nonlinear data, which can lead to wrong classification error and inappropriate prior likelihoods. Nevertheless, the LDA model is still widely used in different areas in the accounting and finance fields (Pohar *et al.*, 2004; Hoque *et al.*, 2015).

3.2.1.5. Decision Tree

The Decision Tree (DT) uses a hierarchical or tree structure (Chang and Chen, 2009). As shown in Figure 3.1, the DT includes the leaf nodes, with tests performed on them being represented as branches, with root nodes. This allows for the identification of potential characteristics and patterns in the dataset, until the optimal output (prediction) is obtained. The core role of DT is to classify input data into two classes of qualified or unqualified audit opinion. The DT is structured as described by previous studies (Chang and Chen, 2009; Gepp *et al.*, 2010; Tsai *et al.*, 2014):

- It is built top-down, starting from the node that includes the two abovementioned classes.
- The node is divided into two sub-sets, symbolising these two potential events.
- A decision function loops on all the divisions to detect the optimal one, and the target sub-tree that represents the best split for unqualified and qualified audit opinion is chosen, based on the if-then rule. This rule identifies classes in terms of their overall error rate and lowest misclassification cost, thus allowing for the construction of a predictive model.

The DT model has the ability to classify a dataset with high accuracy for the following reasons, which have led to it being used across several fields as a classification model (Lin *et al.*, 2015; Dutta *et al.*, 2017):

- Developing the model with classifiers does not demand any previous knowledge or hypotheses.
- The model relies on the if-then classification rule, which requires no previous knowledge, and makes the dataset simple to classify.
- DT is a simple model for decision-making procedures, given that it consists of sets of simple decisions.

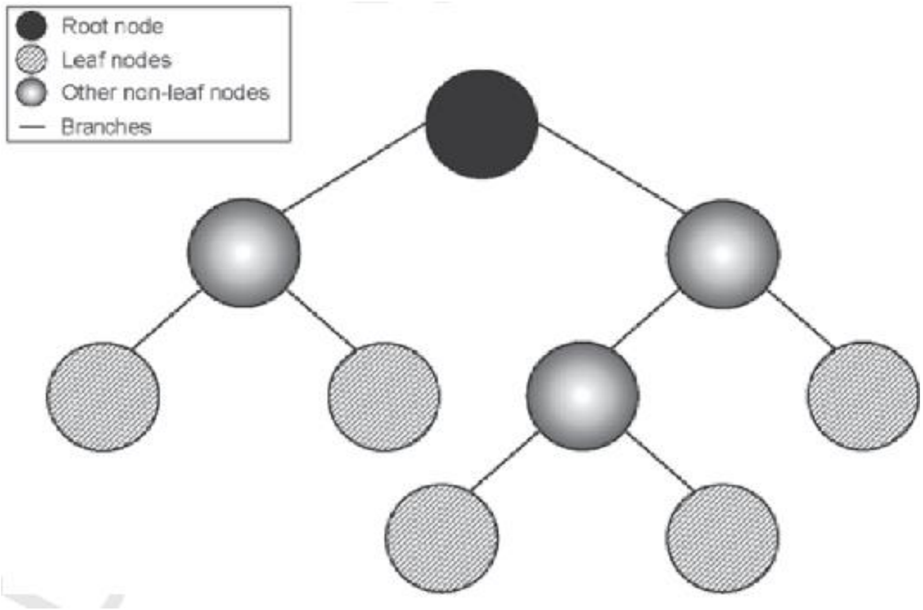


Figure 3.1: DT structure

Source: Gepp *et al.* (2010, p. 539)

3.2.1.6. Artificial Neural Network

Artificial Neural Networks (ANN) is designed to mimic human brain functions in terms of holding complex correlations between output and input values. An ANN can predict output data through complex systems with different efficient input parameters. It consists of incorporated process units called neurons or nodes, which are able to process input data, characterised by building a multi-layer perceptron (input, output, and one or two hidden layers). The numbers of neurons in every layer and nodes are arranged through its feed-forward and back propagation (Lin, 2009). As can be seen in Figure 3.2, the training of an ANN model is in one direction, beginning with the feeding of the dataset X_N into the input layers. These inputs are then forwarded from the input layers to hidden layers by synapses, located according to the weight of the input values. Each hidden layer applies activation equations measuring the outcome of all nodes in these layers, as shown in equations 3.5 and 3.6.

$$net_T = \sum_{N=1}^{M} X_N \, W_{NT} U; \; T = 1, \cdots, V \qquad 3.5$$

$$A_T = \mathfrak{f}_H(net_T); T = 1, \cdots V \qquad 3.6$$

Where net_T is the activation value of the Tth node, A_T refers to the outcome of a hidden layer, and $f_{\boldsymbol{H}}$ is the node activation equation used on the value of every neuron in a hidden layer. The activation equation may select many functions, but the most common one is the sigmoid function (equation 3.7).

$$f_H(X) = {}^{1}/_{(1+e^{-X})} \qquad 3.7$$

The hidden layer outcome with new weights is moved to the output layer, and the final ANN output of the decision (O_J) is measured by equation 3.8.

$$O_J = f_J\left(\sum_{K=1}^{V} W_{KJ}\, A_T\right); J = 1, \cdots, B \qquad 3.8$$

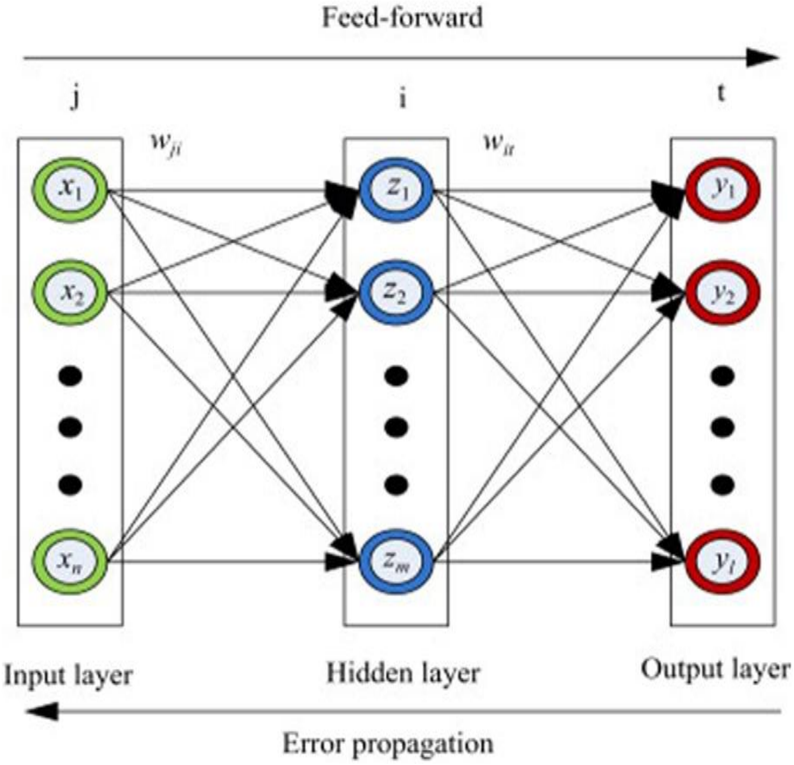


Figure 3.2: Topology of a feed-forward back propagation ANN

Source: Alhnaity and Abbod (2020, p. 6)

When there is a significant variation between an observed actual value and final output decision, the backward learning algorithm is propagated (backwards) in order to update bias and weight values between the output layer and hidden layer, and between the hidden layer and the inputs. The calculation is repeated until the mean squared error between the actual output value and the final network's prediction value is at a minimum. If the final outputs are not optimal, the prediction value is modified through the ANN classifier until it becomes acceptable (Cao *et al.*, 2005; Alhnaity and Abbod, 2020).

ANNs have been shown to be suitable tools in the field of finance. Huang *et al.* (2007), Calderon and Cheh (2002), and Bahrammirzaee (2010) identified their significant attributes as the following:

- They have the ability to discover the underlying functional relevance between input and output, and to identify complex patterns between the two. These qualities can be utilised in the classification and prediction process.
- They can be effectively employed to handle data that is noisy, incomplete, missing, complex, irrelevant or partial. This is because they can treat data as input without any kind of conversion being needed, whilst some algorithms need to convert numerical data into nominal values, which can lead to data loss.
- They do not require any previous assumptions relating to data distribution of the input data, because they have the ability of updating data using suitable parameters, whilst statistical models like LR and LDA require assumptions about the distribution of the input data.

However, auditors still have lack traceability of decision making and how identification of the audit opinion type from ANN model. Moreover, another ANN disadvantage pertains to there being no function that optimises the choice of parameters, which can negatively affect their prediction accuracy (Bahrammirzaee, 2010).

3.2.1.7. Support Vector Machine

SVM is one of the more powerful and commonly used artificial intelligence classification solutions utilised in auditing. It operates based on a statistical learning theory, with learning algorithms being utilised to analyse input data in a linear classifier. SVM classifies binary data using the finest line of the hyperplane that classifies two classes' boundaries. This is obtained by mapping input training vectors (X_m, Y_m) into d-dimensional attribute distance, where x ∊ Rn denotes vectors in a d-dimensional feature space, and Y ∊ {±1} is the class label. SVM then tries to obtain the optimal separating hyperplane that divides input data into two classes through maximised margin width among training data on a margin (support vectors) and the optimal hyperplane. The maximum margin hyperplane ($W*X + B = 0$) splits negative data points from positive ones. The vector W is perpendicular to the diving hyperplane, and parameters B/W represent the space from a coordinates centre to hyperplane (Huang *et al.*, 2018; Padmavathy and Mohideen, 2020).

A linear model is generated with the new area, for which the optimal separating hyperplane is built, as illustrated in equations 3.15 and 3.16. When data are linearly separated, SVM trains the linear model for the optimal hyperplane that splits a data without error into the higher distance between nearest training points and hyperplane. Support vectors are training examples near to the optimal separating hyperplane. Based on their attributes, the required input data belongs to one of the two classes that may be classified. In the case of non-linear input data, SVM utilise various kernel functions, such as the sigmoid linear polynomial and radial basis function (RBF) to map non-linear data into higher dimensional space, whereby the linear model can then classify non-linear classes. A linear model in the new distance represents a non-linear decision margin in actual space (Pai *et al.*, 2011; MIN and Lee, 2005). Figure 3.3 summarises the SVM model framework.

$$(X * W) - B \leq -1 \; if \; Y_m = -1 \qquad 3.9$$

$$(X * W) - B \geq 1 \; if \; Y_m = 1 \qquad 3.10$$

where *X* is the input dataset, *W* refers to the weight, B denotes the base value, and Y is output of point m.

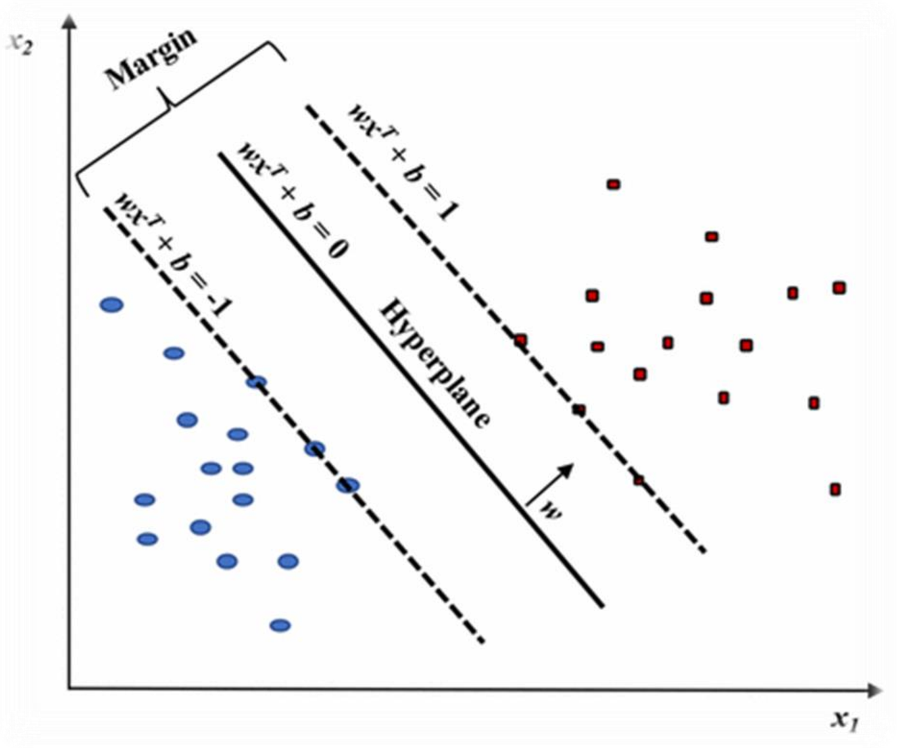


Figure 3.3: SVM model

Source: Huang *et al.* (2018, p. 42)

SVM has numerous advantages that support its use in various areas, and in this thesis (Min and Lee, 2005; Pai *et al.*, 2011):

- It is lean in terms of structural risk decrease, which means that this sort of model decreases the upper bound of actual risk, while other models reduce empirical risk.
- It has just two free parameters, kernel parameter and upper bound, to be selected, and the generalisation capacity of SVM mainly relies on space dimensionality and parameters. Consequently, SVM has powerful inference ability, generalisation capacity, quick learning ability, and capacity for correct prediction.

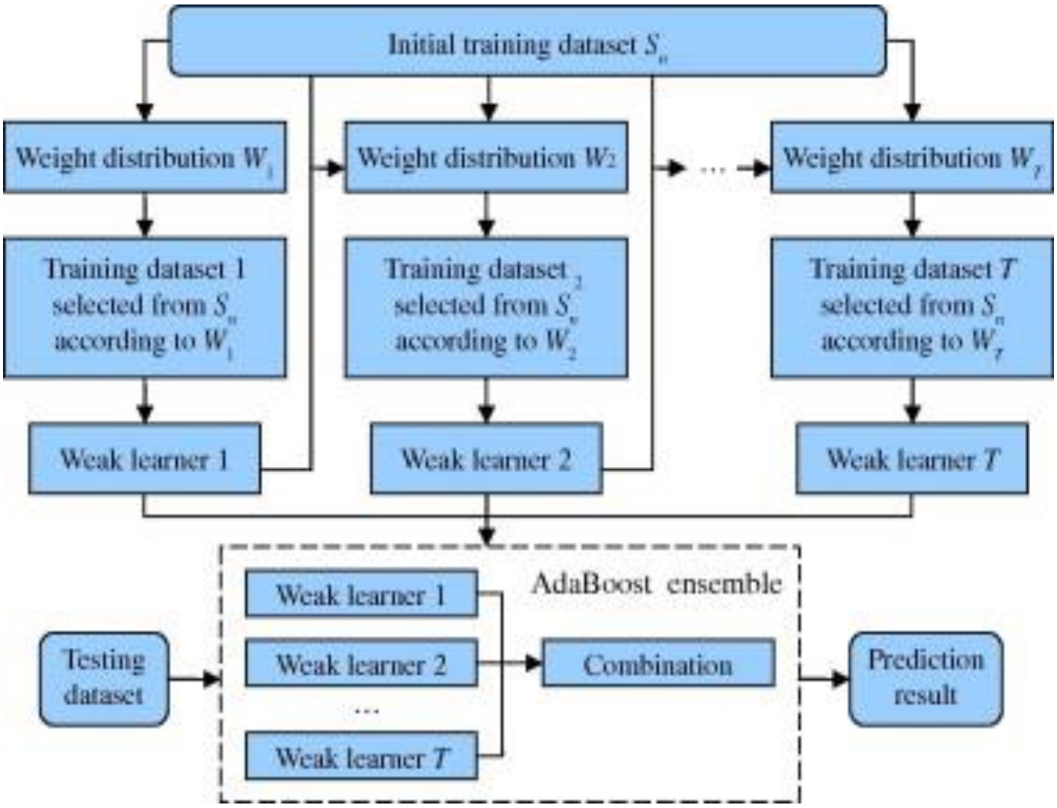
- It is plain enough to be anatomised mathematically, and it may be illustrated to coincide with a linear model in the maximum dimensional attribute space nonlinearly related to input space.

Due to these advantages, SVM can be a promising alternative for integrating the power of conventional statistical models, making it more straightforward and theory-driven, while machine learning models are more data-driven, distribution free, and powerful.

3.2.1.8. Boosting Ensemble Classifier

BEC is a group of techniques for a building an ensemble classifier. A differential characteristic of boosting classifiers is that they combine learning algorithms in series, allowing the improvement of learner performance from many sequentially connected weak classifiers, while other DM models utilise the same species of learning algorithm. Likewise, boosting models gain classifiers and training datasets in a robustly deterministic path (Skurichina and Duin, 2002). For AdaBoost-boosted DT modelling classifier, DTs are weak classifiers, whereby every tree attempts to enhance classifying accuracy and decrease errors of past trees (Lu *et al.*, 2012; Xiao *et al.*, 2017).

The main idea of boosting model is to frequently stratify the base learner to improved version of training data, $W_m = [(X_1, Y_1)\ (X_1, Y_1), \cdots, (X_m, Y_m)]$ and $Y_n[n = 1,2,3, \cdots, h]\epsilon[0,1]$, which illustrate two classes of objective. The boosting model is processed in the path whereby every phase’s data distribution is adapted to put more weight on misclassified classified data, and less weight is specified to correctly classified data, in order to decrease misclassification errors of the past classification tree. In addition, increased weight is assigned to stronger classifiers based on their classification performance, and the final classification output is the weighted mean of all the weak classifiers. An advantage of adding trees sequentially is that BEC learns slowly, making it perform better. Adaboost structures is used to combine all weak leaners in order to get prediction outputs (Abellán and Castellano, 2017). Figure 3.4 summarises the ensemble classifier topology.



Initial training dataset S_n
Weight distribution W_1
Weight distribution W_2
Weight distribution W_T
Training dataset 1 selected from S_n according to W_1
Training dataset $_2$ selected from S_n according to W_2
Training dataset T selected from S_n according to W_T
Weak learner 1
Weak learner 2
Weak learner T
AdaBoost ensemble
Weak learner 1
Weak learner 2
Weak learner T
Combination
Testing dataset
Prediction result

Figure 3.4: Ensemble classifier framework

Source: Sun *et al.* (2011, p. 9307)

3.2.1.9. Deep Learning Model

The Deep Learning (DPL) model, unlike ANN, has more than two hidden layers. Given the depth of the layers and greater number of neurons or nodes, DPL has greater representational power compared to ANN. As a result, DPL has emerged as a robust method for sentiment analysis with massive data. DPL can entail supervised or unsupervised learning. It obtains input data, from which it automatically learns important characteristics, then it identifies unlabelled data availability from the input data by classifier training. The layers in the created hierarchical DPL identify data characteristics that are independent of human knowledge. Other benefits of DPL are that it can eliminate noisy data and unimportant attributes, thereby producing new, cleaner datasets. For these reasons, DPL has become a topic of great interest across different fields, with researchers calling for more studies on it. Moreover, only a few researchers have applied it to the financial and auditing fields (Sohangir *et al.*, 2018; Sun, 2019).

DPL can be utilised as a supervised learning classifier that consists of input units, hidden units and output units, where most of the data points processing work is done. DPL trains the dataset (D) to make the classification using several hidden layers. It uses a hierarchical arrangement comprising a series of non-linear conversions applied to the dataset. Every non-linear transformation is as a layer, where the output of the first transformation is first layer, and so on. The number of the hidden layers signifies the depth of the DPL architecture (Heaton *et al.*, 2017). DPL building depends on different kind of architectures and in this thesis that of the long short-term memory (LSTM) is utilised. One of the benefits of LSTM is that it keeps data for long-term periods, and it has been shown to be the best architecture for DPL prediction compared to other architectures. Likewise, its very deep NN time direction learns sequence and time patterns from data sequences (Bao *et al.*, 2017).

In DPL with LSTM (DPL-LSTM), memory blocks comprising one memory cell and four major portals (input, forget, input modulation gate, and output) replace the hidden layer units (neurons) of the RNN approach. These memory cells and four major portals play a significant role in training long-range dependency, whilst controlling the information store. DPL-LSTM introduces a directional loop that utilises past information to analyse an existing outcome, as the past outcome is related to an existing outcome sequence, and the nodes among the cells are connected. Equations (3.11 to 3.16) present the measured final prediction value (ht) after having received input information at time t (xt) to utilise a past hidden (ht-1) layer. The first phase refers to the forget portal (Fi), which sets out what data should be kept or stripped, with the data from the previous hidden layer being passed by a sigmoid function (σ) together with the information from the existing input. A sigmoid function is used to specify the information to be discarded based on its value. In the next phase, as shown in equations 3.11 to 3.16, tanh and the sigmoid functions are utilised to make the decisions regarding updating the information from the input data. In the final phase, the final predictions are calculated through applying equation 3.16 (Alghofaili *et al.*, 2020, Livieris *et al.*, 2020).

$$f_t = \sigma(W_{f^{x_t}} + U_f h_{t-1} + B_f) \qquad 3.11$$

$$i_t = \sigma(W_{i^{x_t}} + U_i h_{t-1} + B_i) \qquad 3.12$$

$$g_t = tanh(W_{c^{x_t}} + U_c h_{t-1} + B_c) \qquad 3.13$$

$$c_t = i_t \odot g_t + f_t \odot c_{t-1} \qquad 3.14$$

$$O_i = \sigma(W_O[D_{i-1}, E_i\,] + B_O) \qquad 3.15$$

$$h_t = o_i \odot tanh(c_t) \qquad 3.16$$

where O_i represents the output portal, U and W are the input weight metrics of the output portal, B pertains to the biases, σ denotes the sigmoid activation function, x_t denotes the input variables, t denotes the time, i_t is the input, g_t is the input modulation, f_t is the forget portals, $\odot$ represents pointwise multiplication, and c_t is a vector of memory status.

The DPL-LSTM is utilised in this thesis as a single classifier to classify audit opinion and subsequently employed in the dynamic modelling to predict audit opinion one year ahead. Its topology is shown in Figure 3.5.

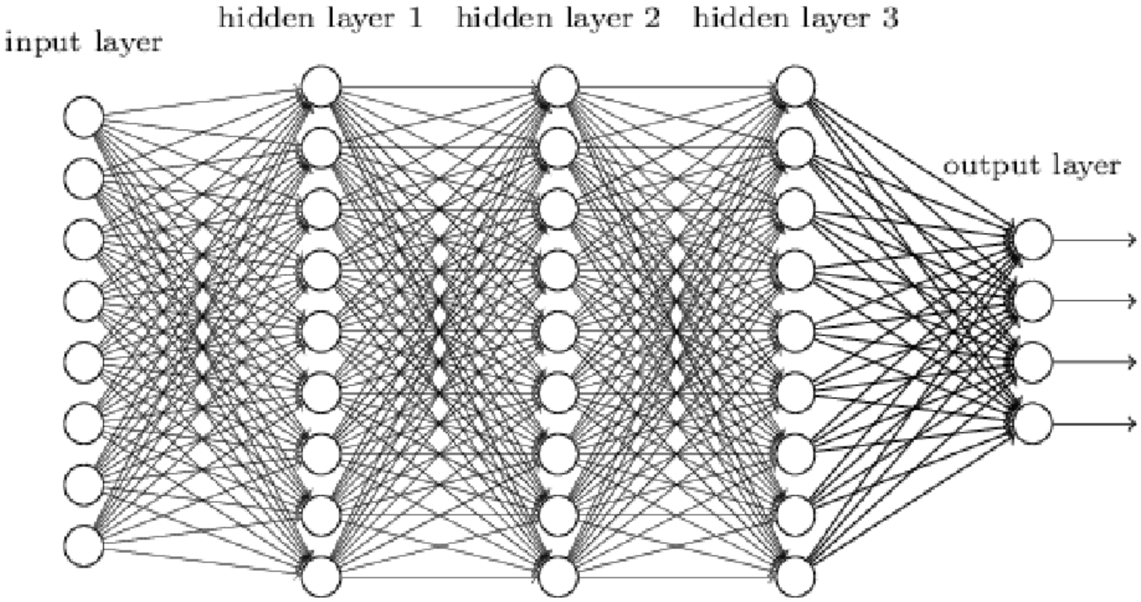


Figure 3.5: Topology for the DPL model

Source: Sun and Vasarhelyi (2018, p. 56)

3.2.2. Committee Combiner Modelling

The committee combiner modelling (CM) essentially seeks to combine the outcomes of several individual classifiers to make a decision on a final output. The main aim of committee modelling is to improve overall classification reliability and reduce estimated error percentage. CM uses multiple classifiers, and committee modelling processes encompasses strengths from each single classifier in order to product the best final output. Likewise, CM seeks to build the group of suppositions and combine single classifiers to produce the desired output, while individual (constituent) models only learn one supposition from the dataset (Woźniak *et al.*, 2014; Yijing *et al.*, 2016).

CM techniques are distinguished from hybrid ones in that the former combine all outcomes of multiple classifiers to award the final decision, while hybrid techniques combine classifiers in sequential operation, whereby just one classifier ultimately awards a final outcome (i.e., the other classifiers' outputs are mere input data for a single classifier to give the final output). The main challenge in constructing CM is to combine different classifiers as effectively as possible to function as a coherent committee group (Woźniak *et al.*, 2014).

CM design includes three main phases: framework topology, classifier generation, and choosing techniques to be utilised for integrating individual classifiers (Zuh, 2010; Woźniak *et al.*, 2014). In first phase, a system topology of constructing CM model can process through serial or parallel structures; this thesis applies parallel structure, which is commonly employed in constructing CM models, training single models on the same input data (Du *et al.*, 2012; Woźniak *et al.*, 2014). Subsequently, the CM model is built heterogeneously, by combining various types of classifiers to make a committee with complementary decision capability, based on several classifiers' views on the same data. Generated classifiers are trained on various features or data of training dataset parts; thus each trained model popularises in several paths (Seijo-Pardo *et al.*, 2017; Papouskova and Hajek, 2019). In the final phase, combining method or rule is used to combine all classifier prediction decisions. This thesis applied classifier fusion to combine all model forecasts trained for a whole issue, with committee methods or rules utilised to integrate all prediction classifiers together. The popular committee rules used to integrate the outcomes of several individual classifiers to make a decision on a final output such as weighted average, average, and majority voting classifier (Du *et al.*, 2012; Woźniak *et al.*, 2014).

As discussed previously, CM applications in auditing remain tentative, being used in few studies, as summarised in Table 2.1. Previous studies used several CM patches to achieve the best data classification performance, achieving better performance compared to individual classifiers. Because of these points, this thesis develops different committee combiner models using average, weighted average, weighted voting, majority voting, minimum rule, and maximum rule with consensus combiner model and fuzzy logic combiner model.

3.2.3. Dynamic Modelling

The nonlinear autoregressive exogenous (NARX), nonlinear autoregressive (NAR) and deep learning-LSTM are most commonly used model in dynamic modelling for two main reasons. First, NAR and NARX models can remove noise from input data and obtain more substantive characteristics for final forecast classification, while DPL-LSTM classifiers can efficiently catch sequence pattern data. Second, NAR and NARX are fully suitable for processing spatial auto correlation data, but it is not usually suitable to accurately to deal with issues such as complex data and long and short periods of time, while DPL-LSTM model can deal with such issues using training set attributes. Time series classifiers using DPL-LSTM can enhance time series forecasting performance (Tealab *et al.*, 2017; Fawaz *et al.*, 2019; Livieris *et al.*, 2020).

NAR and NARX consist of the feedback and the feedforward network. The architecture of feedback networks is characterised by sets the inputs to neurons of past layers, or outcomes of neurons in the layer can be inputs for the same neuron. Feed-forward networks, also called static, employ the nonlinear function of entries, and interconnected neurons function as a group, whereby data flows in a forward way (from inputs to outputs) (Tealab *et al.*, 2017). The DPL framework for prediction time series modelling is processed as presented in Figure 3.6. In this thesis is deployed in the DPL-LSTM. As noted from Table 2.1, dynamic modelling is relatively unexplored with regard to time series prediction for audit opinion. The chapter 6 of the dynamic modelling is presented to develop the performance of three models.

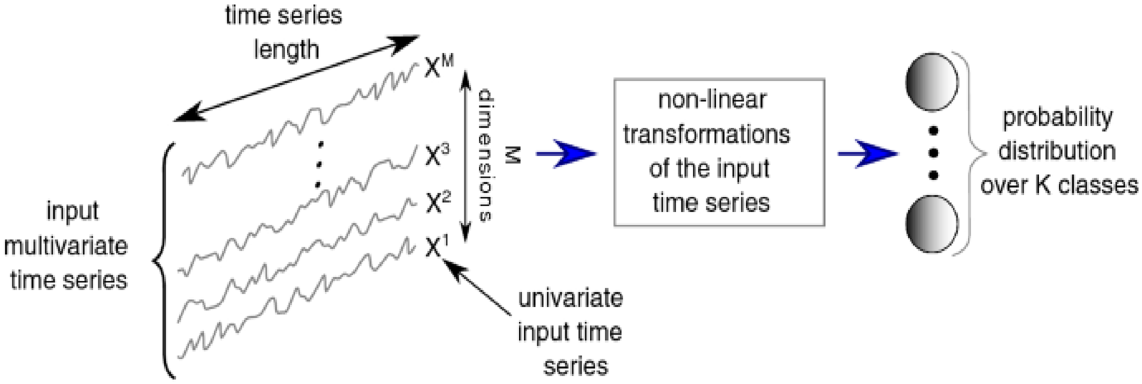

time series length
X^M
X^3
X^2
X^1
M dimensions
input multivariate time series
univariate input time series
non-linear transformations of the input time series
probability distribution over K classes

Figure 3.6: Deep learning framework for time series modelling

Source: Fawaz *et al.* (2019, p. 5)

3.3. Research Strategy

This section explains the process used to collect and itemise datasets utilised in this thesis. It describes the data pre-processing process to get the best model performance, concentrating on imputation, normalisation, and feature selection. The final step is dataset splitting.

3.3.1. Data Collection

The preliminary stage of constructing the audit opinion model is collecting the dataset(s) utilised to implement models, considering their size and number related to aggregation. Real-world datasets are of massive data size. Among the studies summarised in Table 2.1, more than half employed datasets including less than 1,200 companies, and more than half relied on public datasets. However, they observed that there is no gold standard of the suitable size for the dataset, and practical considerations of availability are often the most expedient issue. Some of the reviewed studies validated developed classifiers with multiple datasets. Datasets vary in their attributes, including number of samples, features, and class division, all of which must be factored into study design to reach credible conclusions on developed classifiers' validating capacity in relation to datasets with multiple features.

This thesis collected different public datasets from different sectors for experimental model evaluation. Public datasets were used due to the limited availability of private datasets, due to confidentiality issues (as discussed previously). Datasets from Ireland and the UK were selected by Financial Analysis Made Easy (FAME) software, which is one of the few software packages publishing real companies' data for audit opinion analysis (Fame, 2021). The software makes it easy to access audit opinion types, including large volumes of historical data on around 11 million active and inactive companies, with up to 10 years' worth of data on every firm. The data is largely sourced from the Companies Registration Office (Ireland) and Companies House (UK).

Particular datasets were selected according to certain search criteria. The sampling to validated individual classifier model included active and inactive firms that have qualified and unqualified audit opinion, then all firms with many missing values were excluded, to leave companies that provided most of their available financial and non-financial data reflecting performance and financial position. 45 independent numerical variables were then calculated from this available annual report for each company, along with audit opinions covering unqualified and qualified audit opinion forms. But the procedure to selected sampling to validated dynamic model, were followed same producers to selected datasets to validated individual classifier excepted selected just active companies that have qualified and unqualified audit opinion to valuated dynamic models.

Individual classifier models were validated using four datasets for active and inactive companies over the years from 2019 to 2017. The first three (yearly) datasets presented companies that published their qualified and unqualified audit opinions for the years 2019, 2018, and 2017, while the fourth ("All-Data") dataset combined the former three datasets in order to test classifier ability to test bigger data sizes. Each of the three yearly datasets comprised data for over 10,000 number of firms, to address the limitation of most previous studies. Where use in dynamic modelling to predict audit opinion in advance year (i.e., year 5, 2019, in this thesis), five datasets comprising 6,712 active companies that have unqualified and qualified audit opinion logged as active in the year 2019 (T) and four years before the period 2019 (T-4, T-3, T-2, and T-1). Table 3.1 summarises the datasets used in this thesis, which show number of firms that have qualified and unqualified audit opinion.

Table 3.1: Summary of all datasets

Dataset	**Characteristics (No.)**		
	Number of Firms	Unqualified opinion	Qualified opinion
Year 2019 Dataset	23,817	14,350	9,467
Year 2018 Dataset	18,007	10,644	7,363
Year 2017 Dataset	10,703	6,492	4,211
All-Data Dataset	52,527	31,486	21,041
Dynamic Modelling Dataset for 5 Years	6,712	3,356	3,356

Source: Author

3.3.2. Data Pre-Processing

DM modelling and the improvement of classifier popularisation performance basically relies on data quality (Alasadi and Bhaya, 2017), which is largely determined by the convenience of data to be utilised in connection to the number of samples, and the significance of attributes, utilised in the analysis and detection of outliers (Blake and Mangiameli, 2011). Based on this, data pre-processing is a fundamental stage in auditing area classification issues (Table 2.1).

In general, a raw dataset may be sensitive to inconsistent values, outliers, noise, and incomplete features, and it might include features or samples that are redundant, irrelevant, unreliable, or noisy. These would pose issues in classifier training, rendering knowledge mining and detection harder. Consequently, it is important for datasets to be pre-processed before training (Singh and Singh, 2010; Alasadi and Bhaya, 2017). Proof of dataset quality and portrayal is significant prior to any analyses. Data pre-processing is the most substantial stage to confirm dataset efficiency, and hence pre-processing is an essential machine learning phase, dealing with data transformation and dataset preparation, while seeking to enhance knowledge discovery and make classifier procedure more efficient. Likewise, it is considered as the most decisive stage in developing a classifier that can deal with a raw dataset (García, 2015; Alasadi and Bhaya, 2017).

Data pre-processing is undertaken via various techniques, including feature selection, data transformation, and imputation. After pre-processing, new training datasets are ready for further procedures and analysis (Rizki *et al.*, 2017). The following steps were undertaken in pre-processing for the datasets adopted in the suggested model in this study.

3.3.2.1. Data Imputation

The first step in data pre-processing is data imputation, used to address missing values in collected datasets. Missing values usually arise from companies overlooking some fields when filling financial statements or application. When selecting data in datasets that have incomplete or missing field values, model training may be disturbed. To dealing with this problem, datasets must move through the data imputation step in order to make the dataset simple and functional for learning discovery. Acuna and Rodriguez (2004) stated that the simplest path to deal with missing values is to delete the case consisting of missing values of an attribute; but there are other paths to dealing with this, including exchanging missing values based on some estimations. In the datasets collected in this study, missing values were addressed by the simple data imputation method of exchanging missing categorical data within an average value of attributes holding the missing value (Lessmann *et al.* 2015), using SPSS (version 25) software.

3.3.2.2. Data Normalisation

After overcoming missing values, every feature in the dataset consists of values that differ to some extent. To avert prejudice, data must be fed to the model within the same interval, and values must be converted from a variety of scale values to a mutual one. Normalisation processes raw data and rescales or transforms is such that each attribute has a uniform contribution. Likewise, normalisation helps overcome two major problems of raw data which hinder DM algorithm learning: outlier data and the presence of controlling attributes (Singh and Singh, 2020).

Normalisation techniques set data values in a range between -1 ~ 1 or 0 ~ 1. This technique is helpful for DM models like SVM and ANN, which demand input data in the range 0-1, and in vectors of real numbers (Alasadi and Bhaya, 2017). To obtain these datasets, features must be normalised to numbers in a range 0-1 utilising the most convenient path, because when the plain normalisation technique is utilised (like taking highest number of features in the data and splitting all features via this number), all normalised numbers tend toward 0, which does not mirror the target data. This causes prejudice and inoperative model training. In this thesis, datasets were normalised by utilising min–max normalization method, which scales original data to the predefined upper and lower frontiers linearly (Singh and Singh, 2020). The numbers in between are rescaled based on equation 3.17.

$$Y = \left(\frac{X - max_b}{max_b - min_b} * (Newmax - Newmin)\right) + Newmin \qquad 3.17$$

Where Y is normalization data, X is original data, max_b and min_b represent the maximum and minimum value of variable b (respectively), and $Newmin$ and $Newmax$ denote the minimum value in the feature (given the value of 0) and the maximum value in the feature (given the value of 1), respectively.

3.3.2.3. Feature Selection

Collected datasets generally include various heterogeneous features. Data collection utilised in constructing classifiers is correlated with features, thus datasets can be rendered redundant by irrelevant attributes, which make training classifiers more complex, and make it difficult for them to learn on the dataset; this results in low accuracy and performance. Because of this, analysing attributes and discussing their significance is a necessary and fundamental step for pre-processing data for DM models. Feature selection involves deleting unwanted attributes and choosing the most significant and relevant ones; it is the procedure of selecting a subset of representative attributes germane to developing classifier performance. Feature selection decreases model training time increases performance, decreases dimensionality (to improve forecasting performance), and helps achieve better insight to visualise and understand data.

As seen from Table 2.1, most previous studies used feature selection in their data pre-processing, but with different methods related to their datasets. This thesis applies stepwise regression method with SPSS in the feature selection step, because accounting and financial studies commonly affirm its effectiveness to acquire the most representative features that improve classifier performance. Stepwise regression discovers the best combination of predictor features. There are many variables in the stepwise process to discover individual forecasting variables and the outcome is a combination of such variables, all of which have considerable coefficients (Tsai, 2009). Stepwise regression involves both adding and removing features. The stepwise process chooses important attributes based on mean square root error (MSR), which it changes when specified features are removed from or added to the model, and the significance level (p-value) should be $p < 0.05$ or below to be significant at 95%. With standardized regression coefficient ($B_{K.STD}$), regression analysis on standardised independent and dependent variables would yield standardised coefficients as shown in equation 3.18. When there are vastly variation units involved for the features, this is a path to compare between features.

$$B_{K.STD} = B_K * (A_{X_K}/A_Y) \tag{3.18}$$

Where A_{X_K} and A_Y are the standard deviations for the corresponding Kth independent variable (X) and target variable (*Y*).

Stepwise regression for feature selection illustrated only 34 out of 45 independent variables, as shown in Table 3.2.

Table 3.2: Independent variables (n = 34) after feature selection

Independent Variables	Ratio Number
Liquidity ratio	5
Efficiency ratio	7
Profitability ratio	9
Solvency ratio	3
Converse financial ratio	2
Financial risk ratio	5
Non-financial ratio	3

Source: Author

3.3.3. Data Splitting Method

Data splitting is applied after data pre-processing to split the dataset into training and testing sets, utilised for structuring and assessing classifiers (respectively). Splitting data is significant in building, evaluating, and validating models. Most related studies split their datasets into a visible data training set (to train classifiers) and an invisible data testing set (to validate classifier performance), in order to illustrate how classifiers performed, with implications for use in real-world future cases.

The relative size of data in training and testing sets is an issue. Larger training sets increase the calibration of a classifier to the data, and increase precision and credibility (i.e., better accuracy will be achieved with a training set of 2000 entities than one of 500). In addition, there is problem related to the range of available data and the number of data sample related to every previous class, and utilising a specific partition method significantly affects classifier performance with several data class distribution and dataset sizes. Likewise, fair distribution of a dataset with several classes for testing and training assures that dataset has several classes trained effectively, to have the better classifier popularisation over a testing set.

Several splitting data methods have been utilised in auditing relative to such considerations, as determined by researchers in their particular contexts, but as seen from Table 2.1, the majority of related studies concentrated on K-fold and hold-out methods. This thesis applied two methods potentially suitable for model development: hold-out, used for ANN, DL, and committee combiner models; and K-fold, applied for K-NN, NBN, LR, LDA, SVM, BEC and DT.

3.3.3.1. Holdout Technique

Hold-out method divides the dataset into a training dataset (to train a classifier) and a testing dataset (to validate it). Hold-out technique is a very simple splitting method, widely used by the previous studies. As shown in Table 2.1, all studies that used this method split dataset randomly into 20% testing and 80% training datasets. Therefore, this method can be prejudiced in its reliability outcomes; data may be unsuccessfully utilised, and both testing and training datasets could be non-representative, such as the testing and/ or training dataset having uncomplicated and/ or complicated data. This problem may be averted by repeating a hold-out method to have randomly chosen testing and training datasets in each iteration, which reduces the likelihood of obtaining a lucky testing dataset. While a testing and training dataset can be overlapping, this is not ideal.

3.3.3.2. K-Fold Cross-Validation

In K-fold cross-validation method, premier datasets are split into folds or K subsets of approximately equal size, such as F_1,….,F_k (the number of splits made from the premier dataset). Every split must be individually tested and trained. Table 3.2 presents these operations of training and testing practically, and shows that an operation of K-fold cross-validation suggests that all splits obtainable are for training, except one split for testing. An operation carries on until all splits are tested and trained. Ultimate accuracy is assessed through taking a mean of all folds or K-subsets examined. K-fold method utilises the impact of all data available, hence averting any overlapping from occurring, and it could be more efficient to echo operations multiple times, because of testing and training data as much as possible at every recurrence.

Table 3.3: K-fold cross-validation process

Partitions/Folds	**Training Set**				**Testing Set**
1	F_2	F_3	F_4	F_5	F_1
2	F_1	F_3	F_4	F_5	F_2
3	F_1	F_2	F_4	F_5	F_3
4	F_1	F_2	F_3	F_5	F_4
5	F_1	F_2	F_3	F_4	F_5

Source: Author

Problems may arise regarding how many K-subsets or folds are used for data, and whether K-subsets are numerous and classifier performance is precise, with elevated difference. Small K-subsets may produce biased classifier performance and decreased difference. The optimum number of K-subsets depends on dataset size. All related studies that used K-fold had 10 or 5 folds, which García (2015) reported can be perfect for datasets with various sizes, with recurrences of operation can be handled with switching between testing and training data to avert difference. Consequently, this thesis adopted 5-fold cross-validation for each dataset with operations repeated 10 times for each dataset, awarding a total of 50 experience outcomes used to award the final outcome for every dataset, to obtain robust and accurate inferences relating to classifier performance.

3.4. Model Evaluation Techniques

The evaluation of machine learning over a selected dataset is essential in experimental modelling development. As well as determining outcomes *per se*, evaluation considers whether outcomes are credible for the practical application of models to real-world datasets. There are numerous index measures to evaluate the strength and reliability of model performance and forecasting ability (Lessmann *et al.*, 2015). This thesis adopted nine parameters for performance measurement to evaluate each classifier’s performance and to compare between them. Seven estimate parameters commonly used in auditing studies were derived from the confusion matrix: average accuracy, AUC, specificity, sensitivity, Type I Error, Type II Error, and F-measures. Two additional estimation methods were used: Brier score and AURD. These nine parameters were used to evaluate classifier performance and to determine significant characteristics.

3.4.1. Confusion Matrix

Confusion matrix interprets the performance capability of a developed forecasting model in terms of how much data has been correctly or incorrectly classified, through the ability to distinguish classification performance data as correct or incorrect classification. Various criteria can be derived from the confusion matrix, including sensitivity and specificity, Type I and Type II Error, average accuracy ratio, and F-measure (Dutta *et al.*, 2017), which have been extensively used in previous literature to evaluate developed models. Table 3.4 illustrates the confusion matrix.

Table 3.4: Confusion matrix

		Predicted classifier (%)	
		Qualified	Unqualified
Actual class (%)	Qualified	True Negative (TN) (Specificity)	False Positive (FP) (Type I Error)
	Unqualified	False Negative (FN) (Type II Error)	True Positive (TP) (Sensitivity)

Source: Author

3.4.1.1. Sensitivity and Specificity Ratios

Sensitivity (true positive) rate refers to the correctly classified correct companies, while the specificity (true negative) rate is defined as the correctly classified proportion of unhealthy companies. Specificity and sensitivity are measured as shown in equations 3.19 and 3.20. In this thesis, specificity reflects the number of qualified audit opinion instances that are rightly specified as qualified opinions, while the sensitivity rate is the number of audit opinions correctly classified as unqualified.

$$Specificity = TN/(TN + FP) \qquad 3.19$$

$$Sensitivity = \ TP/(TP + FN) \qquad 3.20$$

3.4.1.2. Type I and Type II Error

Method of error measurements of Type I and II Error rate are opposite to sensitivity and specificity measurement. Type I Error (false positive) refers to percentage of unqualified firms that were incorrectly flagged as another class, while Type II Error (false negative) reflects the number of the incorrectly classified of qualified companies to unqualified class. Type I and Type II Error are calculated as shown in equations 3.21 and 3.22. Models able to correctly forecast qualified firms are more beneficial to auditors than models specialised in correctly forecasting unqualified companies, because it is more significant for auditors to predict qualified companies than missing the opportunity to predict unqualified ones. Likewise, it is important to construct balanced models that are not prejudiced for any category.

$$Type\ II\ Error = FN/(FN + TP) \qquad 3.21$$

$$Type\ I\ Error = \ FP/(\ FP + TN) \qquad 3.22$$

3.4.1.3. Average Accuracy Ratio

The average accuracy rate refers to the percentage of the total number of cases correctly classified (i.e., rightly classified qualified and unqualified), as shown in equation 3.23, and it reflects a degree of proximity between target values and forecasting values (Hsu and Lee, 2020). It is a widespread measure for evaluating the performance of developed models, and it also utilised by most financial fraud and audit opinion literature (Table 2.1), and average accuracy is a very significant parameter to draw conclusions on model performance for accounting and auditing applications. Because of this, in this thesis average accuracy rate is considered an appropriate measure for drawing inferences regarding predicted model performance in terms of detecting the right audit opinion.

A drawback of accuracy measurement is that it does not go in depth into how tasks of various types have individually performed. For instance, with a dataset of 95% unqualified companies and 5% qualified companies, the developed model might correctly classify all unqualified companies and misclassify all qualified companies, yet the model would have the best accuracy performance at 95%. Consequently, it is better to define how effectively a model is able to classify various types, and to acknowledge whether a parameter is prejudiced toward a particular type.

$$Average\ accuracy = (TP + TN)/(TP + TN + FN + FP) \qquad 3.23$$

3.4.1.4. F-Measure

F-Measure is calculated from the weights of mean of recall and precision, as in equation 3.24. In order to avert measurements that purpose error deviation, F-measure combines sensitivity and precision into a metric utilising the weighted harmonic average of sensitivity and precision (Craja *et al.*, 2020). In other words, F-measure is a combination of right classification of unqualified types proportional to all instances classified as unqualified with the number of correct unqualified cases. This means that it measures how accurate and robust a model is to classify unqualified audit opinion cases.

$$\text{F–Measure} = (2 \text{ x Precision x } sensitivity)/\ (\text{Precision} + sensitivity) \qquad 3.24$$

3.4.2. Area Under Curve

Area under the curve (AUC) is the area under the receiver operating characteristics (ROC) curve, consisting of two dimensions: the y-axis (sensitivity rate) and x-axis (specificity rate). The AUC converts likelihoods into 1 or 0, based on the specified cut-off score; in predictive modelling, this measurement is utilised to compare classifiers of two classes. A model with good AUC must tend toward 1 or the upper left corner of the ROC curve, while being under the diagonal line indicates a bad model (Ala’raj and Abbod, 2016). Any increase in sensitivity values occurs at the cost of increasing the false positive rate. AUC illustrates the accuracy of the classifier by a relative trade-off between negative cases (qualified audit) and positive cases (unqualified audit), which is helpful for organizing models and visualising their performance, especially in cases with skewed class distribution and imbalance classification error costs. For instance, with regard to the binary issue, the ROC curve permits visualisation of a trade-off between a percentage of positive classes versus a percentage of false negatives for various portions of the test set.

The ROC curve gives guidelines on setting a cut-off value, depicting model features without consideration of operators like error cost and class distribution. Consequently, ROC efficiently separates model performance based on these features (Zhou, 2013). AUC has different advantages over other evaluation tools related to its capacity to work without being influenced by misclassification costs or distribution classes (Dutta *et al.*, 2017). Studies using AUC (Table 2.1) found that it has preferable performance with imbalanced data distributions, and good measurement to evaluate model performance; Likewise, ROC curves are uninfluenced by any variation in misclassification or distribution classes outcomes from misclassification, due to relying only on the performance of cases. Because of this, AUC is used widely in different fields.

3.4.3. Brier Score

Brier scores use mean square error to calculate the accuracy of the likelihood prediction values of the model, through taking mean squared distances between model prediction outcome values and actual target values, as shown in equation 3.25. Brier score illustrates the square root of possibility of an error, major variation between values, but average accuracy, which transforms classifier prediction into two separate classes (0 and 1) based on a threshold pre-determined value. In additional, brier score does not depend on confusion matrix results as average accuracy rate to measure model accuracy. But brier score is used final floating prediction values for the model and actual target values as show in equation 3.25. Two essential components of Brier scores are accuracy and credibility, measuring how forecasting likelihood is closed to actual values, and how much conditional probabilities differ from forecasting mean (Assel *et al.*, 2017).

$$Brier\ score = \frac{1}{M}\sum_{B=1}^{M}(d_B - x_B)^2 \qquad 3.25$$

Where *M* denotes the number of audit opinion cases, d_B refers to final floating prediction (qualified and unqualified) values model for company B, and x_B is the actual target classes for company B. As illustrated in Table 2.1, Brier score measurement is not generally used in auditing or going-concern opinion studies.

3.4.4. Area Under Reliability Diagram

Reliability diagram graphically presents the accuracy of probabilistic prediction. The reliability feature is visible as essential demand when proving probabilistic production, since a lack of accuracy could introduce the methodical prejudice in subsequent decision making. Reliability diagram includes the plot of an observed relative frequency versus forecast likelihood, presenting visual comparison when tuning a probabilistic prediction system and documenting the final outcome of a model. It shows the range correspondence of prediction likelihoods for M and observed frequencies achieved over the training model. If correspondence between observational frequency values and prediction likelihood values is ideal, then all data points lie on the diagonal line in the reliability diagram (Weisheimer and Palmer, 2014; Pedro *et al.*, 2018; Gweon and Yu, 2019).

In this thesis, for the reliability diagram, N instances (which consist of the classifier prediction $X_N$and target values Y_N) are grouped into bins. To assess expected reliability from finite prediction, target values are grouped into 20 bins at non-overlapping subintervals, which gives better results compared to other bin numbers like 10, 30, and 50 bins. Sorting from 0 to 1 is based on the distance bin range for each bin; for example, the predicted values between 0 < 0.05 drop into the first bin, but the predicted values between 0.05 < 0.1 fall into bin no. 2, etc. After that, mean prediction and target values are calculated for each bin, and are compared to determine each bin's accuracy. Mean target points are plotted on the Y axis, and mean prediction points on the X axis. When likelihoods achieved through a model are accurate, target values and prediction values are similar for all bins, thus target and prediction points are predictable to fall near to a diagonal line. The final step measures miscalibration, which illustrates the deviation of the reliability curve from the diagonal line (equation 3.26).

$$Miscalibration = trapz(MP, DT) \qquad 3.26$$

where MP and DT present mean prediction and mean target value in bin.

3.5. Statistical Significance Testing

The final phase is evaluating the statistical significance of model performance results, as the latter in themselves are not enough to infer the models' relative performance. Statistical significance eliminates random outcomes arising from the methods utilised and performance model measures. For complete performance estimation, statistical testing confirms that experimental variances in performance are meaningful and significant. As shown in Table 2.1, researchers who evaluated model performance using statistical significance testing considered this essential to demonstrate model validity and credibility, and classifier robustness.

Selecting suitable statistical significance testing methods depends on elements like the number of models to be contrasted and a measurement scale of data outcomes (like nominal or binary). Unsuitable statistical tests may lead to deceitful and uncertain inferences. Statistical test methods may be non-parametric or parametric. Non-parametric tests may be more secure because (unlike parametric tests) do not embed homogeneity of difference or normality of data (Tanha *et al.*, 2020). Due to this, non-parametric tests are widely utilised to enhance estimation procedures of classifier performance, and they can be very useful to interpret the significant experimental outcomes achieved by multiple classifiers on various datasets (Yu *et al.*, 2017). Because of this, the normality trial for datasets used in this thesis utilising a statistical software SPSS were examined, as well outcomes presented that the datasets are not ordinarily distributed. Friedman statistic is particularly advised to compare classifier results over multiple datasets (Demšar, 2006; Berrar, 2017; Bhattacharyya *et al.*, 2020; Tanha *et al.*, 2020).

Consequently, this thesis used Friedman statistic to evaluate individual classifier and committee models over four datasets and dynamic model performance. The Friedman statistic process consists of several rankings of forecasting across multiple models for every dataset (separately). Friedman test assigns higher ranking to the best classifier, and lower ranks to classifiers with relatively worse performance. In this case, the null hypothesis is all models from those to be compared perform identically and whereby distinctions are only random fluctuations (Eisinga *et al.*, 2017; Tanha *et al.*, 2020). The following steps were taken during statistical significance testing in this thesis.

1. To use the Friedman test, the model predictions are arranged in the matrix as shown in Table 3.5, where the m columns represent the model predictions, and the K rows illustrate various model outputs on each input test dataset. A Friedman test is realized on ranked data, and each row from floating-point outputs of each model are rank transformed to ranking row. For instance, each raw data row $\left(X_{ji}; j = 1, \cdots, k, i = 1, \cdots, m\right)$ is transformed into ranking entry rows, so if outputs are lower a lower ranking will be applied, and the highest rank will be awarded to the best output. The sum of all the ranks from each row ($R_j = \sum_{i=1}^{k} r_{ji}$) is used to measure Friedman test (X_F^2), as shown in equation 3.27.

$$X_F^2 = \left[\frac{12}{km(m+1)} \sum_{j=1}^{m} R_j^2\right] - 3k(m+1), where\ R_j = \sum_{i=1}^{k} r_{ji} \qquad 3.27$$

Table 3.5: Transforming table of model outputs to rankings during Friedman test

Input	**Model 1**	**Model 2**	**Model 3**	⋯	**Model m**
1	X_{11}	X_{12}	X_{13}	⋯	X_{1m}
2	X_{21}	X_{22}	X_{23}	⋯	X_{2m}
3	X_{31}	X_{32}	X_{33}	⋯	X_{3m}
⋮	⋮	⋮	⋮		⋮
K	X_{k1}	X_{k2}	X_{k3}	⋯	X_{km}
Input	**Model 1**	**Model 2**	**Model 3**	**⋯**	**Model m**
1	r_{11}	r_{12}	r_{13}	⋯	r_{1m}
2	r_{21}	r_{22}	r_{23}	⋯	r_{2m}
3	r_{31}	r_{32}	r_{33}	⋯	r_{3m}
⋮	⋮	⋮	⋮		⋮
K	r_{k1}	r_{k2}	r_{k3}	⋯	r_{km}

Source: Author

2. The level of significance in this case is α= 0.05 and α = 0.1. In addition, the null hypothesis is rejected when the Friedman test X_F^2 outcome is higher than ($X_{m-1;1-\alpha}^2$) chi-square distribution that correspond to the (m−1) degrees of freedom.
3. A decision is made by comparing the critical value (P-value) with the corresponding significance level denoted by α = 0.05 and α = 0.1, according to a common threshold; if the P-value ≤ α, the null hypothesis is rejected.
4. If the null hypothesis is rejected, it is widely recommended to move to perform post-hoc Bonferroni–Dunn pairwise comparison test, in order to determine statistical differences between models in relation to a control model (the best ranking model is selected as the control classification model). Bonferroni–Dunn test considers the performance of multiple models to be significantly various if the corresponding mean of rankings sum is at a certain critical difference (CD), which determines the lowest required variation in rank sums for the pair of groups to differ at the pre-specified α level of significance (Demšar, 2006), as shown in equation 3.28.

$$CD = q_{\propto} \sqrt{{N(N+1)}/{6\,K}} \tag{3.28}$$

where *CD* is critical difference, $q_{\propto}$ is measured based on studentised range statistic with the confidence level α */ (N-1) = α /N* split by √2, and N and K are the number of classifiers and datasets (respectively).

3.6. Proposed Framework Research Design

Section 2.4.2 explained the fundamental phase of structuring the comprehensive audit opinion model. This section outlines the guidelines required to have a rigorous and successful audit opinion model, including utilising more than one dataset, with different sizes; selecting appropriate criteria evaluation techniques for comparing between the performance models; and applying suitable statistical significance testing to prove model performance. The major framework research design steps of the suggested audit opinion model developed in this thesis are displayed in Figure 3.7 and described below.

- **Phase 1 – collecting datasets:** 4 datasets were selected to structure and validate individual classifiers and committee modelling, and another datasets to structure and validate dynamic modelling.
- **Phase 2 – dataset pre-processing:** imputation, normalisation, and feature selection processing of datasets.
- **Phase 3 – dataset splitting:** datasets are split into training and testing datasets for individual classifier and committee combiner model. Clustering datasets to fit with dynamic models.
- **Phase 4 – classifier development and modelling approach:** individual classifier logarithms, Committee combiner models, and dynamic modelling.
- **Phase 5 – applying different evaluation measurement techniques:** to evaluate performance model outcomes' efficiency.
- **Phase 6 – statistical significance testing:** to prove model performance statistical significance.

3.7. Summary

This chapter presented the experimental framework of the proposed audit opinion model designed in this thesis. Developing audit opinion models is not straightforward, requiring dataset selection, data pre-processing and splitting, and training models for validation by evaluation measurements and statistical significance testing. The experimental design framework of the suggested model thus includes sundry fundamental techniques, as described in section 3.6, which will lead to credible and comprehensive experimental design modelling for the audit opinion model. The execution of each phase of the suggested model is expounded in the following chapters.

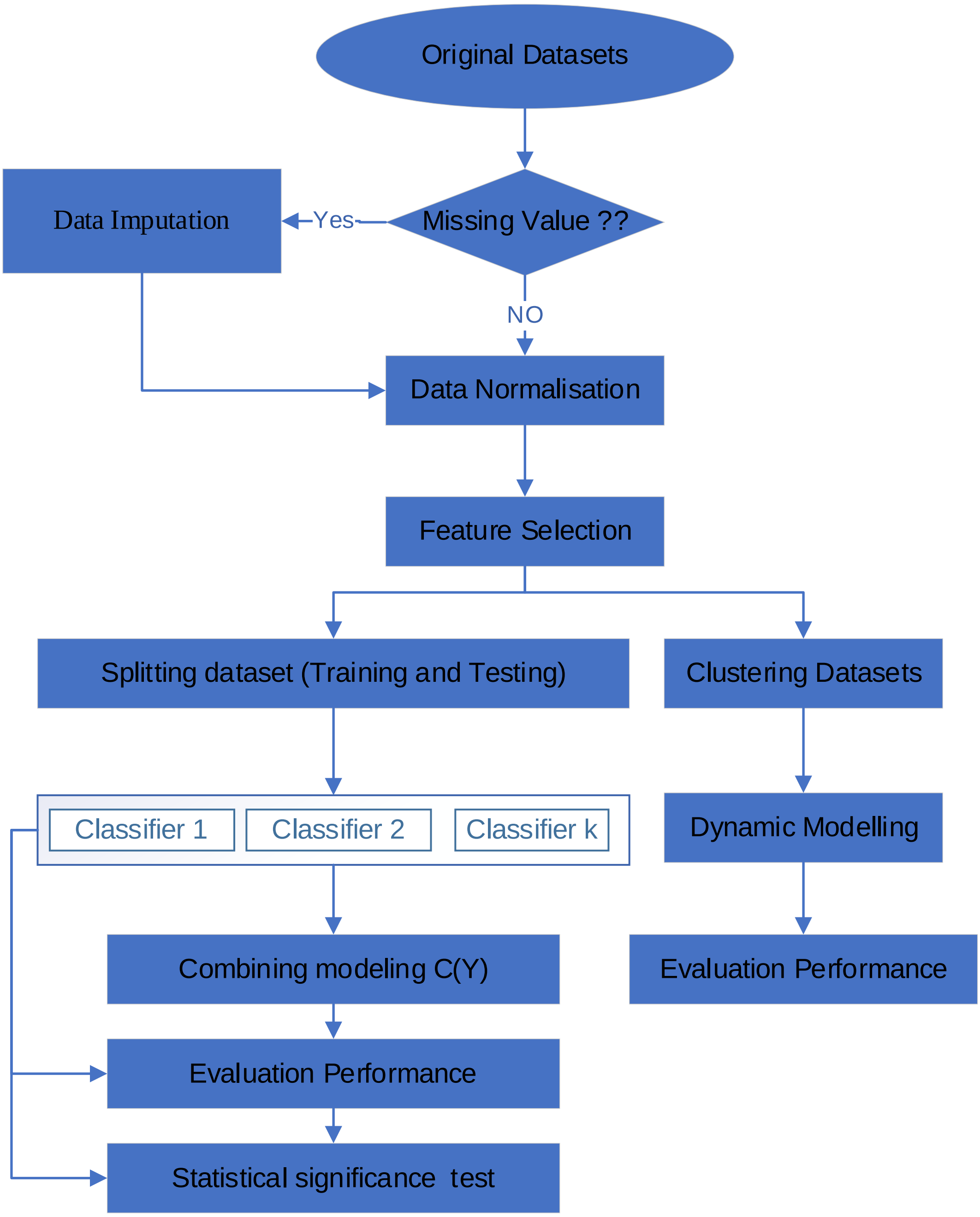


Figure 3.7: Proposed framework research design

Source: Author

Chapter 4
Audit Opinion Using Single Classifier Modelling

4.1. Introduction

Initial BDA implementation regarding DM mechanisms in the audit process remains under-researched, and more investigation is needed to develop classification models utilising classification tools to identify appropriate auditing opinion. Audit opinion models developed by researchers using DM analytics can be used by auditors in practice. As illustrated in the literature review chapter, several studies have contributed to developing predictive models for decision making audit opinion, and testing the ability of ANNs, DTs, LR, NBN, and K-NN yielded significant outcomes, but there is still a need for substantially more effective classification models for audit.

This chapter explains the development and training process of each of the single classifiers (DPL not used in other previous studies, LR, K-NN, NBN, LDA, ANN, SVM, DT and BEC models) and testing abilities based on these classifiers as classification tools to determine audit opinion, by evaluating their performance over four audit opinion datasets. Comparative analysis of evaluation measurement results of the nine classifiers' performance models indicate which classifier has the best capacity to classify audit opinion correctly. Finally, statistical significance testing is applied to the model outputs. The experiments of this study are conducted using MATLAB 2019a version on an 8 GB RAM personal computer with 3.4 GHz, Intel CORE i7, and Microsoft Windows 10 operating system.

4.2. Single Classifiers Development and Experimental Test Results in Classification Audit Opinion

This section presents the details of the simulation and computing platform of each of the nine single classifiers (LR, K-NN, NBN, LDA, DT, ANN, SVM, BEC, and DPL), presenting the performance evaluation test results for each classifier in terms of correctly classifying auditing opinions across the four studied datasets.

4.2.1. Logistic Regression Model

LR is an appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). The *glmfit* function is appropriate for LR classification learning. LR classification model is a specific condition of a generalised linear model, and is more convenient than linear regression for binary data. Because of this, it utilises the fitting procedure that is more the convenient for a binomial distribution to deal with an observed output for a target variable having only two potential classes: 1 (success class), and 0 (failure class). LR coefficients are identified as the statistical weight (mean and variance) of variables by employing the maximum likelihood method through standard optimisation. In addition, logistic link function limits the relationship between the linear predicted probability and the mean and variance of the binomial distribution function (lying between 0 and 1).

The evaluation test comparison results (Table 4.1) reveal that LR achieved the best AUC, average accuracy, F-measurement, sensitivity, specificity, and ROC curve, and the lowest Type I and II Errors and Brier scores for the year 2018 dataset. However, it had higher AURD rates compared to the year 2017 and All-Data datasets, with the worst reliability diagram (Figure 4.2). Furthermore, LR classification results for the All-Data and year 2017 datasets showed a 22.7% difference between sensitivity and specificity rates, and a 21.2% difference between Type I and II Errors (respectively), compared to differences obtained for the year 2019 and year 2018 datasets. The confusion matrix results of the year 2019 and year 2018 datasets exhibited less variance between sensitivity and specificity at 5% and 9.2% rates, respectively. Figure 4.1 shows that the year 2017 dataset and All-Data had worse ROC curves compared to the year 2018 and 2019 datasets, but the best reliability diagrams.

Table 4.1: Evaluation test results of the LR model

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	92.9%	95.4%	95.3%	86.1%
Type II Error	4.9%	1%	1.5%	7%
Type I Error	11%	10.2%	22.7%	29.7%
Specificity	89%	89.8%	77.3%	70.3%
Sensitivity	95.1%	99%	98.5%	93%
AUC	94.5%	96%	94%	89%
F-measurement	92.3%	94.6%	89.1%	83.5%
Brier score	6.9%	4%	4.8%	12.3%
AURD	19.1%	17.9%	11.12%	9.5%

Source: Author

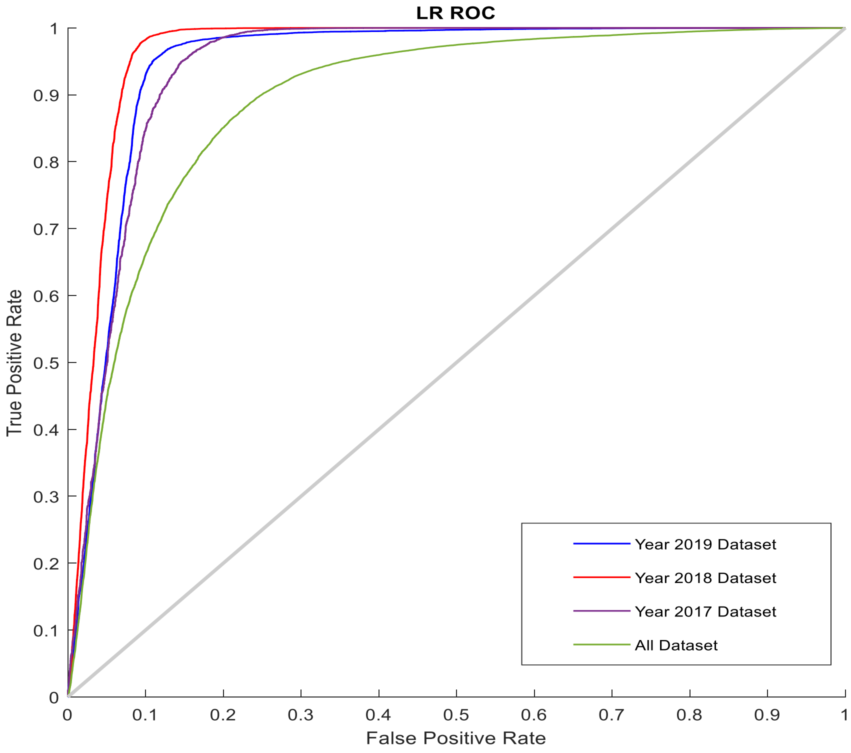


Figure 4.1: ROC curves for the LR model over four datasets

Source: Author

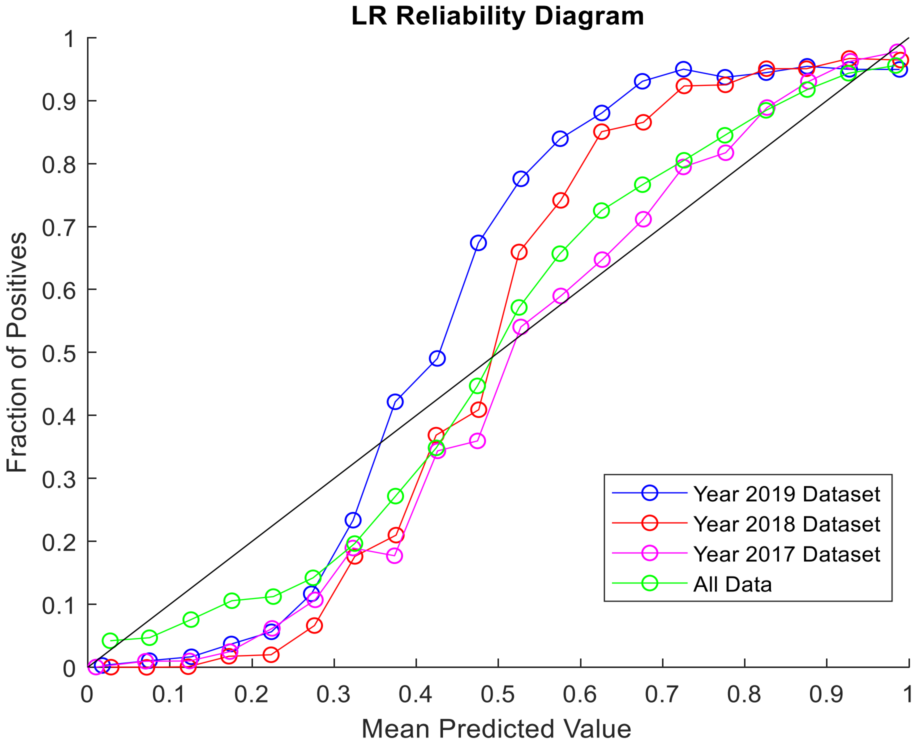


Figure 4.2: LR reliability diagrams over four datasets

Source: Author

4.2.2. K-Nearest Neighbour Model

K-NN model classifies queries according to 10 neighbours, based on their distance to neighbours in the training dataset, offering a simple dynamic path for classifying new points. K-NN classifier training utilises *fitcK-NN* function. For the best performance, the K-NN model type was selected as weighted K-NN, using distance weight squared inverse to identify distance. K-NN is given a group of objects for which a class is known, then the closest K neighbour (K = 1 or K = 0) in an object to a query neighbour or group of neighbours is measured, utilising the distance metric (Euclidean). The model specifies standardised data as true, and specification of K = 1 determines a new observation for the most popular class of that nearest neighbour.

K-NN model results for over four datasets are presented in Table 4.2 that demonstrate incorrectly classified proportions (Type I and II Error rates) lower 9% for qualified and 25% for unqualified audit opinion, higher 91% correct unqualified (sensitivity) and 75% classified qualified firms (specificity). Over the four datasets, K-NN had average accuracy of 87.9-92.3%, F-measurements of 87-91.5%, AUC of 92-93.6%, Brier score under 6.3%, and AURD of under 14.44%. In addition, Figure 4.3 shows K-NN had good reliability diagrams, especially for the year 2019 and All-Data datasets, close to the diagonal line. On the other hand, as seen from Table 4.2, there was a big gap between Type I and Type II Error rates, and ROC curves (Figure 4.3) are not close to the corner. Consequently, K-NN is an unbalanced model, lacking the ability to distinguish between qualified and unqualified audit opinion.

Table 4.2 illustrates that K-NN had better evaluation results for the year 2017 dataset, with the highest F-measurement (91.5%), AUC (91.8%), average accuracy (92.3%), and number of qualified and unqualified companies correctly classified, with the lowest number of incorrectly classified audit opinion, and a Brier score at 6.2%. However, K-NN had the worst AURD for the year 2017 dataset (Figure 4.8). For the year 2019, 2018, and All-Data datasets, K-NN achieved similar results for eight parameters, such as average accuracy rates of 88.2%, 88%, and 87.9%, respectively. Likewise, the year 2019, 2018 and All-Data datasets had reliability diagrams relatively close to the diagonal line (Figure 4.4), and better ROC curves compared to the year 2017 dataset (Figure 4.3).

Table 4.2: Evaluation test results of the K-NN model

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	88.2%	88%	92.3%	87.9%
Type II Error	7%	8.6%	3%	6.6%
Type I Error	20.6%	23%	17.9%	24.8%
Specificity	79.4%	77%	82.1%	75.2%
Sensitivity	93%	91.4%	98%	93.4%
AUC	93.6%	92.6%	91.8%	92%
F-measurement	87.1%	85.3%	91.5%	88%
Brier score	9%	9.25%	6.2%	10.4%
AURD	6.54%	8.67%	14.43%	5.28%

Source: Author

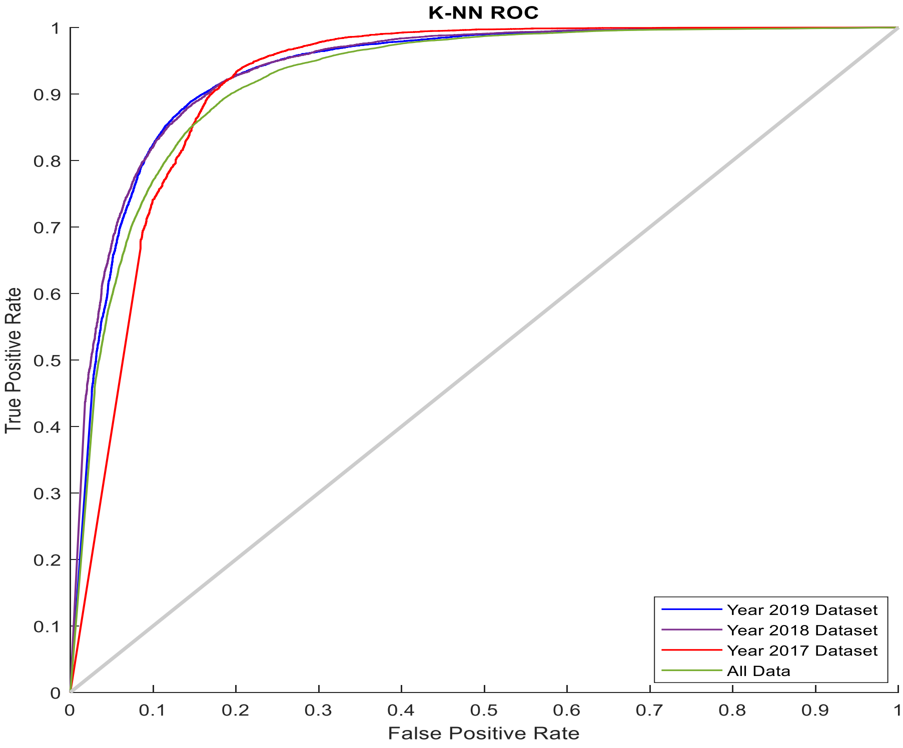


Figure 4.3: ROC curves for the K-NN model over four datasets

Source: Author

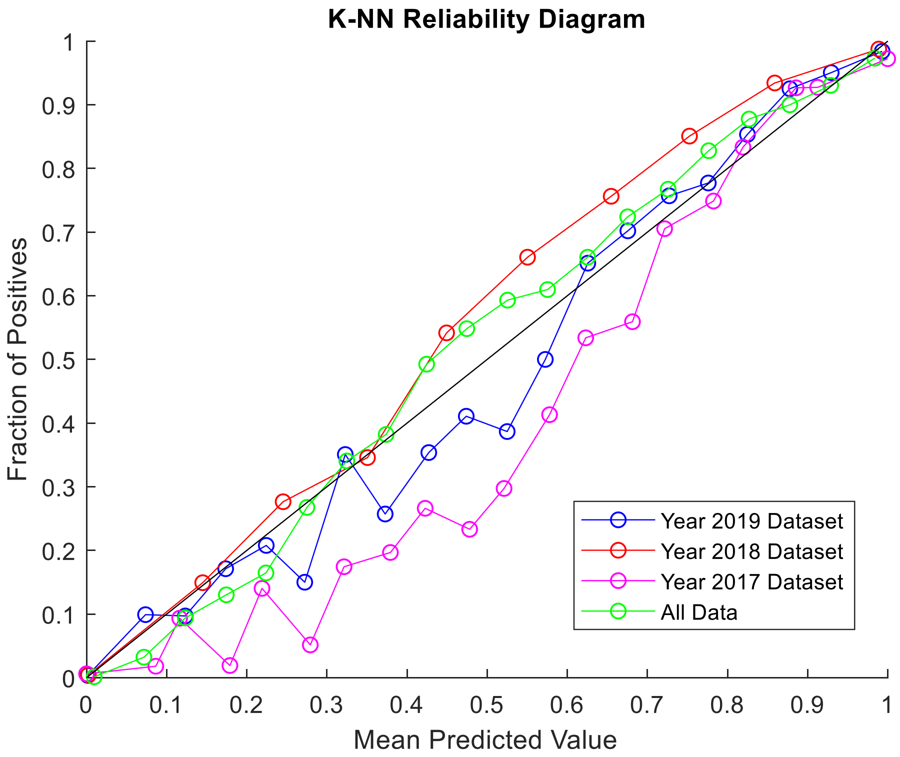


Figure 4.4: K-NN reliability diagrams over four datasets

Source: Author

4.2.3. Naïve Bayes Networks Model

NBN model is a simple probabilistic classifier based on Bayes theory, whereby decisions are made about statistical probability in regard to predicting an event based on learning form past proceedings. NBN classifier uses Gaussian Naïve Bayes, determining optional comma-separated pairs of distribution names per predictor (if numerical predictors), specified as normal (Gaussian distribution); and multivariate multinomial distribution for categorical predictors. If numerical data is passing to *fitcnb* function, the Gaussian (Kernel Function) is substituted. The kernel smoothing type is normal, distribution parameter estimates as mutability 2 (class label) to the number of independent input variables (n = 34) per cell. The first row is the mean and the second row is the standard deviation for the independent input variable. NBN classifier training identified the kernel smoothing density support as the comma-separated pair consisting of unbounded (all real values). In training the classifier specifies the class name as 0 and 1.

The partitioned model is generated using cross-validated DA classifier (*crossval*), utilising five K-fold for predicted groups to obtain pragmatic sensibility of predictive model accuracy in actual practice, and measuring validation scores and predictions through *kfoldPredic*, validation accuracy by 1- *kfoldLoss* (classification loss that calculates a predictive error of the classification model), and accuracy out of 1.

Table 4.3 presents the evaluation test measurement results for NBN over the four datasets. It can be seen that NBN achieved the best results for the 2018 dataset, while it had the weakest for the year 2019 dataset, including 75.8% average accuracy rate, 85% AUC, 54.6% specificity, 90.7% sensitivity, 76.8% F-measurement, and higher percentages for Type I and II Error, Brier score, and AURD. Additionally, Figure 4.5 illustrate that NBN had the worst ROC curve, not near to the corner 1 across all four datasets. Figure 4.6 show that over four datasets the NBN got the worst reliability diagram because not relatively with diagonal line this led to get high rates of AURD that ranging from 22.45-38.4%. In addition, Table 4.3 that year 2019, 2017 datasets and All-Data show that the NBN model as unbalance model to classifying the correct audit opinion due to big gap between the specificity and sensitivity percentages as well as difference between false positive rate and false negative rate. As can observed from the analysis that NBN model is not a good classification tool to classify audit opinion correctly.

Table 4.3: Evaluation test results of the NBN model

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	75.8%	92.3%	91.2%	86.5%
Type II Error	9.3%	2.2%	4.8%	7.5%
Type I Error	45.4%	17.5%	26.7%	23%
Specificity	54.6%	82.5%	73.3%	77%
Sensitivity	90.7%	97.8%	95.2%	92.5%
AUC	85%	94.1%	92.9%	89.6%
F-measurement	76.8%	91.7%	85.8%	85.81%
Brier score	22%	6.9%	7.6%	11.4%
AURD	38.4%	22.45%	26.9%	29.77%

Source: Author

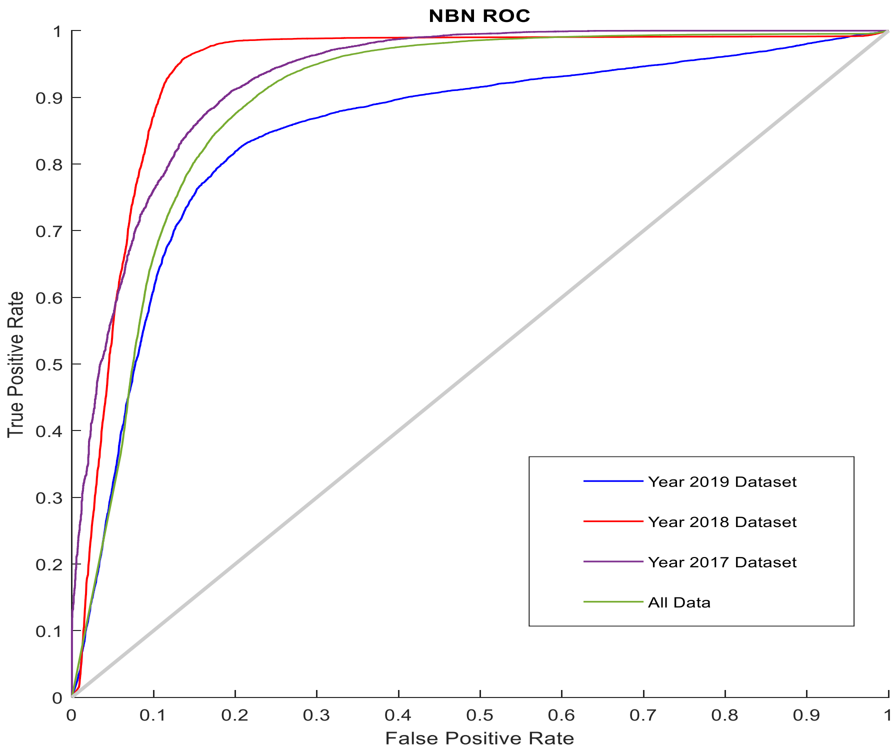


Figure 4.5: ROC curves for the NBN model over four datasets

Source: Author

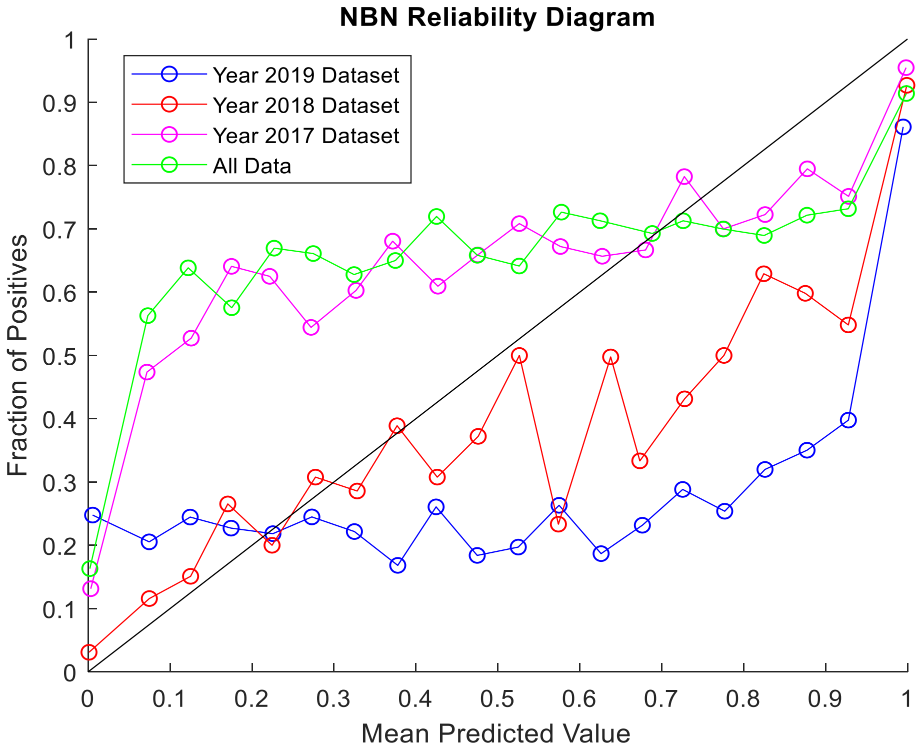


Figure 4.6: NBN reliability diagrams over four datasets

Source: Author

4.2.4. Linear Discriminant Analysis Model

LDA has many assumptions and restrictions compared to LR, and most previous studies showed that it is more accurate than LR when data sample sizes are equal, with homogeneity or equal covariance or variance. To create an interactive and flexible LDA model for classifying, it is trained using fit discriminant analysis classifier (*fitcdiscr*) in a command line. LDA model determines discriminant type linearly, whereby all classes have an equal covariance matrix. In the classifier options, the Gamma is set as 0, which means a model predicts and utilises an unrestricted, experimental covariance matrix. The default *FillCoeffs* is "off". A partitioned model is generated by cross-validated DA classifier (*crossval*), utilising 5 K-fold on predicted groups to obtain the pragmatic sensibility of how accurate the predictive model can be in actual practice. Validation scores and predictions are then measured using *kfoldPredic* and 1- *kfoldLoss* (classification loss calculating the predictive error of the classification model) in order to get the accuracy for the model out of 1.

LDA model evaluation test results are shown in Table 4.4. It can be seen that it achieved the best results for the year 2017 dataset, with higher average accuracy, F-measurement, AUC, sensitivity, and specificity rates, and lower Brier score and Type I and II Error rates. It had the poorest performance for the All-Data dataset, with the lowest AUC, average accuracy, F-measurement, sensitivity, and specificity, and higher Brier score and Type I and II Error rates. For All-Data there was a 33.6% difference between Type I and II Error rates, but the best under reliability curves (9.74%), producing the best reliability diagram (Figure 4.8). LDA had under 17% Type I Error (the number of unqualified firms being flagged as qualified) for the year 2019 and 2018 datasets, and 7% Type II Error (qualified firms identified as unqualified). For these datasets there was 94% AUC, 93% of unqualified and qualified companies correctly identified (sensitivity), and Brier scores between 8.0-8.9%, average accuracy of 90.4-88.6%, AURD of 13.21-17% and F-measurement between 88.95-87.3%.

Figure 4.7 shows that LDA had ROC curves not near to corner 1 across all four datasets, indicating low true positive rates and high false positive rates, and Figure 4.8 shows the reliability diagrams over all four datasets, indicating good S-shapes. The evaluation results indicate that when the data size is increased, LDA model performance declines, as exemplified by the All-Data and year 2018 datasets; LDA has better performance for the smaller datasets (year 2019 and year 2017).

Table 4.4: Evaluation test results of the LDA model

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	90.4%	88.6%	94.2%	81.6%
Type II Error	7%	7%	1%	6.8%
Type I Error	16%	20%	15%	40.4%
Specificity	84%	80%	85%	59.6%
Sensitivity	93%	93%	99%	93.2%
AUC	93.65%	93.6%	94%	86.5%
F-measurement	88.9%	87.3%	92.5%	79.8%
Brier score	8%	8.9%	4.8%	13.3%
AURD	13.21%	17%	24.03%	9.74%

Source: Author

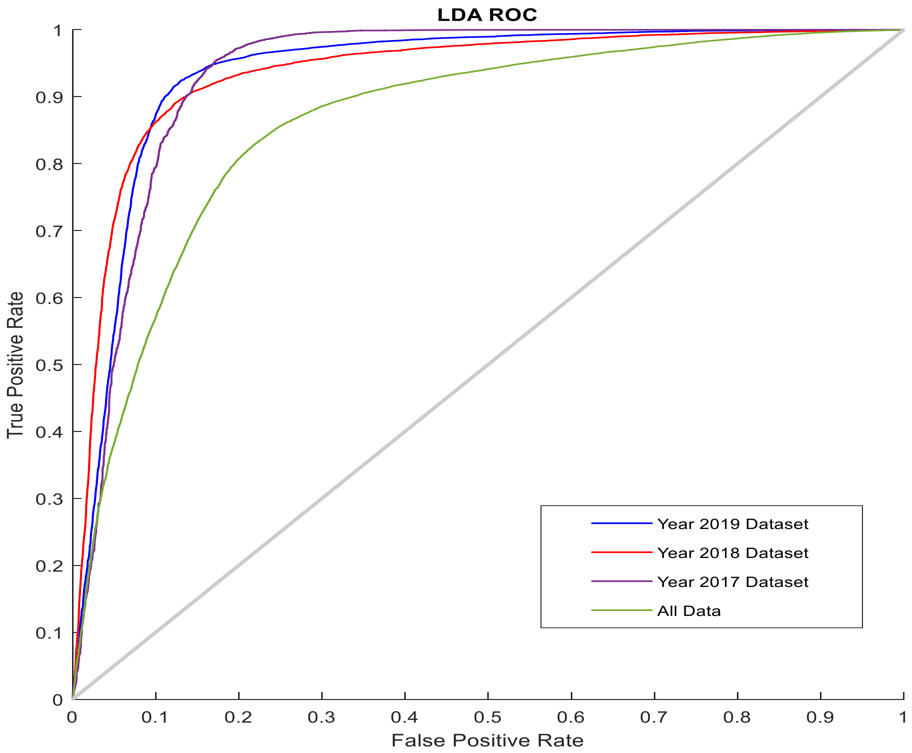


Figure 4.7: ROC curves for the LDA model over four datasets

Source: Author

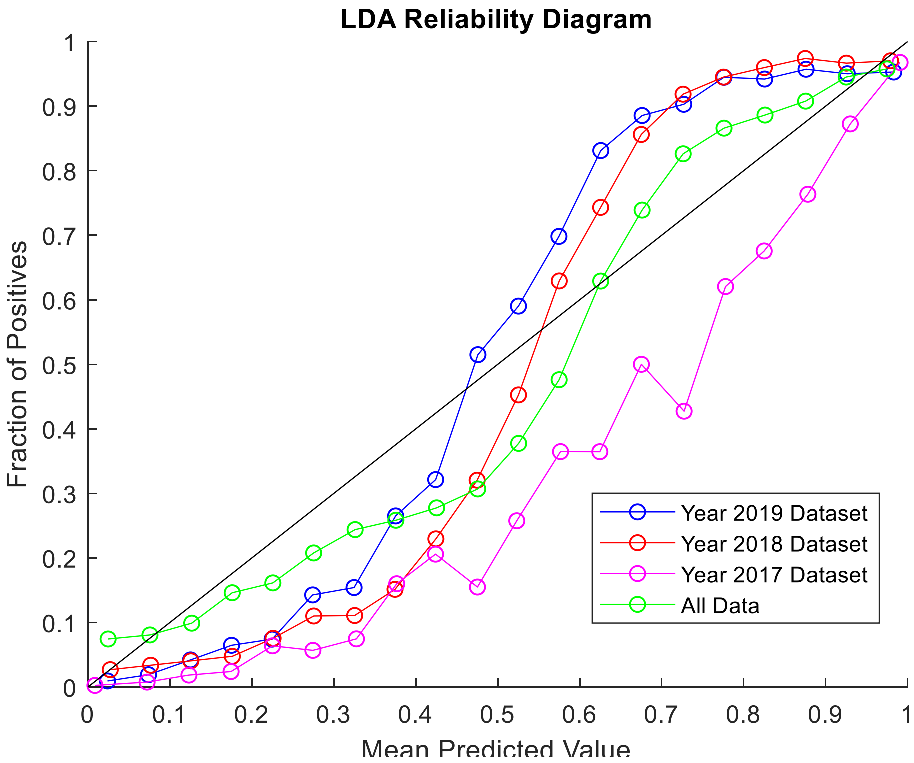


Figure 4.8: LDA reliability diagrams over four datasets

Source: Author

4.2.5. Decision Trees Model

The DT model is one of the easiest and simplest classification models due to its structure. The classification DT model classifier is a medium tree, trained with classifier options utilising *fitctree* (Fit binary DT for multiclass classification). The decision medium tree classifier for deciding when to split nodes is specified by the split criterion of Gini's diversity index (GDI), which is a weighted mean of a classification margin that can deal with two classes in original target data, while utilising surrogate decision splits that may deal with the whole observation in relation to misclassified data, in order to improve predictions. The decision medium tree learner template for all input datasets has 10 MIN parent size and 1 MIN leaf size, with a maximum of 20 branch node splits. Five-fold performance cross-validation is used for the classification DT created from *fitctree* to measure the validation predictions and score for the model, and to compute the accuracy of validation using 1–*kfoldLoss* (incorrect classifications).

Table 4.5 illustrates the evaluation testing results for the DT model performance over the four datasets. These results show acceptable performance in correctly classifying audit opinion with average accuracy and AUC rates greater than 95%, F-measurement from 91.7-95.2%, false negative rate and false positive rates lower than 1.9-16%, sensitivity and specificity rates above 98.1% and 84% (respectively), Brier scores ranging from 2.34-3.71%, and AURD of around 5.54-18.32%.

For the year 2018 dataset (Table 4.5), DT had 96.8% average accuracy, and 99% and 92% unqualified and qualified audit opinion were correctly classified (respectively), with 1% and 8% incorrectly classified unqualified and qualified firms (respectively), Brier score of 2.34%, and AUC of 99% (mean ROC curve near to corner 1, as in Figure 4.9), indicating better DT performance than for the other datasets. On other hand, the year 2018 dataset had a higher AURD, with the worst reliability diagram compared to the year 2019, year 2017, and All-Data datasets (Figure 4.10). DT evaluation test results showed that it had weaker performance for the All-Data dataset, with lower average accuracy, specificity, sensitivity, and AUC, in addition to higher Type I and II Errors and Brier score, and the worst ROC (Figure 4.9). Conversely, the best reliability curve and AURD was achieved for All-Data.

Table 4.5: Evaluation test results of the DT model

	Dataset in 2019	**Dataset in 2018**	**Dataset in 2017**	**All-Data**
Average accuracy	96.3%	96.8%	96%	95.1%
Type II Error	1%	1%	1.4%	1.8%
Type I Error	9%	8%	14.1%	15.9%
Specificity	91%	92%	85.9%	84.1%
Sensitivity	99%	99%	98.6%	98.2%
AUC	98%	99%	97.8%	95.6%
F-measurement	95.2%	95.6%	93%	91.7%
Brier score	3.1%	2.34%	3.2%	3.71%
AURD	15.21%	18.32%	12.15%	5.54%

Source: Author

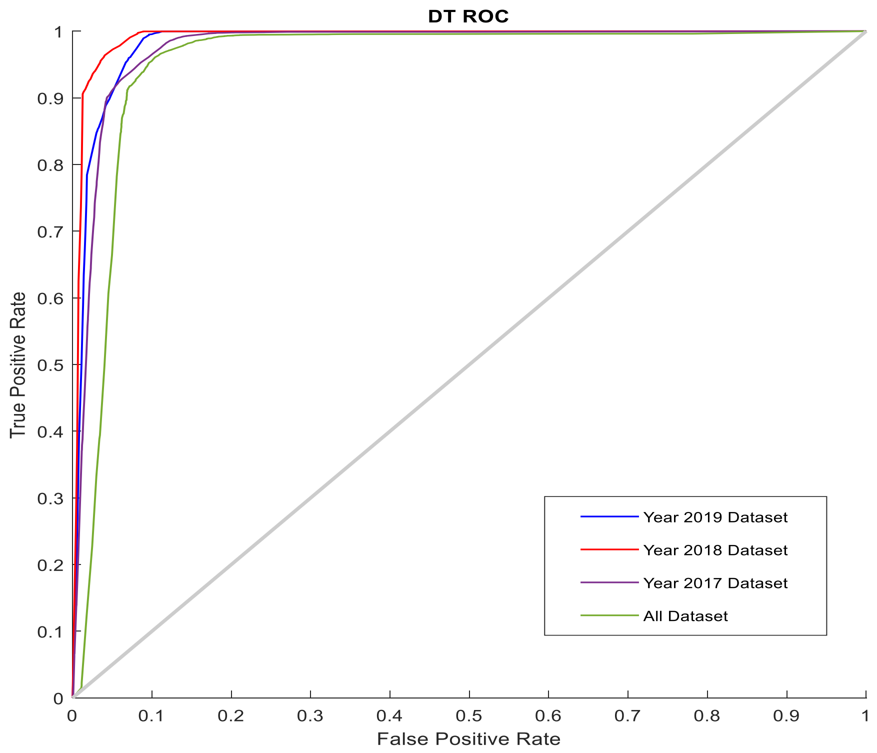


Figure 4.9: ROC curves for DT model over four datasets

Source: Author

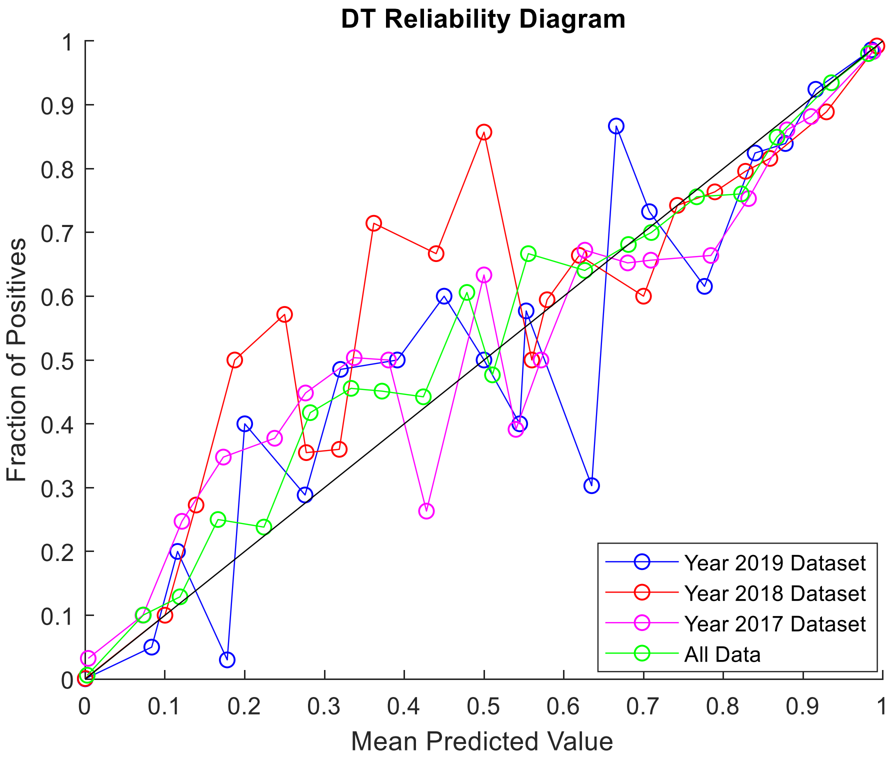


Figure 4.10: DT reliability diagrams over four datasets

Source: Author

4.2.6. Artificial Neural Network Model

A major problem in enhancing the ANN classifier is determining an appropriate order of training, transfer, learning function, learning speed, and network construction topology in terms of hidden neuron numbers in hidden layers. ANN classifier was enhanced using back propagation learning algorithm, choosing different parameters and settings based on input data attributes. The transfer functions (sigmoid symmetric and sigmoid positive transfer, which measure layer outcomes from net inputs) from the hidden layer were selected to be hyperbolic tangent sigmoid transfer function (*tansig*), the most popular transfer functions (equation 4.1) utilised in ANN models to compute output layers from net inputs.

$$A = \frac{2}{(1+exp(-2*n))-1} \qquad 4.1$$

The training function aims to train a network by updating input weight and bias input variables in order to obtain an optimal performance output value. The training functions *trainscg*, *trainbr*, and *trainlm* were used from MATLAB toolbox. For the four datasets, the default *trainscg* was used by changing the training function type, by which the NN is trained to have better performance for these datasets. *Trainscg* can train any network as long as its net input, weight, and transfer functions have derivative functions. For the NN structure shown in Figure 4.11, one hidden layer is utilised, whose size is based on model complexity. For the year 2019, 2018, and 2017 datasets, the hidden layer size was selected as 38, while it was 36 for the All-Data dataset, which gives better performance by being calibrated for the input. Setup division function was used to divide mode type as a sample for three partitions: training, testing, and validation. NN model was then performed by cross-entropy.

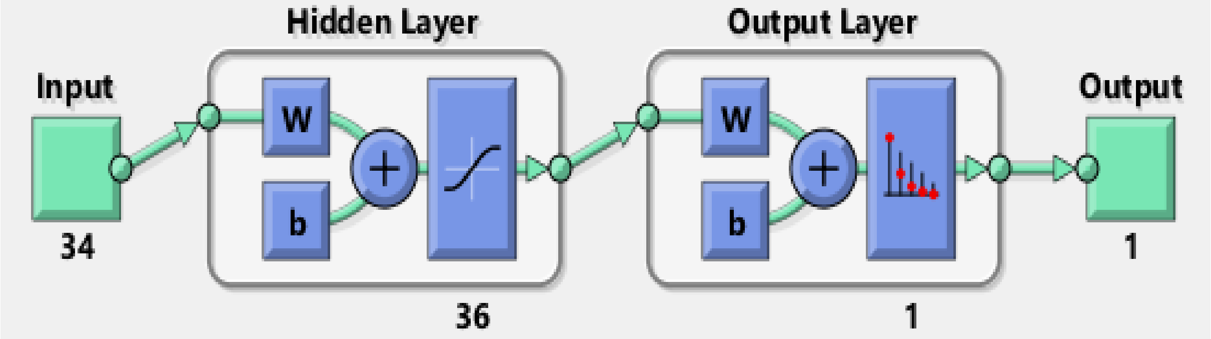


Figure 4.11: ANN structure

Source: Author

Table 4.6 illustrates the evaluation results over the four datasets using ANN model to classify audit opinion. For the All-Data, year 2017, year 2018 to year 2019 dataset there was an increase the evaluation results such as the average accuracy rates from 92.3% to 95.9%, AUC rates 94.8% to 96.73%, F-measurement 89.9% to 95.3% and specificity percentages

81.3% to 90.9%, and the number of unqualified firms incorrectly classified to qualified class (Type I Error) decreased from 18.7% to 9.1% and Brier scores from 6-3.5%. In addition, Table 4.6 illustrates that the ANN model has good ability to classify audit opinion, because over the four datasets it achieved above 92.2% average accuracy, 94.7% AUC, 89.8% F-measurement, and 97% sensitivity rates; furthermore, less than 3% of qualified firms were incorrectly classified as unqualified, and it had under 6.1% Brier scores, and under 8% AURD. Likewise, for the year 2019 and 2018 datasets, it had under 10% Type I Error and above 90% correctly qualified companies, with lower differences between specificity and sensitivity rates (approximately 8.5%).

Likewise, the ANN reliability diagrams (Figure 4.13) illustrate that ANN dedicated prediction values close to the actual target (indicated by proximity to the diagonal line). Figure 4.12 shows that ANN has good ROC curves for the year 2019 and 2018 datasets, but poorer ones for the year 2017 and All-Data datasets, which have higher FP rates.

Table 4.6: Evaluation test results of the ANN model

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	95.9%	95.6%	95.3%	92.3%
Type II Error	0.7%	1.5%	0.8%	2.9%
Type I Error	9.1%	9.7%	18.7%	18.7%
Specificity	90.9%	90.3%	81.3%	81.3%
Sensitivity	99.3%	98.5%	99.2%	97.1%
AUC	96.73%	96.7%	94.9%	94.83%
F-measurement	95.3%	94.6%	91.1%	89.9%
Brier score	3.5%	3.7%	3.98%	6%
AURD	6.8%	7.9%	6.74%	5.55%

Source: Author

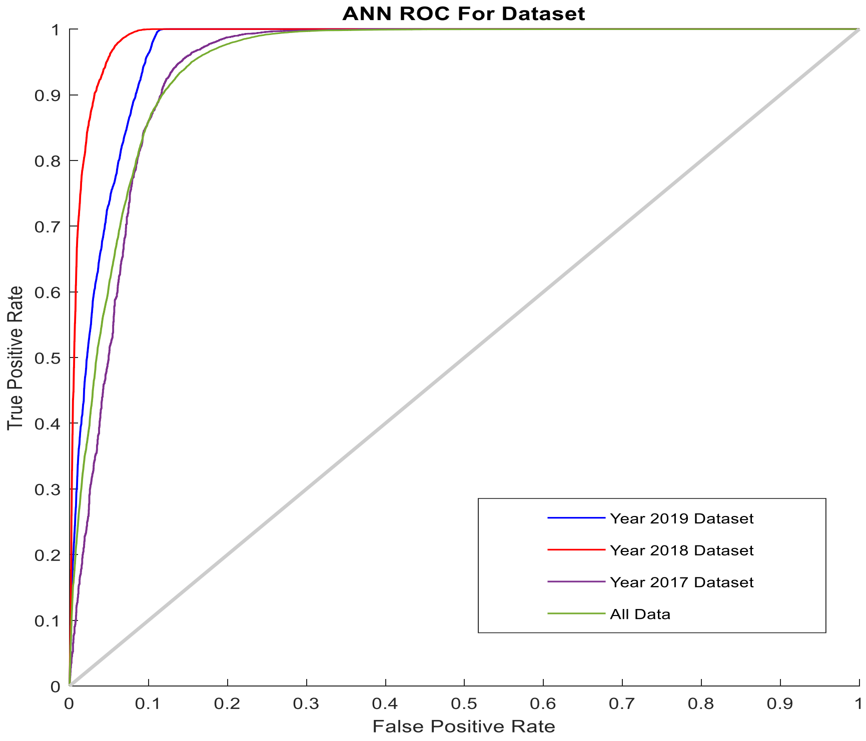


Figure 4.12: ROC curves for ANN model over four datasets

Source: Author

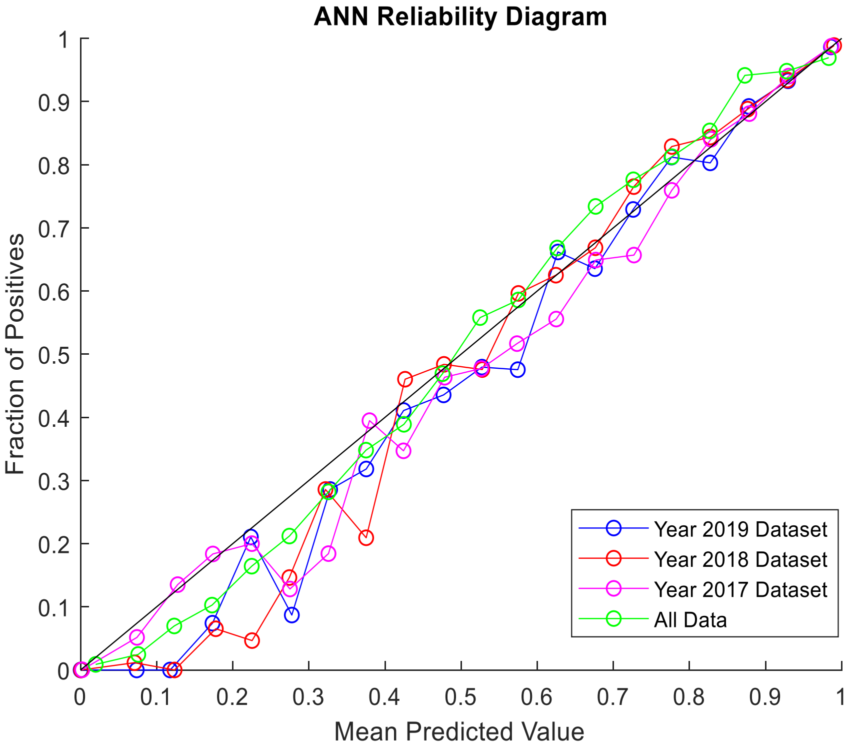


Figure 4.13: ANN reliability diagrams over four datasets

Source: Author

4.2.7. Support Vector Machine Model

SVM classifier classify data by identification of linear optimal hyperplane split data, with a maximum margin between the hyperplane and the nearest point to procedure binary classification (0 and 1), utilising kernel function as summarised in the flowing points:

- Type of SVM: Quadratic
- Kernel function: polynomial
- Polynomial order: 2
- Kernel scale: automatic
- Box constraint level: 1
- Standardise data: true

SVM function (sigma) selected the suitable kernel scale automatically, thus each dataset had different kernel scale parameters: 2.9239 (2019), 2.9153 (2018), 3.3053 (2017), and 5.8 (All-Data). Creating performance cross-validation SVM was used for training classification SVM utilising 5-fold cross-validation in order to measure the validation prediction and score prediction for the model, and accuracy of validation was measured using 1 – *kfoldLoss* (incorrect classifications).

Table 4.7 displays that for the year 2018 and 2017 datasets, SVM had similar average accuracy, and correctly and incorrectly unqualified classification. For the year 2019 and 2018 datasets, SVM achieved 96% AUC, Brier scores of around 3%, sensitivity around 99% and AURD around 22%. For All-Data it had weaker performance, with the lowest average accuracy (90.8%), AUC (94%), sensitivity (97%), and specificity (77%); and higher Brier score (6.5%), false positive (23%), and false negative (3%) rates. All-Data realised the best AURD (5.33%) compared to the other datasets.

Figure 4.14 displays the ROC curves, illustrating that SVM achieved better curves for the year 2019 and 2018 datasets, with the best Type I Error and sensitivity rates. Figure 4.15 shows SVM reliability diagrams, showing that the best was for the All-Data dataset (close to the diagonal line), due to which All-Data has lower AURD, but the other reliability diagrams for the yearly datasets are under the diagonal line, indicating over-forecasting model.

Table 4.7: Evaluation test results of the SVM model

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	95.4%	95.8%	95.7%	90.8%
Type II Error	1.5%	1%	1%	3%
Type I Error	10.3%	12%	20%	23%
Specificity	89.7%	88%	80%	77%
Sensitivity	98.5%	99%	99%	97%
AUC	96%	96%	95%	94%
Brier score	4.27%	3.2%	3.3%	6.5%
F-measurement	94.3%	93.8%	90.4%	88.2%
AURD	22.02%	22.4%	26.8%	5.33%

Source: Author

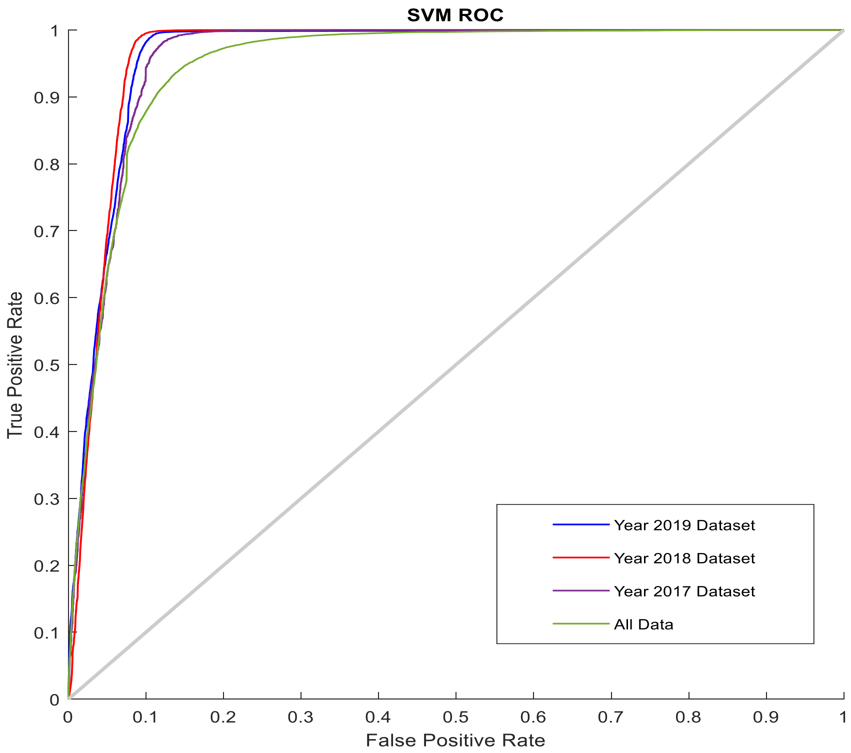


Figure 4.14: ROC curves for the SVM model over four datasets

Source: Author

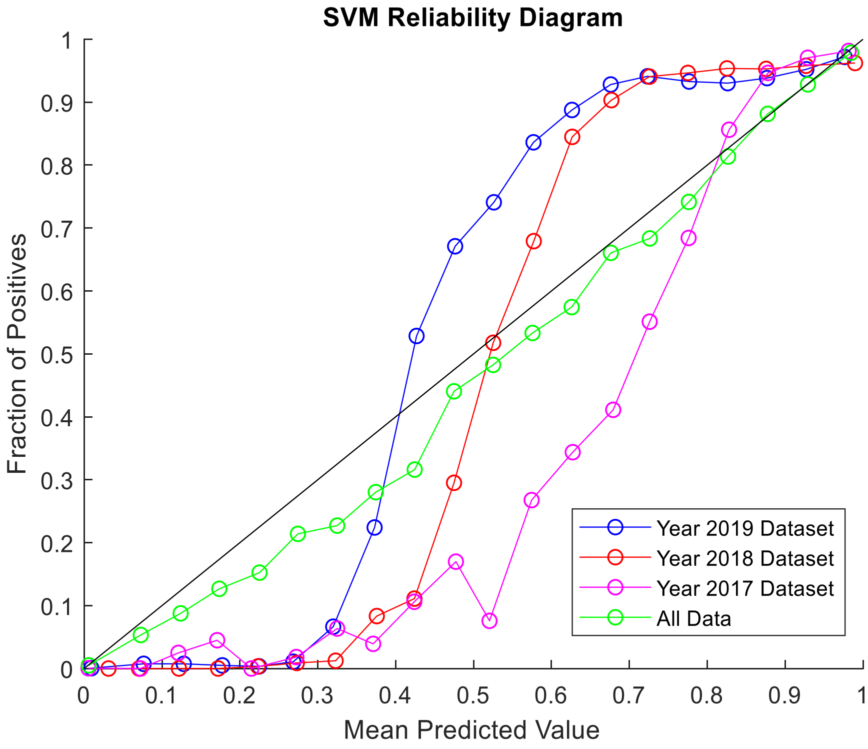


Figure 4.15: SVM reliability diagrams over four datasets

Source: Author

4.2.8. Boosting Ensemble Classifier Model

BEC model assists to enhance the performance of the individual classifier machine learning outcomes through combining the outputs of weaker classifier models to produce better predictions. The ensemble boost tree model training classification used *fitcensemble* with AdaBoostM1 method, one of the best boosting algorithm methods that concentrates on strengthening weak learners by combining the outcomes of their weighted sum classifications to represent the final output of the ensemble boost tree classifier. A boosting algorithm identifies a weak learner as a “tree” and trains 30 learners by satisfied 0.1 learning rate, because of a slower rate of convergence to high-standard solution. AdaBoostM1 predicts new data for weak learners, identifying if a sum is 1, or 0 class is predicted. The ensemble boost fits the DT learner template for all input debates in boosted DT during training classification as 2 MIN parent size and 1 MIN leaf size, with a maximum if 20 splits. The final step is creating cross-validation from the classification ensemble boosted tree utilising 5-fold cross-validation, then the validation predictions and model scores for accuracy are measured using 1 – *kfoldLoss* (incorrect classifications).

Over all four datasets, the evaluation test results for BEC model (Table 4.8) illustrated that it achieved above 95.2% average accuracy, 97.9% AUC, 93.2% F-measurement, 98.7% sensitivity, and 85.8% specificity, with lower 14.2% Type I Error rates and 1.3% Type II Error rates, and Brier scores and AURD of less than 3.7% and 12% (respectively). Likewise, there were fewer differences between the sensitivity and specificity rates for the datasets: 2019 (7.8%), 2018 (6.5%), 2017 (13.4%), and All-Data (11.8%). In addition, there were smaller gaps between the number of correct classified qualified and unqualified companies: 2019 (7.8 %), 2018 (6.5%), 2017 (13.4%), and All-Data (11.8%). These results indicate that the ensemble classifier model has significant performance ability to classify audit opinion correctly.

Figure 4.16 shows that BEC has powerful ROC curves near to corner 1, due to lower Type I Error rates across all four datasets, with high percentages of sensitivity rates. Figure 4.17 represents that BEC reliability diagrams across all four datasets are relatively close to the diagonal line. These results support BEC model’s ability to deduce correct audit opinion.

Table 4.8: Evaluation test results of the BEC model

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	96.8%	97%	96.6%	95.3%
Type II Error	0.6%	0.5%	0.7%	1.2%
Type I Error	8.4%	7%	14.1%	13%
Specificity	91.6%	93%	85.9%	87%
Sensitivity	99.4%	99.5%	99.3%	98.8%
AUC	98.7%	99%	98%	98%
F-measurement	95.7%	96.4%	93.1%	93.3%
Brier score	2.9%	2.3%	2.95%	3.6%
AURD	11.39%	11.48%	11.93%	11.33%

Source: Author

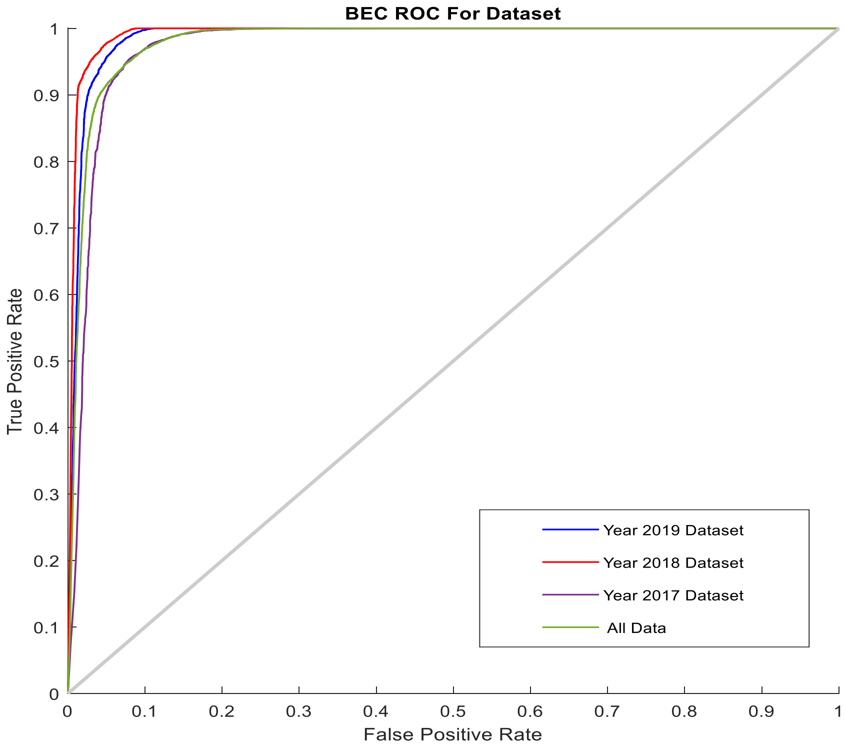


Figure 4.16: ROC curves for the BEC model over four datasets

Source: Author

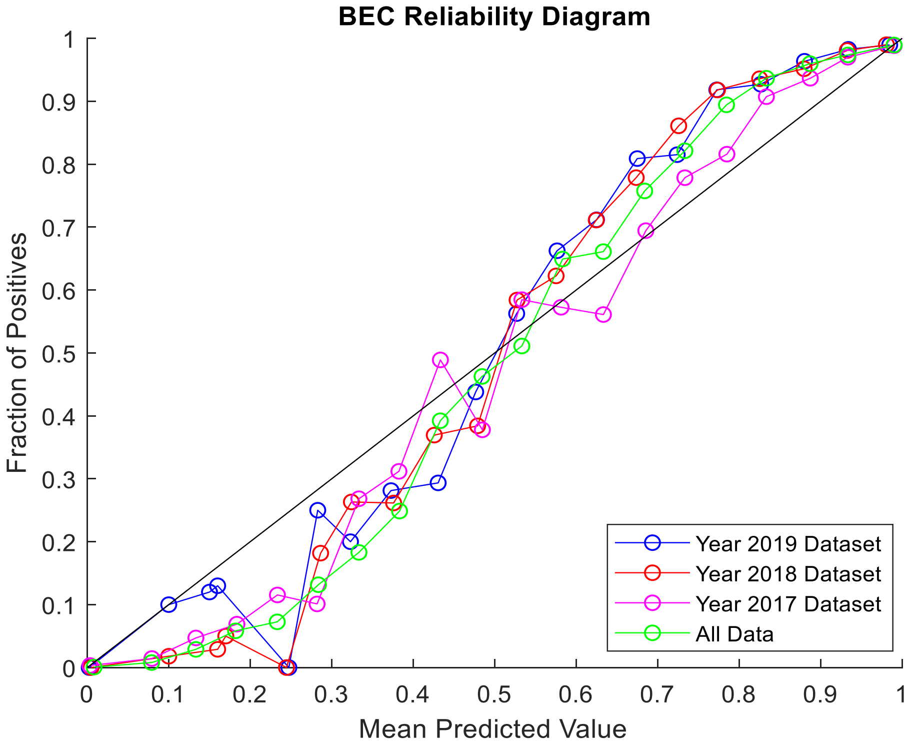


Figure 4.17: BEC reliability diagrams model over four datasets

Source: Author

4.2.9. Deep Learning Model

Deep learning (DPL) algorithm uses deep or multi-layer attributes to extract ingrained features in data with loud representation of characteristic classes, through the attribute of minimum levels feature, to detect massive amounts of construction in data. For an enhanced simulation and computing platform, DPL uses LSTM characteristics with eight layers. The hidden sequence input layer includes 34 independent variables. Two LSTM layers (LSTM layer unit 1, n = 34, LSTM layer unit 2, n = 40), with specified output mode as a sequence for LSTM layers recall only the significant sides of the input sequence. Two dropout layers at 0.2 probability in order to defining the next layer input elements to 0 and to avoid network sentience of the tiny group of neurons in the layer. Connecting all the neurons in a previous layer by using a fully connected layer and then a *softmax* layer (to normalise the fully connected layer's output) creates prediction possibility output for each class, consisting of positive figures that sum to one. The classification layer then utilises the predictions from the *softmax* layer for each input, to specify an input to one of the mutually exclusive classes, to calculate the final error class.

After building the LSTM layers, DPL training options are constructed by generating a set of DPL training options utilising Adam optimiser, with a defined MAX number of 3000 and a min batch of 1000 observations at each iteration, with a gradient threshold of 1. The longest sequence length is used to make each mini batch, to hold the same length for the longest sequence, and the dataset is recalled holding the same length for the longest sequences, assuring that a dataset stays arranged by sequence length, set to every epoch shuffle in order to avoid ignoring the same data in every epoch. In training options, the execution environment is specified to be *Auto* (if the GPU is available it is utilised for training; otherwise, CPU is used). After building the training network and options, the DPL network is created by training the LSTM network with training options, input dataset and targets data by using *trainNetwork*. To predict the final output from the DPL classifier model, the *classify* function is applied to input data with the output of *trainNetwork.*

Table 4.9 illustrates the evaluation test results of the DPL model across the four tested datasets, for which the model is balanced, with few differences between sensitivity and specificity rates for the studied datasets: 2019 (4%), 2018 (4.9%), 2017 (11%), and All-Data (10.2%). Moreover, the results show that DPL has powerful classifier performance ability to classify audit opinion, obtaining average accuracy rates ranging between 97.8-95.5%, sensitivity rate above 99%, specificity rate 88.9%-95.7%%, AUC rates of 98.2-99.3%, and F-measurement from 94.3-97.7%, with lower Type I and II Error rates (incorrectly flagged as other classes), and Brier scores and AURD of less than 3.5% and 7%, respectively. The ROC curves for DPL were close to 1 (Figure 4.1), and the reliability curves were relatively proximal to the diagonal line (Figure 4.2). In sum, DPL is effective in distinguishing audit opinion.

Overall, Table 4.9 shows that DPL had the best performance for the year 2019 dataset, with higher average accuracy rate (97.8%), specificity (95.7%), sensitivity (99.7%), and F-measurement (97.7%), with lower Type I and II Error rates (4.3% and 0.3%, respectively), and Brier score (1.17%). It has the best reliability diagram for the All-Data dataset, but the latter had the worst evaluation test results, including its ROC curve, which is attributable to the larger size of the dataset undermining model performance.

Table 4.9: Evaluation test results of the DPL model

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	97.8%	97.6%	97.2%	95.5%
Type II Error	0.3%	0.4%	0.1%	0.8%
Type I Error	4.3%	5.3%	11.1%	11%
Specificity	95.7%	94.7%	88.9%	89%
Sensitivity	99.7%	99.6%	99.9%	99.2%
AUC	98.9%	99.3%	98.8%	98.2%
F-measurement	97.7%	97.2%	94.7%	94.3%
Brier score	1.17%	1.5%	1.9%	3.4%
AURD	5.72%	6.1%	6.56%	4.63%

Source: Author

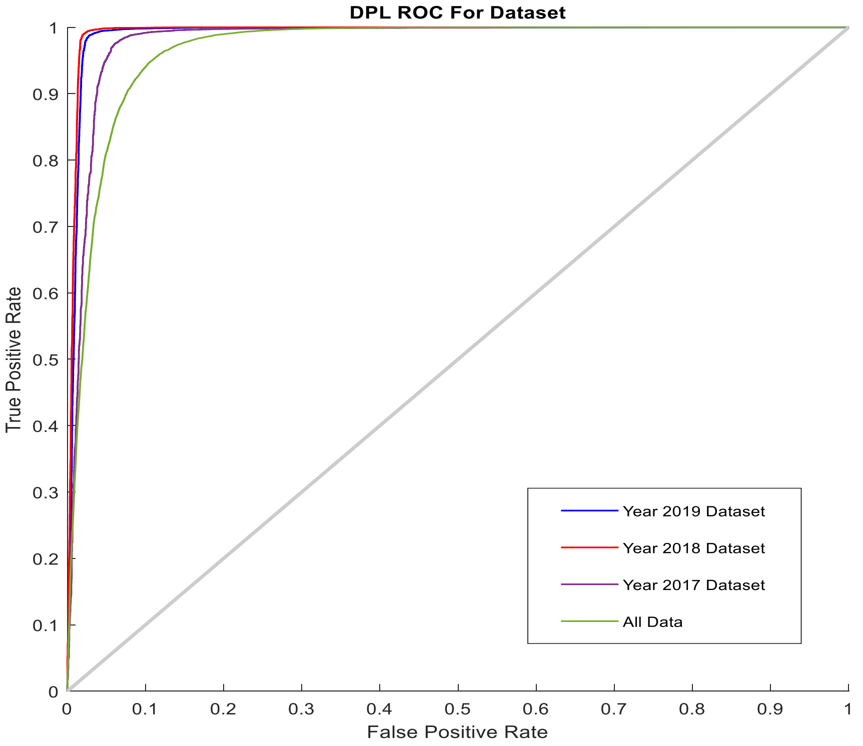


Figure 4.18: ROC curves for DPL model over four datasets

Source: Author

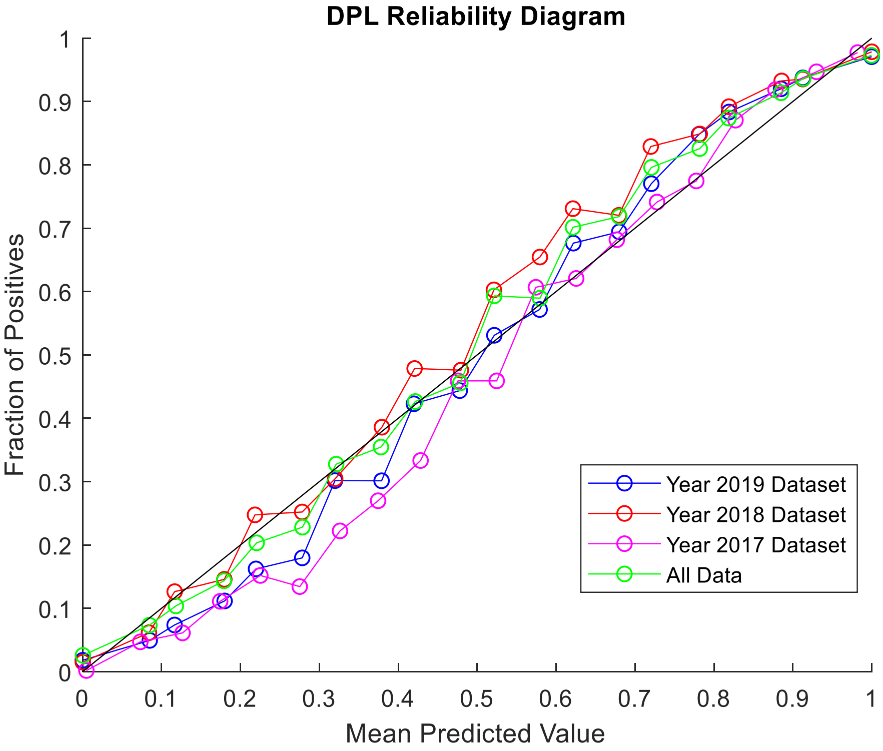


Figure 4.19: DPL reliability diagrams over four datasets

Source: Author

4.3. Comparative Analysis and Discussion

This section analyses the valuation test results achieved by the DPL, ANN, BEC, SVM, K-NN, NBN, DT, LDA, and LR models, to ascertain which predicted audit opinions better. Tables 4.10-4.13 illustrate the valuation test results used to compare between the performance of all nine models.

Table 4.10: Comparing evaluation test results of all models with 2019 dataset

	Year 2019 Dataset								
	Aver Acc.	Type II err.	Type I err.	Specificity	Sensitivity	F-measure	AUC	Brier score	AURD
LR	92.9%	4.9%	11%	89%	95.1%	92.3%	94.5%	6.9%	19.1%
K-NN	88.2%	7%	20.6%	79.4%	93%	87.1%	93.6%	9%	6.54%
NBN	75.8%	9.3%	45.4%	54.6%	90.7%	76.8%	85%	22%	38.4%
LDA	90.4%	7%	16%	84%	93%	88.9%	93.65%	8%	13.21%
DT	96.3%	1%	9%	91%	99%	95.2%	98%	3.1%	15.21%
ANN	95.9%	0.7%	9.1%	90.9%	99.3%	95.3%	96.7%	3.5%	6.8%
SVM	95.4%	1.5%	10.3%	89.7%	98.5%	94.3%	96%	4.27%	22.02%
BEC	96.8%	0.6%	8.4%	91.6%	99.4%	95.7%	98.7%	2.9%	11.39%
DPL	**97.8%**	**0.3%**	**4.3%**	**95.7%**	**99.7%**	**97.7%**	**98.9%**	**1.17%**	**5.72%**

Source: Author

Table 4.11: Comparing evaluation test results of all models with 2018 dataset

	Year 2018 Dataset								
	Aver Acc.	Type II err.	Type I err.	Specificity	Sensitivity	F-measure	AUC	Brier score	AURD
LR	95.4%	1%	10.2%	89.8%	99%	94.6%	96%	4%	17.9%
K-NN	88%	8.6%	23%	77%	91.4%	85.3%	92.6%	9.25%	8.67%
NBN	92.3%	2.2%	17.5%	82.5%	97.8%	91.7%	94.1%	6.9%	22.45%
LDA	88.6%	7%	20%	80%	93%	87.3%	93.6%	8.9%	17%
DT	96.8%	1%	8%	92%	99%	95.6%	99%	2.34%	18.32%
ANN	95.6%	1.5%	9.7%	90.3%	98.5%	94.6%	96.7%	3.7%	7.9%
SVM	95.8%	1%	12%	88%	99%	93.8%	96%	3.2%	22.4%
BEC	97%	0.5%	7%	93%	99.5%	96.4%	99%	2.3%	11.48%
DPL	**97.6%**	**0.4%**	**5.3%**	**94.7%**	**99.6%**	**97.2%**	**99.3%**	**1.5%**	**6.1%**

Source: Author

Table 4.12: Comparing evaluation test results of all models with 2017 dataset

	Year 2017 Dataset								
	Aver Acc.	Type II err.	Type I err.	Specificity	Sensitivity	F-measure	AUC	Brier score	AURD
LR	95.3%	1.5%	22.7%	77.3%	98.5%	89.1%	94%	4.08%	11.12%
K-NN	92.3%	3%	17.9%	82.1%	98%	91.5%	91.8%	6.2%	14.43%
NBN	91.2%	4.8%	26.7%	73.3%	95.2%	85.8%	92.9%	7.6%	26.9%
LDA	94.2%	1%	15%	85%	99%	92.5%	93.6%	4.8%	24.03%
DT	96%	1.3%	14.1%	85.9%	98.7%	93%	97.8%	3.2%	12.15%
ANN	95.3%	0.8%	18.7%	81.3%	99.2%	91.1%	94.9%	3.98%	6.74%
SVM	95.7%	1%	20%	80%	99%	90.4%	95%	3.3%	26.8%
BEC	96.6%	0.7%	12.7%	87.3%	99.3%	93.1%	98%	2.95%	11.93%
DPL	**97.2%**	**0.1%**	**11.1%**	**88.9%**	**99.9%**	**94.7%**	**98.8%**	**1.9%**	**6.56%**

Source: Author

Table 4.13: Comparing evaluation test results of all models with All-Data dataset

	All-Data								
	Aver Acc.	Type II err.	Type I err.	Specificity	Sensitivity	F-measure	AUC	Brier score	AURD
LR	86.1%	7%	29.7%	70.3%	93%	83.5%	89%	12.3 %	9.5%
K-NN	87.9%	6.6%	24.8%	75.2%	93.4%	88%	92%	10.4%	5.28%
NBN	86.5%	7.5%	23%	77%	92.5%	85.81%	89.6%	11.4%	29.77%
LDA	81.6%	6.8%	40.4%	59.6%	93.2%	79.8%	86.5%	13.3%	9.74%
DT	95.1%	1.8%	15.9%	84.1%	98.2%	91.7%	95.6%	3.71%	5.54%
ANN	92.3%	2.9%	18.7%	81.3%	97.1%	89.9%	94.83%	6%	5.55%
SVM	90.8%	3%	23%	77%	97%	88.2%	94%	6.5%	5.33%
BEC	95.3%	1.2%	13%	87%	98.8%	93.3%	98%	3.6%	11.33%
DPL	**95.5%**	**0.8%**	**11%**	**89%**	**99.2%**	**94.3%**	**98.2%**	**3.4%**	**4.63%**

Source: Author

Tables 4.10-4.13 compare the valuation measurement test results of all models across the four datasets, illustrating that the nine models have acceptable performance to correctly classify audit opinion, as indicated by the ranges of average accuracy (75.8-97.8%), F-measure (76.8-97.7%), and AUC (above 84.9%). Additionally, all models had sensitivity greater than 90%, and below 9.5% Type II Error, 22.1% Brier scores, and 38.5% AURDs. However, all models' audit opinion classification performance declined for the All-Data dataset, which had the largest data size, with decreasing average accuracy, F-measure, and AUC rate, and generally increased Type I and II Error and Brier scores.

Overall, all tables comparing the valuation metrics results of all models with the four datasets show that the DPL model outperformed NBN, LR, ANN, K-NN, LR, LDA, BEC, and DT models in terms of obtaining reliability diagrams that calibrated probabilities relatively near to the diagonal line, with under 6.7% AURD, fewer Type I and II Errors (incorrectly flagged as other classes), lower Brier scores (less than 3.5%), and higher average accuracy, F-measure, specificity, sensitivity, and AUC rates (with ROC near to corner 1; DPL had the highest sensitivity rates with the lowest Type I Error rates). Consequently, it can be deduced that DPL model has the most powerful classifier performance to correctly classify audit opinion. In addition, DPL is more balanced, with higher F-measure, and fewer differences between Type I and Type II Error for the datasets: 2019 (4%), 2018 (4.9%), 2017 (11%), and All-Data (10.2%). As seen from Table 4.13, for larger dataset sizes other models had egregiously reduced performance, while DPL continued to have the best capacity to classify audit opinion correctly and distinguish between audit opinion.

As discussed above, over all four datasets, DPL model revealed superior ability to the other models in classifying audit opinion correctly; it is more balanced; and it is better able to distinguish between audit opinions. The valuation measurement outcomes illustrate that the mechanisms of DPL outperform the other models for the following reasons:

- DPL classification training used multiple layers of neural networks, enabling feature extraction and transformation for input into the next hidden layers, and permitting more simple performance of interaction with input data.
- DPL classification training development using LSTM layer for sequence classification created various forecasts for each signal time step of a series data.

Tables 4.10-4.13 presented that BEC model across the four tested datasets achieved better performance in classifying correct audit opinion (after DPL model) compared to DT, K-NN, SVM, ANN, LR, LDA, and NBN, as indicated by average accuracy (above 95.2%), F-measure (93.1-96.4%), specificity (87-93%), sensitivity (99.5-98.8%), AUC (97-99%), and lower Brier scores, and Type I and II Errors. According to these results, the BEC model has good ability to distinguish between audit opinions, and assists to enhance the performance of DT classifier machine learning outcomes through combining the outputs of weaker DT models to produce better predictions. Due to this, BEC model performance indicated better evaluation results than the DT model, but the latter had better performance in correct classification of audit opinion compared to K-NN, SVM, ANN, LR, LDA, and NBN across the four tested datasets, with higher rates in most evaluation parameter results, including accuracy rates, F-measure, sensitivity rates, and AUC rates, with lower Brier scores and Type II Error rates.

The valuation measurement outcomes illustrate that the mechanisms of BEC and DT outperform the other modes as K-NN, SVM, ANN, LR, LDA, and NBN models for the following reasons:

- BEC model enhances machine learning outcomes through combining the output of weaker classifier models to produce better predictions. The training classification ensemble model using *fitcensemble* with AdaBoostM1 algorithm decreases partiality and variance in the ensemble classification model. Additionally, AdaBoostM1 algorithm concentrates on the transformation of weak learners (by integration) into stronger ones.
- DT develops a predictive model or tree structure that does not demand any previous knowledge or hypothesis, as DT provides a meaningful method to obtain knowledge. It simply utilises if-then classification rules, thus model procedures comprise sets of straightforward decisions.

Comparing valuation metrics results tables across the four datasets revealed that the ANN and SVM models outperformed NBN, LR, LDA, and K-NN, with higher average accuracy, F-measure, AUC, specificity, and sensitivity rates, and lower Brier scores and Type I and II Error rates. SVM and ANN models had better ROC (near to corner 1) compared to the NBN, LR, LDA, and K-NN models. The ANN and SVM models' F-measure results (e.g, year 2018 dataset achieved 94.6% and 93.8%, respectively) indicate that they are more accurate and robust to classify audit opinion cases compared to NBN, LR, LDA, and K-NN.

LR, which determines the conditional likelihood of the particular observation relating to a class, presented as a value of input variables, resolves the binary classification issue of identified likelihood in cases where there is a binary output target variable, which consists of only two possible values (no/yes, 0/1, or false/true) for predicting the variable. For the year 2019 and 2018 datasets, LR showed better evaluation measurement performance in terms of predicting correct auditing opinions compared to LDA, NBN and K-NN models, with higher accuracy, AUC, F-measure, sensitivity, and specificity rates, and lower Brier scores, and Type I and Type II Error rates. Additionally, for the year 2017 dataset, LR had better performance than the NBN and K-NN models, with superior results for all nine evaluation parameter results. This means that the LR model is more effective in identifying qualified and unqualified opinion correctly than LDA, NBN, and K-NN. On the other hand, the LR and LDA models had weaker performance than NBN and K-NN for the All-Data dataset, with lower accuracy, AUC, F-measure and specificity rate, and higher Brier score and Type I Error. Consequently, these models have weak performance for larger datasets compared to machine learning.

Compared to the LR, NBN, and K-NN models, the LDA model delivered better performance for classification rates in terms of predicting 99% correct auditing opinions for the year 2017 dataset, with only 1% unqualified opinions incorrectly flagged as qualified. In addition, the LDA model achieved better performance compared to NBN and K-NN for the year 2019 and 2017 datasets, but for the All-Data dataset it had lower F-measure (79.8%) and significant variance between evaluation for correctly flagging qualified and unqualified classes (33.6%) for the All-Data dataset, being outperformed by BEC, NBN, SVM, ANN, LR, DPL, K-NN, and DT. Overall, LDA has the worst ability to distinguish between audit opinion correctly for larger dataset sizes.

Tables 4.10-4.13 display that for the All-Data dataset the K-NN model outperformed the LR, NBN, and LDA models in seven of the nine evaluation parameters (AUC, sensitivity, Type II error, Brier score, accuracy, F-measure, and AURD). Likewise, for the year 2019 and 2018 datasets and All-Data, K-NN had better AURDs compared to the BEC, NBN, SVM, ANN, LR, LDA, and DT models. K-NN had the best reliability diagrams (with bins close to the diagonal line) compared to the reliability diagrams for the BEC, NBN, SVM, ANN, LR, LDA, and DT models. On the other hand, compared to the LDA model, K-NN had a higher gap between the sensitivity and specificity rates for the yearly datasets: 2019 (13.6%), and 2017 (14.9%).

The tables comparing valuation metrics results for all four datasets showed that NBN model had weaker performance in terms of correctly and incorrectly flagging qualified or unqualified class compared to other models. NBN valuation results for the year 2019, 2018, 2017 and All-Data datasets had a high gap in identifying different types of audit opinion, with the highest gap between specificity rate and sensitivity rate (2019 (36.1%), 2018 (15.3%), 2017 (21.9%), and All-Data (15.5%)), and difference between Type I and II Errors, compared to the gaps for the DPL, LR, ANN, K-NN, LR, LDA, SVM, EBC and DT models. NBN classifier performance evaluation revealed higher variation in average accuracy rates (75.8-92.3%) and Brier scores (6.9-22%), with the highest AURD over the four datasets. These results lead to the conclusion that NBN model is a more imbalanced predictive model in terms of detecting the right audit opinion, compared to the other models.

4.4. Statistical Significance Testing

This section explains the Friedman statistical test results applied on all nine individual models to determine the statistical significance of classifier performance over the four datasets, and the Bonferroni-Dunn test to rank all individual models from best to worst performance.

Table 4.14 presents the Friedman test results for all nine classifier performance outcomes, and for the performance of the best five classifiers (DL, DT, EBC, SVM, and ANN), across the four datasets.

Table 4.14: Friedman test results

Datasets	**Year 2019**	**Year 2018**	**Year 2017**	**All-Data**
Friedman X_F^2 (All classifiers)	4,114	4,982	6,494	1,582
Friedman X_F^2 (Best classifiers)	3,720	4,208	3,148	643

Source: Author

In order to make inferences more scientifically robust, Friedman test ranked the best classification models for the four datasets separately, and the classification model with the best rank was selected as the control classification model, with the null hypothesis that there are no statistically significant variations between the performance of these best classifiers' rankings, which are identically generated, and the significance of each model is random; the alternative hypothesis is that one model outperformed others. For all four datasets the null hypothesis is rejected, and the alternative hypothesis is approved, with the critical value (P-value) ≤ significance level denoted as α = 0.05 and α = 0.1. In addition, a null hypothesis is rejected when the Friedman results (X_F^2) ≥ chi-square outcome ($X_{0.05}^2(4)$)= 9.5 and if X_F^2 ≥ $X_{0.1}^2(4)$= 7.8.

Utilised pairwise statistical t-test along with Friedman test evaluates which of the five best individual classifiers attained the best performance. Tables 4.15-4.18 present the results of pairwise comparison statistical tests for the best individual classifiers across the four tested datasets. Pairwise statistical t-test outcomes illustrate if a pair of models has performed in the similar path. The obtained data illustrates low p-value for all the five classifiers across all four datasets, and hence the performance of each model is distinct and proportional to its accuracy.

Based to the pairwise comparison and Friedman results, the tables indicate that the null hypothesis is rejected for four classification models over four datasets, because the Friedman results achieved were higher than 9.5 and 7.8, and at α = 0.05 and α = 0.1 the level of significance is higher than the P-value for the four datasets (around 0). The tables show pairwise t-tests for each duo of models to discover which models did not perform, and which proceeded in a comparable path. The pairwise comparison testing P-value results between the models are low, indicating that the significance of each model is proportionate to its average accuracy. Likewise, all results approved that there was a significant difference between the best classifiers.

Table 4.15: Pairwise comparison results for 2019 dataset

FriedmanX_F^2 = 3,720 P-Value= 0.00004	**Accuracy**	**BEC**	**DT**	**ANN**	**SVM**
DPL	98.1%	0	0	0	0
BEC	96.8%	-	0	-	0
DT	96.3%	-	-	0.38	0
ANN	95.9%	0	-	-	0
SVM	95.4%	-	-	-	-

Source: Author

Table 4.16: Pairwise comparison results for 2018 dataset

FriedmanX_F^2 = 4,208 P-Value= 0.0000009	**Accuracy**	**BEC**	**DT**	**ANN**	**SVM**
DPL	98.6%	0	0	0	0
BEC	97%	-	0	-	0
DT	96.8%	-	-	0	0
ANN	95.6%	0	-	-	0
SVM	95.8%	-	-	-	-

Source: Author

Table 4.17: Pairwise comparison results for 2017 dataset

Friedman X^2 = 3,148 P-Value = 0.00006	**Accuracy**	**BEC**	**DT**	**ANN**	**SVM**
DPL	97.5%	0	0	0	0
BEC	96.6%	-	0	-	1
DT	96%	-	-	0	0
ANN	95.3%	0.279	-	-	0.01
SVM	95.7%	-	-	-	-

Source: Author

Table 4.18: Pairwise comparison results for All-Data dataset

FriedmanX_F^2 = 643 P-Value = 0.0001	**Accuracy**	**BEC**	**DT**	**ANN**	**SVM**
DPL	95.5%	0	0.894	0	0.002
BEC	95.3%	-	0	-	0
DT	95.1%	-	-	0	0.003
ANN	92.3%	0.28	-	-	0
SVM	90.8%	-	-	-	-

Source: Author

After rejecting the null hypothesis, post-hoc Bonferroni–Dunn comparison test was performed to discover any significant variances between individual models. Individual classifier models are significantly different if variation in their mean ranks from the Freidman test is not lower than the critical difference (CD) at the significance level α = 0.05 and α = 0.1. In our case:

- $CD_{0.05} = 5.3$ where $q_{0.05} = 2.724$
- $CD_{0.1} = 4.85$ where $q_{0.1} = 2.498$

Figure 4.20 summarises the mean rank for each individual classification model obtained from Freidman test by bars, and the CD of Bonferroni-Dunn's procedure with α = 0.05 and α = 0.1 indicated by two horizontal lines, which represent cut-off lines that move through all bars. Based on these two lines, it is possible to determine which model has the best performance: the higher classifiers' average rank value is above the two lines, the worse their performance; and the lower their average rank value, the better their performance. The two cut-lines are calculated by the sum of CD at α = 0.05 and α = 0.1, with the lowest rank presenting the model with the best performance. The line at α = 0.05 is equal to 6.7 (5.3+1.4), and the line at α = 0.1 is 6.25 (4.85+1.4).

Figure 4.20 shows that the average rank values of DPL, DT, BEC, SVM, and ANN are below the two cut-lines, which means they have better performance to classify audit opinion correctly compared to LR, LDA, K-NN, and NBN, whose average rank values are above the cut-lines. DPL model had the lowest rank at 1.4, and it was the furthest below the two cut-lines at α = 0.05 and α = 0.1. This statistically proves that the DPL model has superior ability to classify audit opinion correctly compared to the other individual models. On the other hand, K-NN, LDA, and NBN had average ranks of 7.5, 7.09, and 8.68 (respectively), comprising poor performance (above the two cut-lines), but LR was the worst in relation to the cut-line at α = 0.05. Statistical testing results affirmed that DPL outperformed the other models, and K-NN, LDA, and NBN did not have acceptable performance to classify audit opinion rightly.

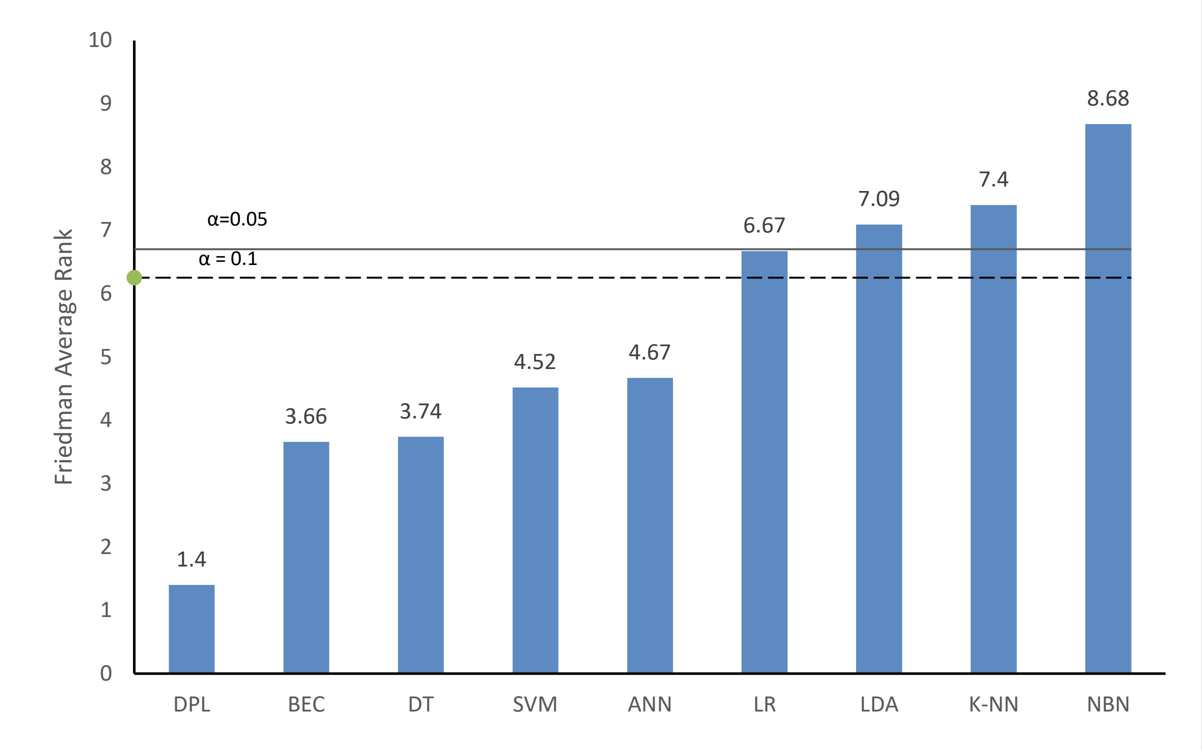


Figure 4.20: Bonferroni-Dunn correction for Individual classifier models, with significance levels

Source: Author

4.5. Summary

This chapter explored the details of the simulation and computing platforms of single classifiers and the capability of DM based on DPL, ANN, BEC, SVM, K-NN, NBN, DT, LDA, and LR models as classification tools for auditing opinion. The evaluation testing outcomes for the four datasets illustrate that the nine classification models have acceptable performance to correctly classify audit opinion, as indicated by achieving good ROC and reliability diagrams, with good average accuracy rates ranging from 75.8-97.8%, F-measure 76.8-97.7%, and an AUC of above 84.9%. Additionally, all models achieved good Type II Error classification rates, Brier scores, and AURD (below 9.5%, 22.1%, and 38.5%, respectively). On the other hand, analysis of the comparative evaluation results for all models showed that all model performances had declining performance with the All-Data dataset (BD size), with reduced evaluation results such as average accuracy and AUC rate, and increased Type I and II Error rates.

The empirical test results indicate that from over four datasets, DPL model revealed superior ability in classifying audit opinion correctly, outperforming NBN, LR, ANN, K-NN, LR, LDA, BEC, and DT models, achieving higher results for all nine evaluation parameters. It is also a balanced model, able to distinguish between audit opinions, with fewer differences between Type I and II Error rates compared to other models. In addition, DPL had the best ability for any size of dataset, especially with the BD dataset (All-Data), while other models' performance abilities decreased at this size. The BEC model had the next best performance (after DPL) in terms of lower incorrect classification rate, with the highest rates for AUC, accuracy, specificity, F-measurement, and sensitivity in classifying audit opinion, compared to the NBN, LR, ANN, K-NN, LDA, and DT models' evaluation test results over all four datasets. In addition, K-NN model indicated lower AURD in all four datasets compared to NBN, BEC, LR, LDA, and DT.

On the other hand, the evaluation metrics results over all four datasets indicated that NBN model was imbalanced regarding its detection of the right audit opinion, and it had the poorest ability to distinguish between audit opinion correctly, indicated by a large gap between the Type I and II Error rates in the evaluation results for all datasets compared to the gap between Type I and II Error rates for the other models. Additionally, NBN model achieved the highest Brier scores and AURDs at all four datasets. Likewise, all comparative evaluation test results analysis over the four datasets indicated that LDA, K-NN, and NBN had worse performance to classify audit opinion correctly compared to DPL, BEC, LR, DT, and ANN models.

Statistical significance testing approved that DPL model has superior ability to classify audit opinion correctly compared to the other individual classifiers, and NBN has the worst performance. Likewise, statistical significance testing supports all results reached from comparing evaluation test results, particularly the finding that DPL outperforms other models.

Chapter 5
Committee Machine Classifiers Models Combiner

5.1. Introduction

Chapter 4 presented the test ability of nine individual classifiers as classification tools in determining audit opinion by evaluating their performance across the four tested datasets. It concluded that some of the classifiers (DPL, BEC, and DT) have good ability to indicate correct audit opinion, but it remains to determine whether utilising nine classifier predictions together in combination can achieve more certainty in audit opinion classification tools. Consequently, this chapter tests CM, combining nine individual classifiers in a compound committee model, to evaluate outputs and determine how these individual classifiers work together. There are several committee combination methods in relation to how the committee functions, which play an instrumental role in model outputs and generalisability. The committee methods used in this chapter are CON, FC, AVG, WAVG, MED, MajVot, MIN, and MAX methods. The experiments of this study are conducted using MATLAB 2019a version on an 8 GB RAM personal computer with 3.4 GHz, Intel CORE i7, and Microsoft Windows 10 operating system.

The following section illustrates the performance of each of the eight committee combination methods, followed by experimental test results for each. Comparative analysis and discussion of the committee methods' abilities to indicate the correct audit opinion is accompanied by comparison with each individual classifiers' performance. The final section presents the statistical significance testing to approve the capability of committee model performance.

5.2. Combination Method Development

This section discusses the improvement of six traditional committee combiner methods and two new committee modelling approaches – CON and FC methods –used to combine all predictions from each single classifier used in Chapter 4 in order to enhance prediction accuracy, based on analysis of statistical functions and illustrating their weaknesses and strengths. This section presents statistical functions for each traditional committee combiner modelling via diagrams and consideration of data type suitability for committee method utilisation.

5.2.1. Consensus Combiner Model

The consensus combiner model (CON) works as a decision maker to consider individual classifiers as a cooperative group of agents that transfer their decision of input admittance to a combiner decision maker to adjust judgments about inputs, harmonising opinions in order to raise the consensus level of a set. The decision maker then reaches consensus on the estimation of a best decision classification, through integrating individual classifiers' predictions, achieving more efficient decision making. The developed consensus combiner method is explained with its procedure and dataset-specific enhancement in the following subsections.

5.2.1.1. Measuring Classifier Rankings and Constructing Decision Profiles

This phase involves constructing the decision profile for each individual classifier in ensemble E = [e_1,⋯,e_H], then this ensemble (E) is trained on the same input data points to express its own predictable decision. As after training ensemble, the answer $\Gamma = [b_1, \cdots, b_n]$ is selected from the group of potential answers. Considering the assessment function RL_l for each single classifier, the function is associated with the positive number for every probable response Γ_l. After that, an outcome of an assessment function RL_l value in a range of 0 and 1 illustrates the desirability of a corresponding output (equation 5.1).

$$\sum_{K=1}^{n} RL_l(b_k) = 1;\ where\ \forall l\ \in [1\ \cdots F] \qquad 5.1$$

After measuring the ranking of each individual model, uncertainty must be evaluated between each single classifier by the operation DP matrix 5.2 for testing the group (input by input), arranged in j columns and l rows. As seen from the DP matrix (equation 5.2), e_j is the j^{-th} input ensemble classifier and $\mathrm{RL_l}(\mathrm{e_j})$; and l $\in [1\ \cdots 9]$ is the l^{-th} classifier ranking level for the j^{-th} input. The final decision may be represented as evaluating a common set ranking level $\mathrm{RL_s}\colon \Gamma \to [0,1]$ to aggregate the predictable ranking levels for all individual models.

$$DP = \begin{bmatrix} RL_1(e_1) & RL_1(e_2) & RL_1(e_3) & RL_1(e_4) & \cdots & RL_1(e_j) \\ RL_2(e_1) & RL_2(e_2) & RL_2(e_3) & RL_2(e_4) & \cdots & RL_2(e_j) \\ RL_3(e_1) & RL_3(e_2) & RL_3(e_3) & RL_3(e_4) & \cdots & RL_3(e_j) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ RL_9(e_1) & RL_9(e_2) & RL_9(e_3) & RL_9(e_4) & \cdots & RL_9(e_j) \end{bmatrix} \qquad 5.2$$

5.2.1.2. Measuring Uncertainty Assessment Matrix of Classifiers

This phase finds a function to measure the uncertainty of each individual classifier. A function specifies less weight to classifiers with higher uncertainty performance, and vice-versa; weight values must reflect inequality in classifiers' performance judgements. During this phase, the uncertainty measurement is divided into self-uncertainty (also called local uncertainty), and global uncertainty (also called conditional uncertainty), as discussed below.

Local uncertainty

Local uncertainty illustrates the fineness of a classifier's classification related to its own judgment base and uncertainty about its own judgment. Local uncertainty of a classifier e_j is measured by equation 5.3. The n in the equation is equal to the number of classes.

$$U_{ll} = -\sum_{k=1}^{n} RL_l(b_k)\, log_n(RL_l(b_k)) \qquad 5.3$$

Where U_{ll} represents local uncertainty for the l^{-th} classifier, n is the number of the class [0,1], and $RL_l(b_k)$ is the l^{-th} agent ranking level of answer of b_k.

Conditional uncertainty

Global or conditional uncertainty relates to the degree of classifiers' certainty in their own judgments after observing other classifiers' decisions. This reflects the relative value of information concluded in relation to other classifiers' decisions. A new collaborated decision profile exchange is generated in which each classifier is able to reveal its uncertainty level, in an attempt to produce more certain outcome about a firm's status of all individual classifiers together. In this phase, the classifier has the capability to rehearse its uncertainty level and adjustment is based on other classifiers' decisions and its own, enhancing judgment when the decisions of others classifiers are available. The ranking level (equation 5.4) is used to measure the conditional uncertainty of classifiers using equation 5.5.

$$\sum_{k=1}^{n} RL_l\left(b_k \middle| \Gamma_j\right) = 1; where\ \forall_l \in [1, \cdots, F] \qquad 5.4$$

$$U_{lj} = -\sum_{k=1}^{n} RL_l\left(b_k \middle| \Gamma_j\right) log_n(RL_l\left(b_k \middle| \Gamma_j\right) \qquad 5.5$$

Where U_{lj} reflects the conditional uncertainty of the lth classifier's classification when it realises a j^{-the} classifier's ranking level. To estimate classifiers' uncertainty, consider $RL_l\left(b_k \middle| \Gamma_j\right)$is l^{-th} agent ranking level of response b_k when it informs a decision weight of vector of j$_{-th}$ agents.

Firstly, equations 5.1 and 5.4 are verified, and then equations 5.3 and 5.5 will be used. The matrix of the classifier's uncertainty stated in equation 5.6 is displayed as a matrix (U) that is evaluated using equations 5.3 and 5.5:

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} & U_{14} & \cdots & U_{1j} \\ U_{21} & U_{22} & U_{23} & U_{24} & \cdots & U_{2j} \\ U_{31} & U_{32} & U_{33} & U_{34} & \cdots & U_{3j} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{91} & U_{92} & U_{93} & U_{94} & \cdots & U_{9j} \end{bmatrix} \qquad 5.6$$

5.2.1.3. Measuring Classifier Weights

After calculating the uncertainty matrix, all classifiers have the potential to specify weights for themselves and other classifiers in an ensemble. The weights of uncertainties are assessed utilising equation 5.7, with uncertainties weight is displayed in matrix W, consisting of the column (j) and row (l), as shown in matrix 5.8. The weights are measured through equation 5.7.

$$W_{lj} = 1 \div \left(U_{lj}^{2} * \left(\sum_{j\in F} U_{lj}^{-2}\right)\right) \qquad 5.7$$

Where $\mathrm{W_{lj}}$ is the wieghted l$^{\text{th}}$ classifier, when it knows a ranking level of the j$^{\text{th}}$ classifier, $\mathrm{U_{lj}^{-2}}$ is uncertainty for the l$^{\text{-th}}$ classifier, when it knows an uncertainty of the j$^{\text{-the}}$ classifier, and j is the number of the classifier (F).

$$\mathrm{W} = \begin{bmatrix} \mathrm{w}_{11} & \mathrm{w}_{12} & \mathrm{w}_{13} & \mathrm{w}_{14} & \cdots & \mathrm{w}_{1\mathrm{j}} \\ \mathrm{w}_{21} & \mathrm{w}_{22} & \mathrm{w}_{23} & \mathrm{w}_{24} & \cdots & \mathrm{w}_{2\mathrm{j}} \\ \mathrm{w}_{31} & \mathrm{w}_{32} & \mathrm{w}_{33} & \mathrm{w}_{34} & \cdots & \mathrm{w}_{3\mathrm{j}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{w}_{91} & \mathrm{w}_{92} & \mathrm{w}_{93} & \mathrm{w}_{94} & \cdots & \mathrm{w}_{9\mathrm{j}} \end{bmatrix} \qquad 5.8$$

An outcome of the weight matrix is the N×N stochastic matrix, a one-step transition likelihood matrix of Markovian chain with individual classifiers (DeGroot, 1974). It is possible to employ a limited theorem of Markovian chains to convert the ensemble into a distinctive ensemble consensus common ranking (Berger, 1981). DeGroot (1974) demonstrated that such ensembles can be converted to common ranking only when the situation of the vector π is as in equation 5.9, which enables a common set consensus ranking using equation 5.10.

$$\left(\begin{array}{l} \pi * W = \pi \\ \sum_{j=1} \pi_j = 1 \end{array}\right. \qquad 5.9$$

$$LR_s(b_k) = \pi * RL_j(b_k) \qquad 5.10$$

5.2.1.4. Update Calculations and Final Decision

DeGroot (1974) explored outcomes when e_i wishes to alter weights specified to other individual models after acknowledging their initial judgments, particularly to determine how far judgments vary from the consensus decision. In this case, each $e_j \in E$ updates a weight identified for other classifiers after learning their primary decision. Utilising this update, e_j has the ability to go over all ranking levels assigned to classifiers. These new ranking levels subsequently generate a new uncertainty matrix and weight matrix calculation. This update is done with 30 loops in this case, which can repeatedly execute a previous phase. In this phase an aggregate final result is assessed through utilising the following steps. However, the final loop control, determines the final update, and leaves no room for further decisions, after which the final consensus prediction is made. The following steps are undertaken during the updating process.

Let the initial consensus decisions be presented in the vector Γ = $[b_1, \cdots, b_n]$ and the classifier decisions be presented by C_j = $[C_1, \cdots, C_n]$, then for each e_j the final aggregate prediction is measured for the consensus combiner, when there is no room for update decision, as in equation 5.11.

$$FC_j = \frac{\sum C_j * LR_s(b_k)}{\sum LR_s(b_k)} \tag{5.11}$$

In the second step, for each FC_j value, the error performance value must be measured using equation 5.12.

$$ER_j = 1 - \left|\sum(FC_j - (c_j * LR_s(b_k)))\right| \tag{5.12}$$

After finding the error, the new weights modification factor is calculated for each classifier, using equations 5.13 and 5.14.

$$Z_j = \left(0.1 \times ER_j\right) \div \left|\sum ER_j\right| \tag{5.13}$$

$$w_j = LR_s(b_k) + Z_j \tag{5.14}$$

where W_j is the final weights modification factor to each classifier, and Z_j is the sigma of each classifier.

Finally, each ej has the ability to give new rankings to fellow individual models, mirroring an update that has been received.

Algorithm 5.1 displays the pseudo code epitomising the adopted classifier consensus operation (CON).

Algorithm 5.1: CON pseudo code based on committee method (CON)

```
Input:
Predictions values for each single classifier (F)
$l = 1 \dots 9$
$RL_l = 1$

Output:
    a) For $l = 1$ to F do
    b) For $j = 1$ to F do
    c) For each $e_j \in E$ do
    d) if $(l == j)$ $DP_{ll} \leftarrow$ classify $(RL_l, e_j)$
       Else
       $DP_{lj} \leftarrow$ classify $(RL_l, e_j)$
       End if
       End for
       End for
       End for
    e) if$(l == j)$then $U_{ll}$ = Measured through equation 5.3
       Else
       $U_{lj}$ = Measured throuh equation 5.5
       End if
    f) $W_{matrix}$ = Measured through equation 5.8
    g) Calculate the aggregate consensus ranking LRs by utilising equation 5.10
    h) Specify ensemble final aggregate reply utilising equation 5. 11
    i) Calculate the error by equation 5.12
    j) Calculate the weights modification factor by equation 5.14
```

Source: Author

5.2.2. Fuzzy Logic Combiner Method

Fuzzy logic maps input space to output space, and FC model controls the range point dimension by fuzzy membership functions. Outputs are defuzzified to convert fuzzy inference set outputs into crisp outcomes. A new FC method uses the reliability of the predictions of each classifier in combination, using a rule base to find a more accurate answer. The combined set is defuzzified to produce the final outcome. The development of the new FC process is explained below.

5.2.2.1. Building Fuzzy Logic Combiner Method

The new FC method is intended to deal with uncertainty. First, a reliability diagram is generated for each single classifier's prediction. For any particular predicted datapoint, the prediction is fed to the reliability plot, to obtain the confidence level (how far the graphs are from the diagonal line; the closer they are, the higher the confidence), and the confidence standard deviation. Classifier predictions (CPs) are divided into 20 bins, into which predictions fall based on optimal mean (OP), ranging from 0 to 1. For example, the classifier predicted values between 0 and 0.05 drop into the first bin, and those between 0.05 and 0.1 fall into bin no. 2, etc. Mean and standard deviation values are calculated based on prediction value and the actual target values in each bin. After that, the confidence level (CL) and confidence standard deviation (CSD) are calculated using equations 5.15 and 5.16.

$$CL_k = 1 - \frac{\sqrt{(CPM_k - OP)^2 + (CTM_k - OP)^2}}{B} \qquad 5.15$$

$$\sigma_k = \frac{\sqrt{\sigma_{CPk}{}^2 + \sigma_{CTk}{}^2}}{B} \qquad 5.16$$

where K is the number of queries acquired for the classifier, *CPM$_k$* refers to classification prediction mean for *kth* classifier input data, *CTM$_k$* is the classification target mean for kth classifier input data, and B is the classifier input data size.

5.2.2.2. Fuzzy Logic Membership Functions and Rules Processing Inference

After getting the confidence standard deviation from all the classifiers, they are changed into a fuzzy set, combined using a rule base. The fuzzy set (H) is defined as unclear, and it may include elements with only partial membership degree. The membership function (μ) specifies how each point in a universe of discourse (*Y*) is mapped to the membership value range from 0 to 1, whereby a number falling between 0 and 1 belongs to a fuzzy set only partly, due to which the fuzzy set can represent infinite numbers of membership functions. Therefore, if an element y is a member of fuzzy set H, this mapping can be represented by $H = \{y, \mu_H(y) | y \in Y\}$. Membership functions are denoted by μ_H and $\mu_H(y)$ discover where the membership degree of an element y in H; when universe discourse *Y* and outputs mapped each element of *Y* to the membership numbers range from 0 to 1.

The fuzzy logic toolbox of MATLAB has 11 built-in membership functions based on Gaussian distribution function, Sigmoid curve, quadratic and cubic polynomial curves, and piece-wise linear functions. In this case, the Gaussian distribution function is used to measure all classifiers' membership function values, as shown in equation 5.17.

$$G_{H_k} = EXP\left(\frac{-(Y - PC_k).^2}{2\sigma_k^2}\right) \qquad 5.17$$

where G_{H_k} is the output for each classifier membership by gaussian function, Y is the universe of discourse, PC_k is the classifier prediction for kth, and σ_k^2 is the confidence standard deviation for *kth*.

After getting the Gaussian distribution for each single classifier's membership value, weights rule is used for the modification of fuzzy sets, while the position of each fuzzy set is unchanged. This modification is applied through multiplying each fuzzy set output by the weight rule (D*), as illustrated in equation 5.18. The weight rule measures the reliability of each classifier, and in this case, the CL is the weighted rule, which can be simply adjusted to modify fuzzy set performance (the position of each fuzzy set) without modifying each classifier's membership function. Figure 5.1 illustrates as example for two classifier fuzzy sets:

$$G_{H_k} = EXP\left(\frac{-(Y - PC_k).^2}{2\sigma_k^2}\right) \times \text{D*} \qquad 5.18$$

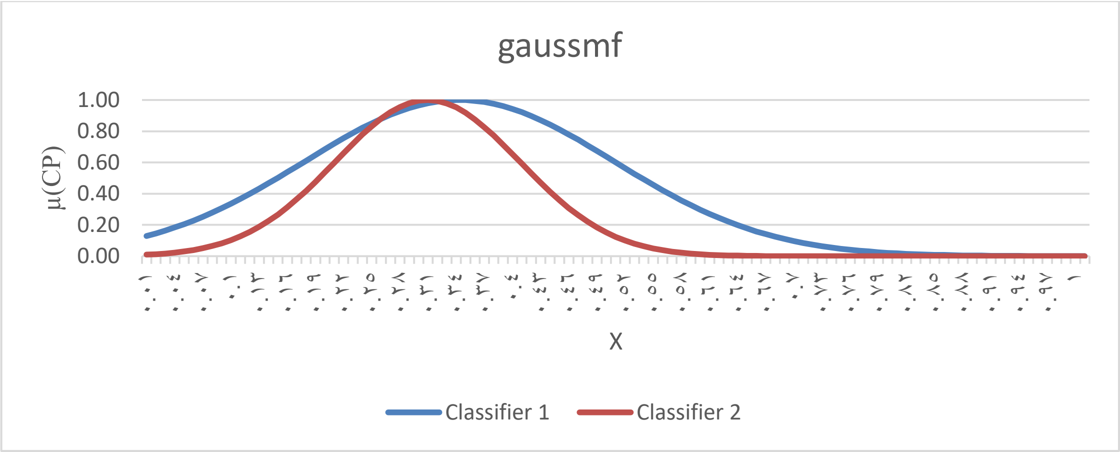


Figure 5.1: Two fuzzy sets with different widths and centre positions

Source: Author

5.2.2.3. Fuzzy Logic Combiner Operation

After getting each classifier fuzzy set, all fuzzy sets are combined together into one fuzzy set, integrating the operations of each classifier fuzzy set by combining their parallel threads. The average method is used to combine all fuzzy sets together by taking and combining all the fuzzy sets of each individual classifier in columns, and returning a column vector containing the mean of all the fuzzy sets of classifiers in each row. Equation 5.19 illustrates the final output of the process of integrating all fuzzy sets.

$$CFS = \frac{\sum_{k=1}^{n} G_{H_k}}{n} \qquad 5.19$$

where *CFS* is the combined fuzzy set, *n* is the number of the classifier fuzzy set, and G_{H_k} refers to the output for each classifier fuzzy set.

5.2.2.4. Defuzzification

Once functions are inferred and combined, the final fuzzy set output needs to be defuzzied into crisp outcome data. The final step requires a custom defuzzification method for defuzzifying the output fuzzy set. Defuzzification is a procedure of combining the successful outputs of the fuzzy set process via an inference mechanism, converting fuzzy inference outputs into crisp outcomes. In other words, the internal representation of the output data in the fuzzy system (usually a fuzzy set) needs to be defuzzified through the decision-making algorithm that specifies a crisp number based on the outputs of the fuzzy set. This final output includes the final output prediction values classifying each firm as qualified or unqualified. There are various defuzzification methods, including middle of maximum (MoM), smallest of maximum (SoM), largest of maximum (LoM), bisector of area (bisector), centre of gravity (CoG), and centroid of area (centroid).

In this case, the popular centroid method was used, which gives better results compared to the other methods in this case. Centroid defuzzification creates crisp values by returning on the centre of gravity of a fuzzy set. The overall area of a μ(y_k) (membership function distribution for the point y_k) is utilised to symbolise a combined control action, split into a number of subareas. A centre of gravity and an area of each subarea is computed, after which a summation of all these subareas is taken to reach defuzzied values for the detached fuzzy set, as shown in equation 5.20. The centroid of the fuzzy set is then computed by $value$ = *defuzz (Y, CFS, 'centroid').*

$$xcenteroide = \frac{\sum_{k=1}^{n} \mu_H(y_k) y_k}{\sum_{k=1}^{n} \mu_H(y_k)} \qquad 5.20$$

where n refers to number of points in the dataset, $\mu(y_k)$ represents a membership function of an aggregated fuzzy set H with respect y_k, y_k is the dataset element, k^{th} is sub-area, and *xcenteroide* is the defuzzified values.

The pseudo code shown in Algorithm 5.2 is used to optimise the classifiers FC operation.

Algorithm 5.2: FC pseudo code based on committee method

```
Input:
    Y = 0:0.01:1
    optimal mean = [0.025:0.05:1]
    $PC_k$
    $TC_k$
    $CPM_k$
    $CTM_k$
    $\sigma_{CP_k}$
    $\sigma_{CT_k}$
Output:
    $CL_k$
    $\sigma_k$
    bin_number = max(ceil([$PC_k$*20), 1)
    $D_k$*= $CL_k$ (bin_number)
    width=$\sigma_k$ (bin_number);
    $G_{H_k}$=gaussmf(Y, [width $PC_k$])*$D_k$*
    CFS=mean([$G_{H_1}, \cdots, G_{H_k}$])
    value=defuzz(Y,CFS,'centroid')
```

Source: Author

5.2.3. Average Method

The average method AVG is designed through picking a mean or average value ranking level of all nine classifiers to be selected as the final decision. Figure 5.2 presents the mechanism of the AVG, whereby average combining takes all the predictions of each individual classifier in a column, and returns a column vector containing the mean of all prediction classifiers in each row. AVG does not require changing mean values with a threshold one (0.5), as if all individual models' value distributions are balanced, the final outcome of average model is commensurately balanced.

One of the drawbacks of the AVG is that if there is a large gap between single classifiers' performance, the method will not achieve better performance than the best individual classifier. However, average method has many advantages, including that it often produces better performance results, because the multiple errors of single classifiers are averaged, and AVG design tends to give weaker network fewer rankings. Likewise, the AVG gives better performance than MIN and MAX methods due to the output of the AVG being evenly based on all classifier outputs. Furthermore, when a dataset has equal numbers of 1 and 0, the AVG has good performance to balance between true positive and true negative.

AVG identifies if the ranking changes of classifiers are nonlinear, relying on the confidence of the classifier in the output. For instance, single models' awards rankings between 0.3 to 0 show almost the same level of confidence in outcome results, whilst from 0.4 and above confidence dramatically decreases. If the average model calculates all rankings as equal, misclassification can arise during the final output.

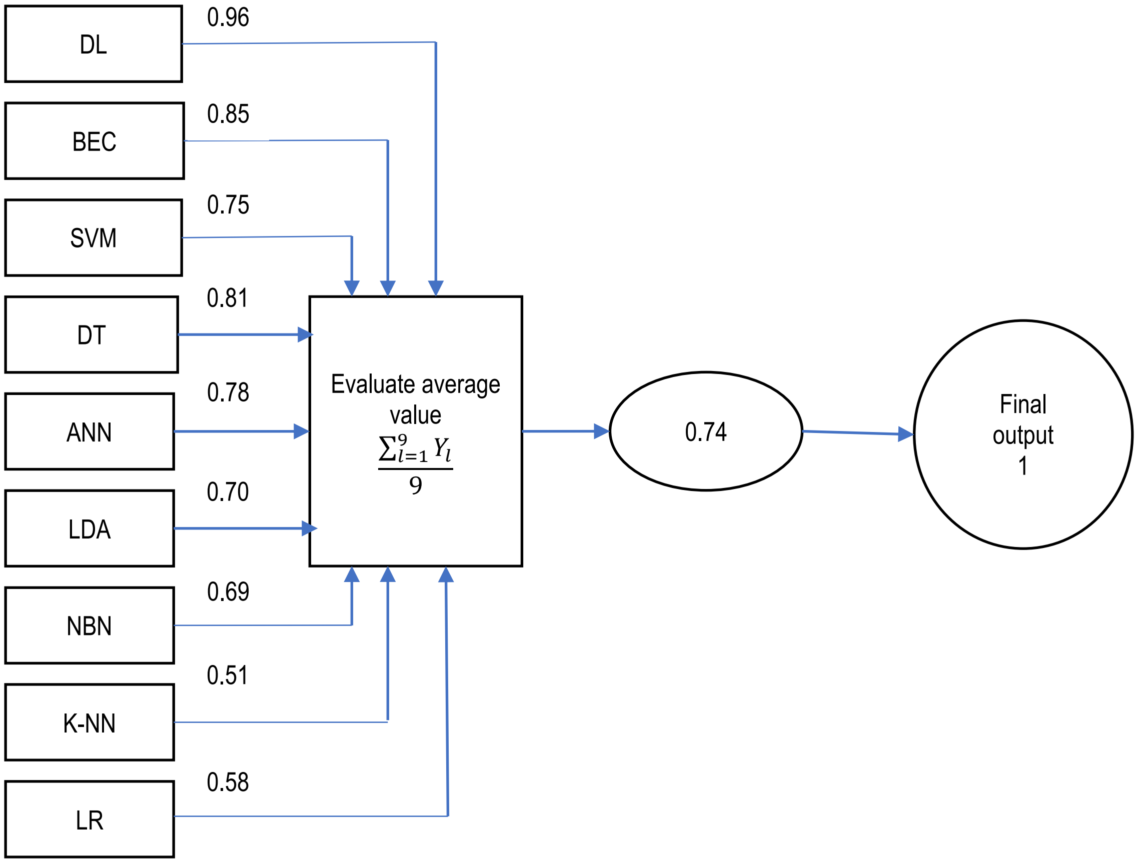


Figure 5.2: AVG method mechanism

Source: Author

5.2.4. Weighted Average Method

The weighted average method (WAVG) takes the mean value of all classifier predictions with the weights associated with the significance of the performance of each classifier to be select as the final output. Figure 5.3 shows the mechanism of the WAVG process to classify the correct class. The weighted average measures the mean for all prediction values of the nine classifiers with weights associated with each single classifier's accuracy. Weighting coefficients are evaluated based on individual classifier's performance global accuracy over the training set. The highest weight coefficient assigned to the classifier has the best accurate performance is on training set and lowers weight coefficient assigned to the classifier has the lower accurate performance. Because of this, WAVG can achieve better performance than most the individual classifiers' performances combined.

The most important advantages to this mechanism include the potential to make decisions by giving higher weight to single classifiers with better performance, and lower weights to single classifiers with less accurate performance. This incorporates more impact of accurate single models' judgments, while minimising the contribution of less accurate judgments to the final output. On the other hand, the weighted average has disadvantages, including that some individual models tend to be over-trained with training datasets, and it can give superior performance outcomes over the training dataset compared to over the testing dataset. These classifier weights will have more impact on WAVG performance. In order to solve this type of issue, training can be increased until the training set performance of all individual models is of suitable and equitable accuracy in relation to the testing dataset.

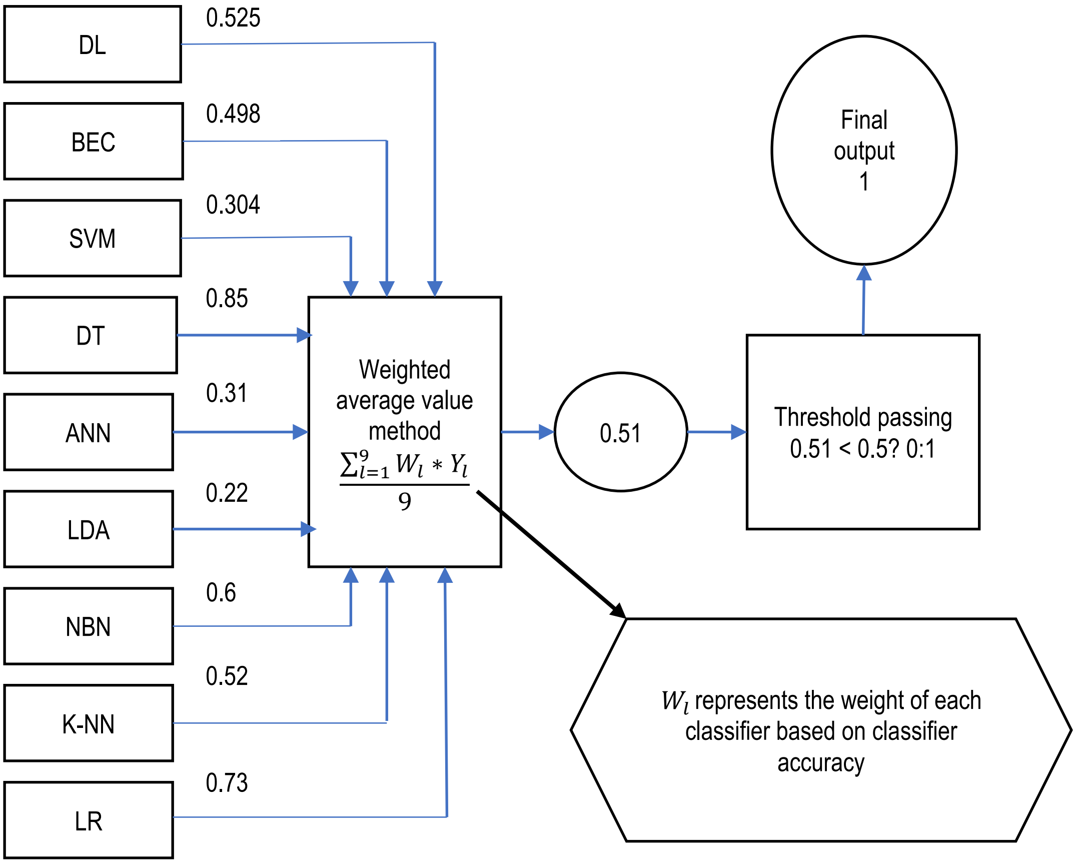

DL
0.525
BEC
0.498
SVM
0.304
DT
0.85
ANN
0.31
LDA
0.22
NBN
0.6
K-NN
0.52
LR
0.73
Weighted average value method
$\frac{\sum_{i=1}^{9} W_i * Y_i}{9}$
0.51
Threshold passing 0.51 < 0.5? 0:1
Final output 1
W_i represents the weight of each classifier based on classifier accuracy

Figure 5.3: WAVG method mechanism

Source: Author

5.2.5. Median Method

Median method (MED) coordinates values in descending or ascending order, and then picks middling values. Figure 5.4 shows the mechanism of the MED process to predict the correct class. The median classifies a value by putting all nine individual classifiers prediction values (qualified and unqualified) in the column, then the MED produces the matrix and computes a value in each row, holding a middle position in each row, due to which MED has no need of a threshold.

One of the disadvantages of MED that the accuracy performance of the nine combined classifiers can be undermined due to selecting the median value. For example, when MED coordinates values are arranged in descending or ascending order, and then picks the middle value for a classifier that has bad performance, the MED performance will be affected. On the other hand, the MED's holding of the middle value reduces negative impacts from extreme values.

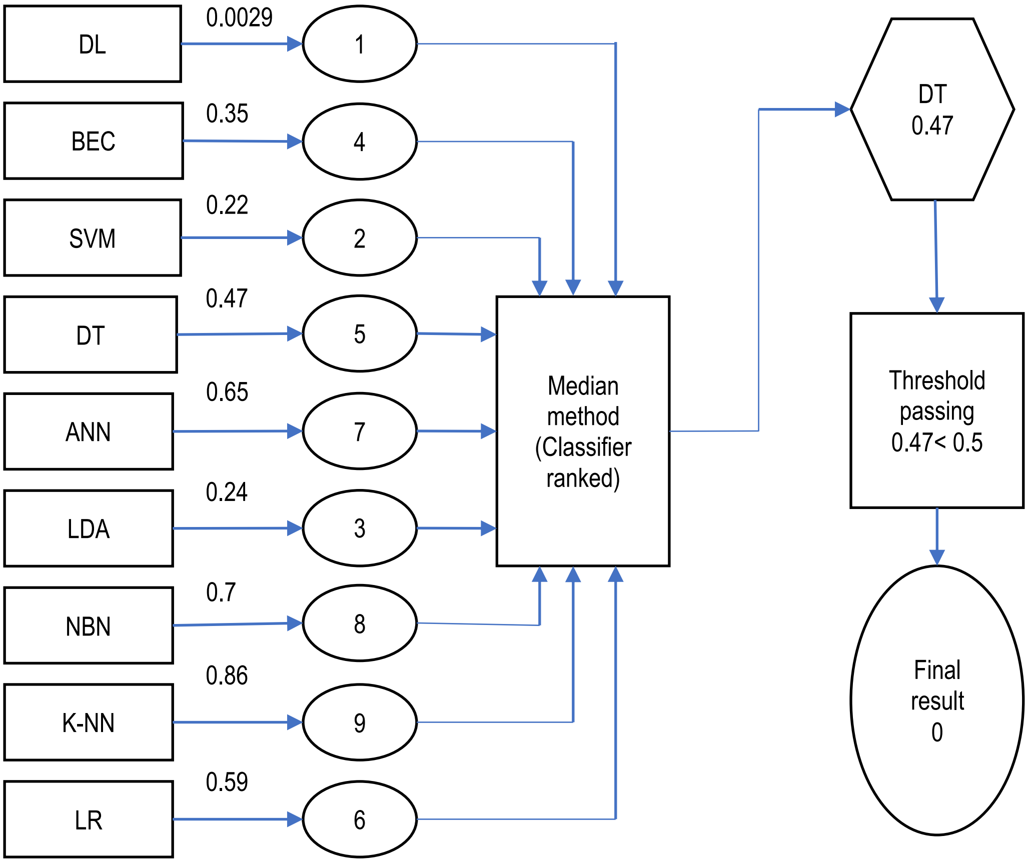


Figure 5.4: MED method mechanism

Source: Author

5.2.6. Majority Voting Method (MajVot)

The majority voting method (MajVot) takes final decisions based on a label or class on which a majority of classifiers correspond. Figure 5.5 presents the mechanism of the MajVot process to classify the correct class. Majority voting puts all nine predictions of the individual classifiers in the column, then the mode method produces the matrix and computes the value in each row, to holding the most frequent class in each row (the highest overall vote is selected through ensemble classifier), in order to make the final decision based on the class to which a majority of the classifiers correspond. Consequently, MajVot does not need to define a threshold.

One of the disadvantages of MajVot method is that its accuracy can be undermined when there are equally frequent numbers of voting, whereby the majority voting picks up the minimum value of the multiple values, but in our case, there are nine classifiers, which means this is not possible, as the mean cannot get an equal number of votes. In order to resolve this problem, an intricate prediction classifier (input data in majority voting) renders complex input into simpler values. MajVot performance is not affected if one of the individual classifiers has weak performance, in contrast to MAX, MIN, and AVG methods, because majority voting holds the frequency value. Accordingly, when a classifier's outcome is independent, MajVot will permanently improve the total performance of the classifier.

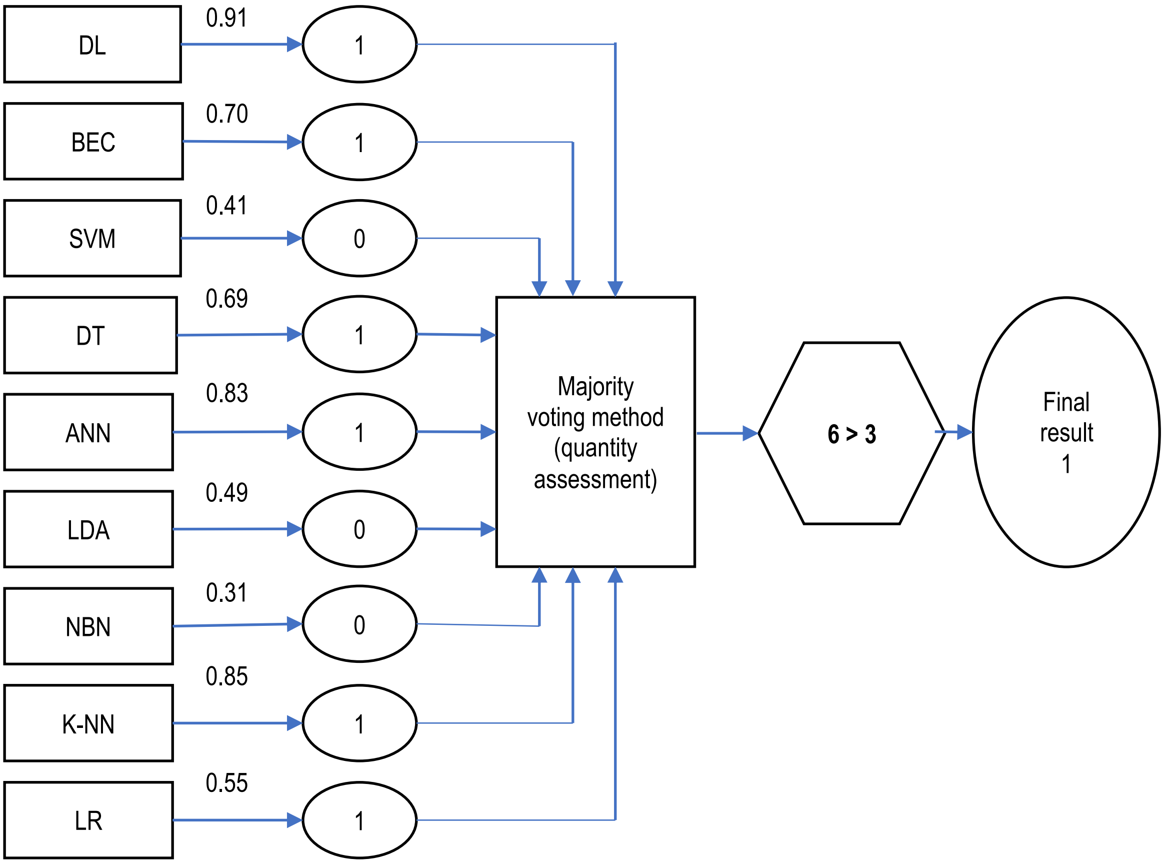


Figure 5.5: MajVot method mechanism

Source: Author

5.2.7. MIN Method

MIN method design holds the smallest value of all classifiers to be picked for final ranking. All nine the individual classifier prediction values (qualified and unqualified) in the column are assigned a column vector holding a minimal value in each row. MIN prediction is the minimal value from all nine individual classifiers' predictions. MIN predicts correct classification for datasets with a higher proportion of the qualified class (actual target value as 0 class), and when all the single classifiers have better performance in predicting qualified companies than unqualified ones, due to the prediction values near to 0, which is the actual target of qualified companies. Conversely, MIN has weak performance when all or most of the individual classifiers have good ability to predict the actual target 1.

Figure 5.6 presents the mechanism of MIN, with improved performance through reducing the threshold lower than 0.5 (the regular threshold) using trial and error; if the threshold was kept as 0.5, the MIN would predict most of the values as 0, but if it was reduced, the performance of MIN to predict unqualified companies would increase. The optimal threshold is chosen by trying different values below the regular threshold. This optimal threshold helps the MIN model to reach best performance to classify data point correctly.

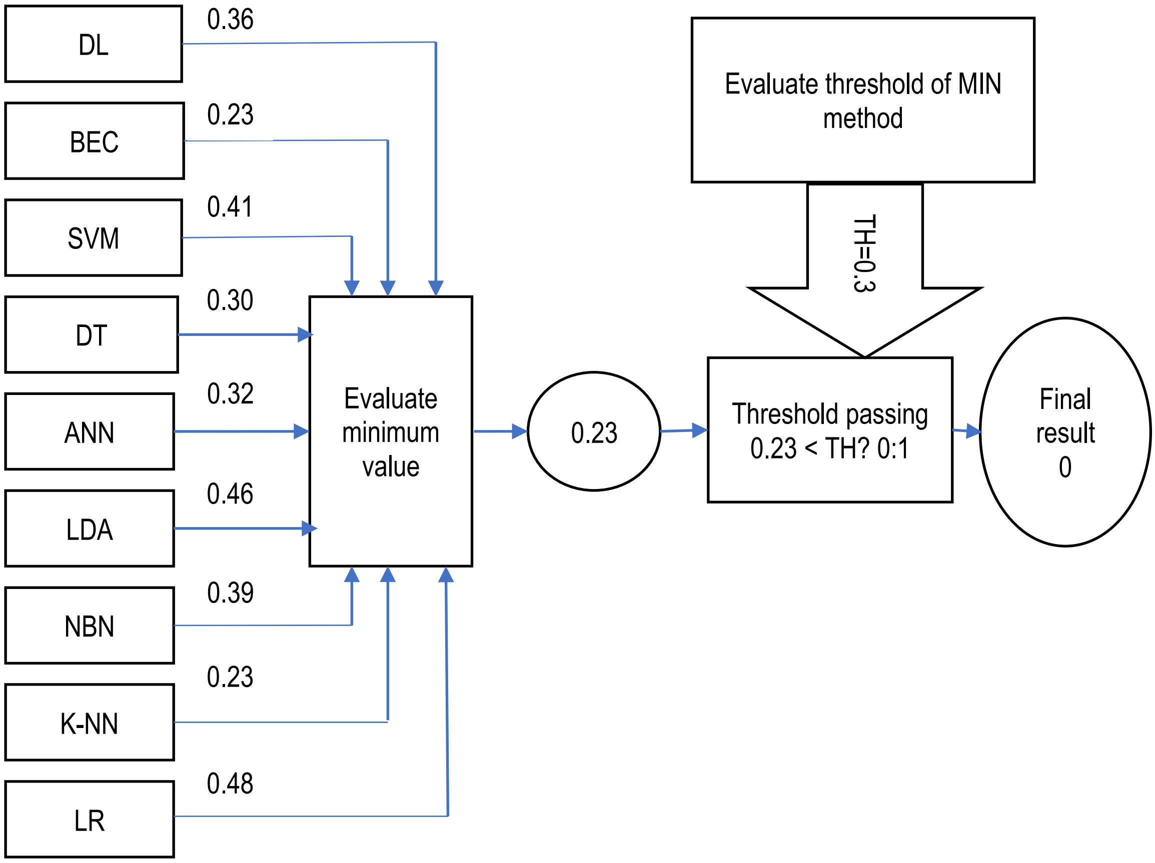


Figure 5.6: MIN method mechanism

Source: Author

5.2.8. MAX Method

MAX method works the opposite way to the MIN method mechanism. All nine predictions (qualified and unqualified) of individual classifiers in the column are used by MAX to produce a matrix and compute a value in each row, in order to pick the highest value in each row. MAX has good classifier performance in classifying unqualified firms with actual targets of 1 than classifying qualified firms (0 class). Because of this, MAX illustrates better performance results when the dataset has proportionally more healthy companies than qualified ones, and if all the single classifiers have better performance to predict unqualified companies (with a target of 1) than qualified ones (with a target of 0).

Figure 5.7 shows the mechanism of the MAX process to predict the correct class. MAX predicts the maximum ranking level for each individual classifier with an enhanced threshold of more than 0.5, using trial and error. This improves the accuracy of MAX performance to detect correct audit opinion prediction, as the regular threshold (0.5) would result in MAX predicting most values as positive (unqualified companies). The optimal threshold is attained by trying different values over the 0.5 threshold. This optimal threshold assists the MAX model to reach best performance to classify data point correctly.

Both MAX and MIN methods have significant disadvantages, which can have major negative impacts on accuracy. For example, if after training one of the single classifiers has achieved weak performance in classifying the correct class, the MAX and MIN methods, when computing values for all classifiers in each row, would pick up the wrong minimum or maximum value in the row, increasing the likelihood of misclassification of the correct prediction value.

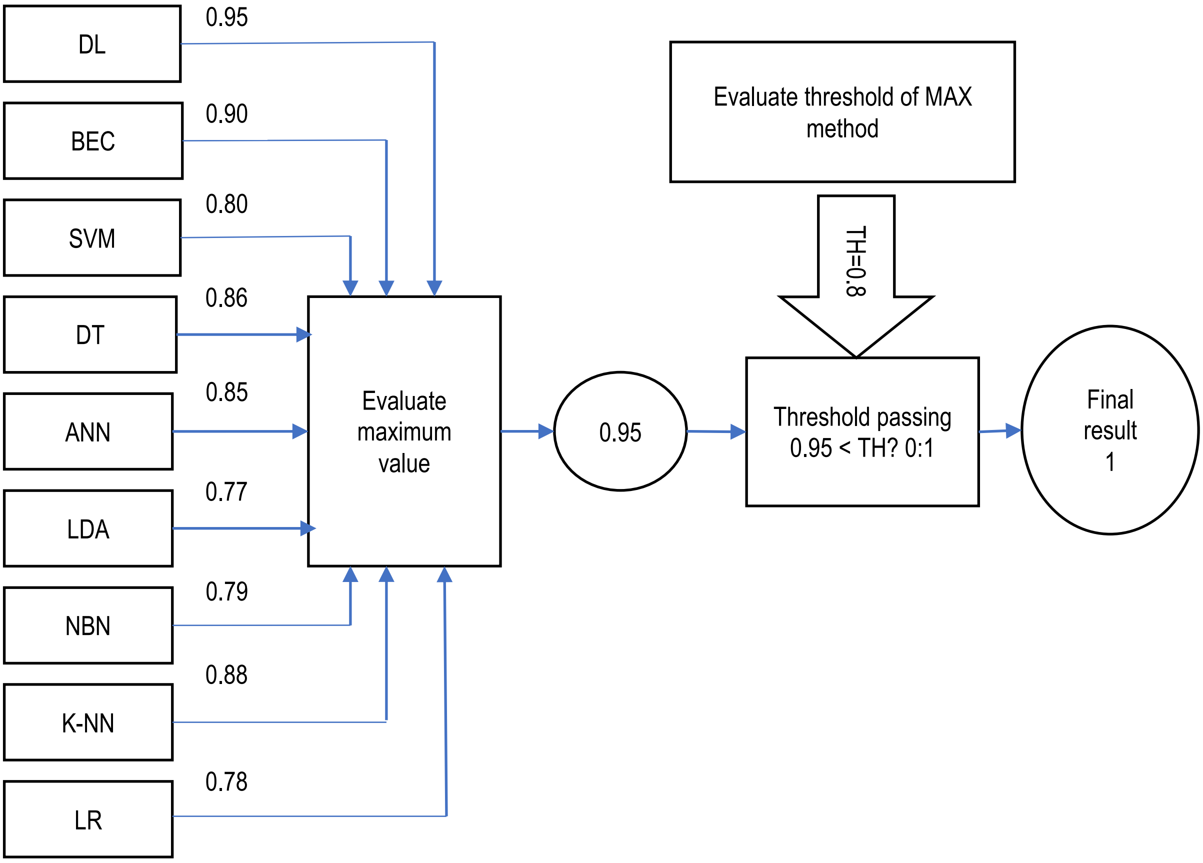


Figure 5.7: MAX method mechanism

Source: Author

5.3. Combination Method Results

The predictions of all nine individual classifiers were integrated to analyse testing committee methods results. This section illustrates the evaluation test measurement performance results utilised to measure the AVG, WAVG, MIN, MAX, MED, FC, MajVot, and CON methods' performance in classifying correct auditing opinions across the four tested datasets.

5.3.1. Consensus Combiner Model

Table 5.1 shows that consensus method classifier has the most powerful performance over the four datasets, as indicated by the evaluated parameters: average accuracy (95.1-98.6%), Brier scores (1.2-2.3%), F-measure (93.3-98.3%), specificity (89.3-97.1%), and sensitivity (99.5-99.9%). Figure 5.8 shows that the ROC curves for the four datasets are near to the corner, indicating that the CON model has good performance, especially in areas of skewed class distribution and balance classification error costs, based on high percentages of sensitivity, with low Type I Error rates.

Figure 5.9 shows reliability diagrams across the four datasets, showing that CON overestimated between 0 to 0.46 mean prediction value, after which the reliability line becomes closer to the diagonal line over the four datasets. Additionally, CON model is balanced in its detection of right audit opinion, by shown smaller gaps between Type I and II Error Rates across the datasets: 2019 (3.8%), 2018 (2.8%), 2017 (10.6%), and All-Data (12.5%).

However, CON has reduced performance in classifying correct audit opinion for the All-Data dataset, with reduces percentages for average accuracy (95.7%) and F-measure (93.3%), increased Brier score (2.3%), and an increasing gap between sensitivity and specificity rates (12.5%). CON obtained the best performance for the year 2017 dataset.

Table 5.1: Evaluation test results of CON

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	98.1%	98.6%	97.5%	95.7%
Type II Error	0.1%	0.1%	0.1%	0.5%
Type I Error	3.9%	2.9%	10.7%	13%
Specificity	96.1%	97.1%	89.3%	87%
Sensitivity	99.9%	99.9%	99.9%	99.5%
AUC	98.9%	99.5%	98.2%	98.8%
F-measure	98%	98.3%	94.9%	93.3%
Brier score	1.9%	1.2%	2.1%	2.3%
AURD	18.3%	23%	23.8%	16.9%

Source: Author

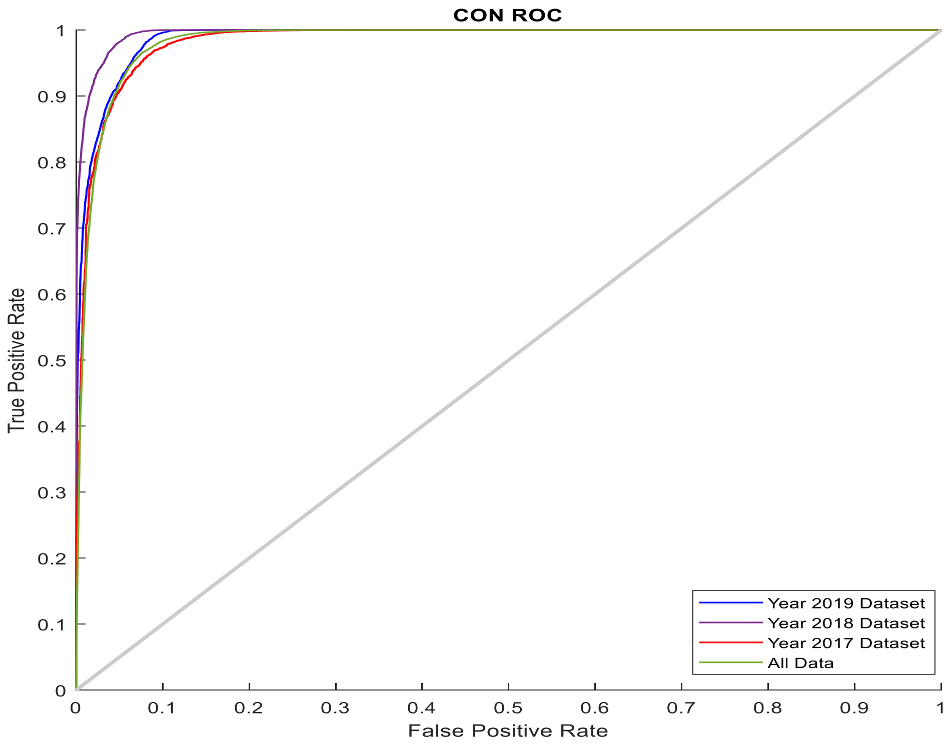


Figure 5.8: CON ROC curves over four datasets

Source: Author

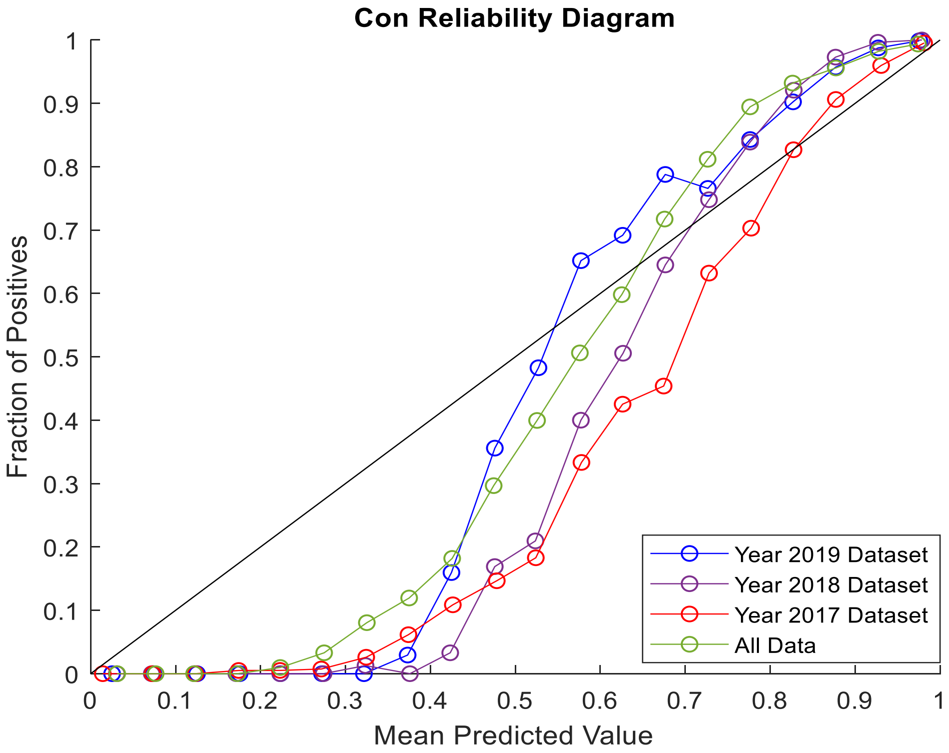


Figure 5.9: CON reliability diagrams over four datasets

Source: Author

5.3.2. Fuzzy Logic Combiner Model

FC method had better performance than MajVot, MIN, MAX, and MED methods. As illustrated in Table 5.2, FC method has good performance for correct audit opinion classification for the year 2019 and 2018 datasets, with fewer differences between the number of the incorrect unqualified and incorrect qualified firms compared to the year 2017 and All-Data datasets, with the highest rates of average accuracy, specificity, F-measure, and AUC, with lower Type I and II Errors and Brier scores. On the other hand, FC displayed weak performance for All-Data, with dramatically decreased performance accuracy (93.5%), increased Brier score (5.2%) and Type II Error (1.8%), and lower rates of sensitivity (98.2%) and specificity (81%). Figure 5.10 shows that the FC has better ROC curves for the year 2019 and 2018 datasets compared to the year 2017 and All-Data datasets.

The reliability diagrams of FC method (Figure 5.11) illustrate that it had better reliability diagrams across all four datasets compared to the AVG, WAVG, MIN, MAX, MED and CON methods, with the prediction values being relatively close to the actual target values. As shown in Table 5.2 FC achieved AURD of less than 11%. Figure 5.11 shows that fuzzy method has best reliability diagram at All-Data but at year 2019 dataset has the worst reliability diagram compared to year 2018, year 2017 and All-Data.

Table 5.2: Evaluation test results of FC

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	96.5%	96.8%	96.4%	93.5%
Type II Error	0.6%	0.2%	0.2%	1.8%
Type I Error	9%	6.6%	18.9%	19%
Specificity	91%	93.4%	81.1%	81%
Sensitivity	99.4%	99.8%	99.8%	98.2%
AUC	98.7%	98.9%	98.5%	96.8%
F-measure	95.4%	96.7%	91.2%	90.4%
Brier score	2.9%	2.48%	2.98%	5.2%
AURD	10.8%	6.6%	9.2%	6.5%

Source: Author

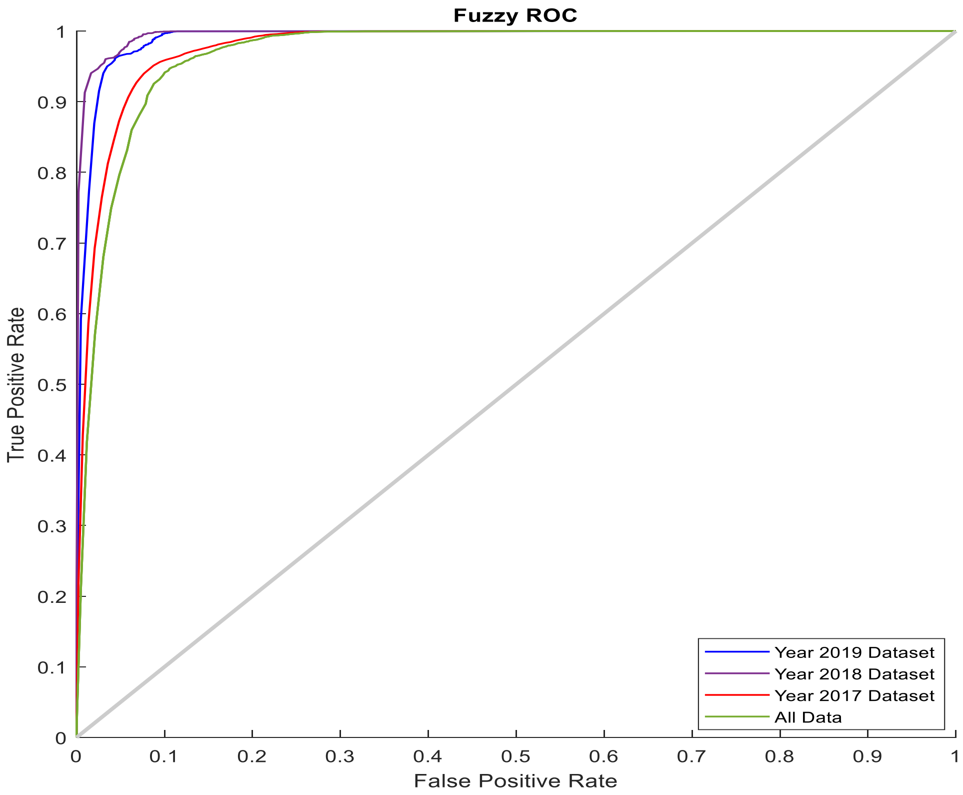


Figure 5.10: FC ROC curves over four datasets

Source: Author

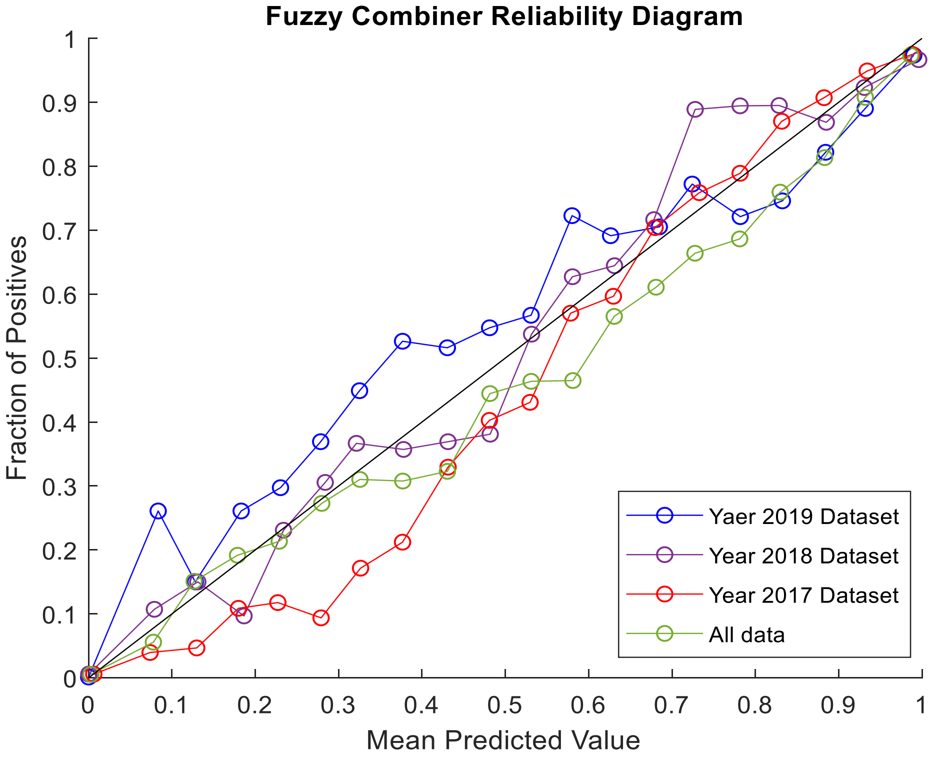


Figure 5.11: FC Reliability diagrams over four datasets

Source: Author

5.3.3. Average Method

Table 5.3 shows that AVG has good performance across the four datasets to indicate correct audit opinion types according to the nine performance measurements, including F-measures of 89.9-95.3%. AVG achieved average accuracy of around 96% for the year 2019, 2018, and 2017 datasets, but 93.8% for All-Data. Across the four datasets, AVG achieved 2.9-4.8% Brier score, 80-92% specificity, above 98.7% sensitivity (i.e., lower 1.3% incorrectly qualified firms flagged as unqualified class).

Figure 5.12 shows ROC curves relative to the Y-axis (true positive rate) and X-axis (false positive rate), reflecting that AVG had good ability to distinguish between qualified and unqualified class across the datasets, because all curves are near to the corner (1), with AUC rates of 98.5% (2019), 99.4% (2018), 97.8% (2017), and 97.4% (All-Data).

AVG had AURD percentages of 17.6-24%; Figure 5.13 illustrates that there were around eight bins against 0 fraction actual targets, with mean prediction values from 0 to 0.4. The mean predictions from AVG foundation bins may struggle for forecasting probabilities near 1 and 0, due to variation in the underlying base model prejudicing predictions. Due to predictions being restricted to an interval between 0 and 1, errors caused via variation to be near 1 and 0.

Table 5.3: Evaluation test results of AVG

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	96.3%	96.6%	96.2%	93.8%
Type II Error	0.7%	0.5%	0.2%	1.2%
Type I Error	9.1%	8%	20%	17.6%
Specificity	90.9%	92%	80%	82.4%
Sensitivity	99.3%	99.5%	99.8%	98.8%
AUC	98.5%	99.4%	97.8%	97.4%
F-measure	95.3%	96.1%	90.8%	89.9%
Brier score	3.6%	2.9%	3.7%	4.8%
AURD	18.26%	23.12%	24%	17.6%

Source: Author

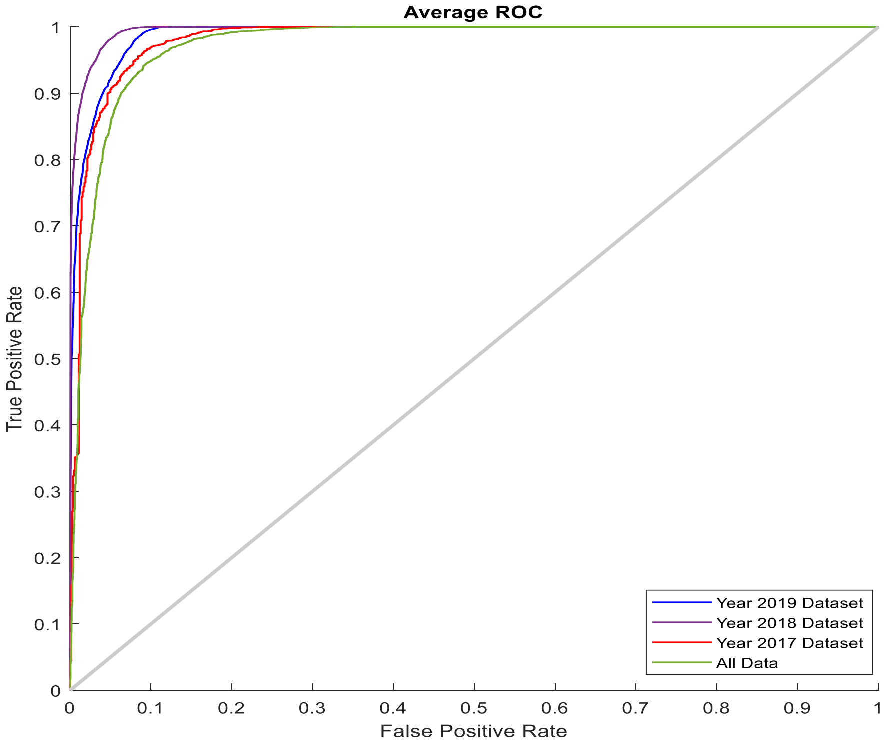


Figure 5.12: AVG ROC curves over four datasets

Source: Author

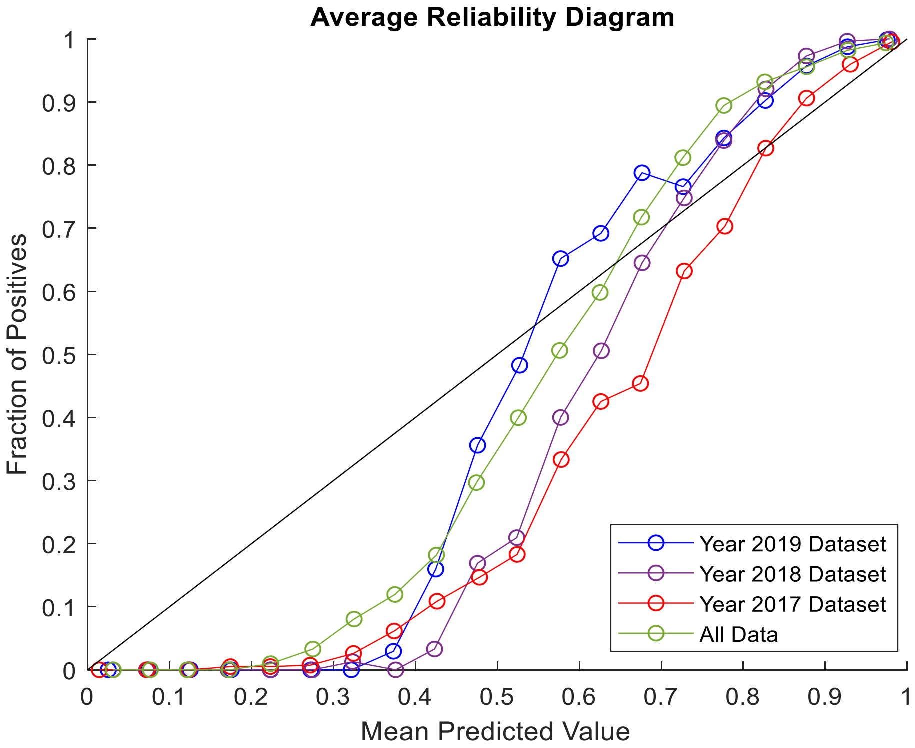


Figure 5.13: AVG reliability diagrams over four datasets

Source: Author

5.3.4. Weighted Average Method

WAVG gives weight to each classifier, as shown in Table 5.7, which determine which single classifiers have better or less accurate performance, because the weights affect the final results of WAVG. For instance, DPL, BEC, and DT have higher weights over the four datasets due to having better performance than NBN, K-NN, ANN, LR, LDA, and SVM over all four datasets. Weights thus correctly represent the performance of each classifier.

The empirical results shown in Table 5.4 indicate that across the four tested datasets, WAVG method revealed superior ability to classify audit opinion correctly, attaining above 95.4% average accuracy, 92.6% F-measure, 98.7% sensitivity, 84.9% specificity, and 98.2% AUC, with fewer qualified and unqualified incorrect classifications for the year 2019 and 2018 datasets. WAVG is thus a good model to distinguish between audit opinions and classify audit opinion correctly.

Table 5.4 illustrates that for the year 2019 and 2018 datasets, WAVG classifier has good ability to detect the right audit opinion, as indicated by average accuracy (96.9-97%), F-measure (96.2-96.9%), Brier score (2.8-2.3%), AURD (5.4-6.3%), Type I Error (6.1-6.5%), Type II Error (1.2-0.2%), specificity (93.5-93.9%), and AUC (98.7-99.2%). Additionally, WAVG was balanced in its detection of right audit opinion, with smaller gaps in the valuation results for the 2018 and 2019 datasets. For instance, as shown in Table 5.4 there were 5.3% and 5.9% gaps between the Type I and Type II Error rates for the 2019 and 2018 datasets, respectively. Additionally, Figure 5.14 shows that WAVG had ROC close to corner 1 for 2019 and 2018, and the reliability diagrams indicate that the predicted value is close to actual targets.

On the other hand, WAVG method evaluation results showed reduced ability to correctly classify audit opinion for the year 2017 and All-Data datasets, with decreased average accuracy, F-measure, AUC, and specificity rates, increasing Brier scores, Type I Error, and AURD, and a dramatic rise in the gap between the false negative and false positive rates to 13.8% for the year 2017 dataset and 14.4% for the All-Data dataset. Figure 5.15 shows the reliability diagrams for the year 2017 and All-Data datasets, presenting that WAVG reliability diagrams are over-forecasting.

Table 5.4: Evaluation test results of WAVG

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	96.9%	97%	96.7%	95.5%
Type II Error	1.2%	0.2%	0.2%	0.6%
Type I Error	6.5%	6.1%	14%	15%
Specificity	93.5%	93.9%	86%	85%
Sensitivity	98.8%	99.8%	99.8%	99.4%
AUC	98.7%	99.2%	98.5%	98.3%
F-measure	96.2%	96.9%	93.4%	92.7%
Brier score	2.8%	2.3%	2.98%	4.3%
AURD	5.4%	6.3%	26.5%	17.2%

Source: Author

Table 5.5: Weighted average coefficients for nine classifiers across four datasets

	Year 2019	Year 2018	Year 2017	All-Data
LR	0.000543	0.001286	0.00013	0.00031
K-NN	0.0009	0.000485	0	0.0005729
NBN	0.000061	0.000668	0	0.0003584
LDA	0.00094	0.00041548	0.000463677	0
DT	0.00727	0.001864	0.17349	0. 18
ANN	0.000417	0.00033287	0.00903	0.0023161
SVM	0.00052	0.000284	0.00035393	0.000244
BEC	0.01254	0.092	0.2911	0. 2915946
DL	0.976809	0.90266465	0.525849702	0.524604

Source: Author

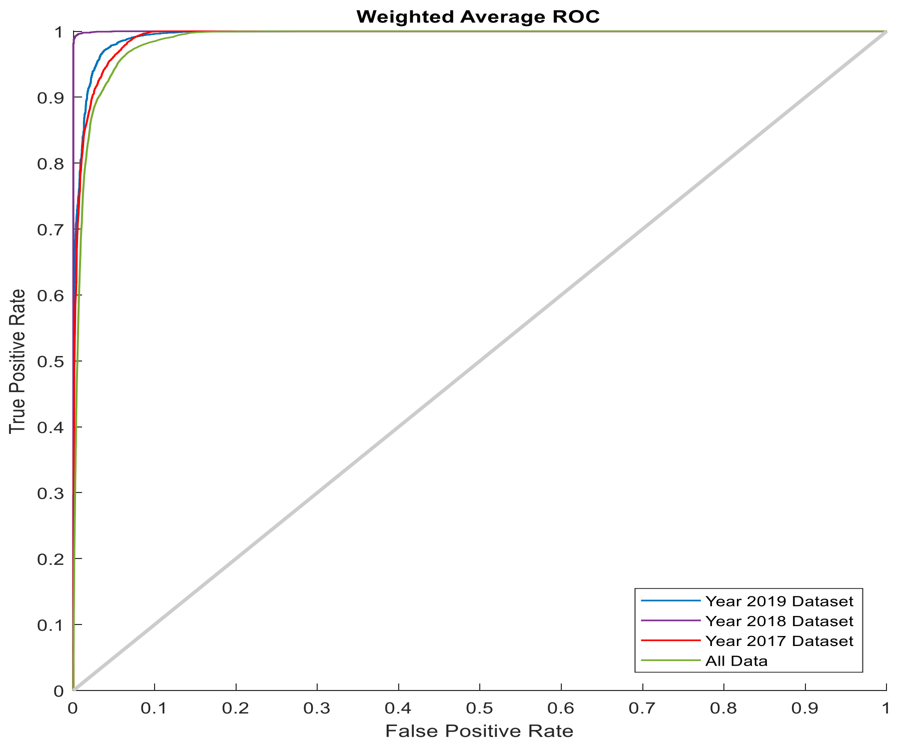


Figure 5.14: WAVG ROC curves over four datasets

Source: Author

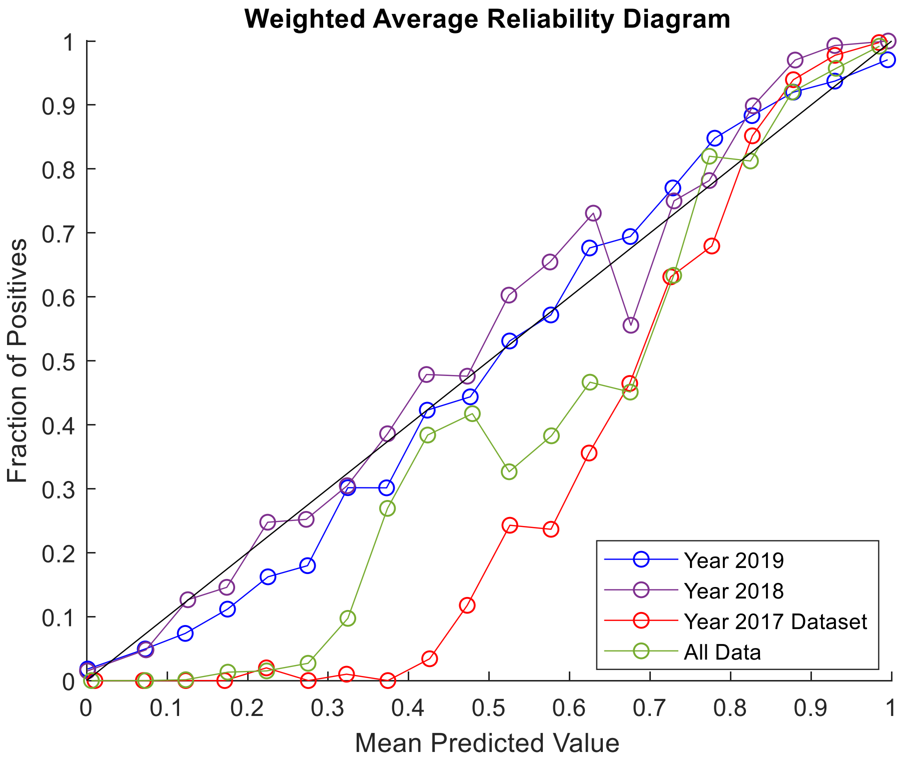


Figure 5.15: WAVG reliability diagrams over four datasets

Source: Author

5.3.5. Median Method

Table 5.6 indicates that MED had the best performance evaluation results for the year 2019 and 2018 datasets, with average accuracy above 96.1%, F-measure above 95%, AUC above 98.4%, and Brier score from 3.61-3%. Likewise, for these two datasets, around 10% of healthy firms were incorrectly classified, with 99.3-99.8% sensitivity rates; and 90.5-90.2% of qualified firms were correctly classified, with 0.7-0.2% incorrectly classified qualified opinion. AUC percentages were approximately 99% (with ROC curves near to 1), and AURDs 13.99-20.3%. MED method thus has good performance evaluation for the year 2019 and 2018 datasets.

However, MED had reduced performance for the year 2017 and All-Data datasets, albeit with respectable accuracy, sensitivity, Type II Error, F-measure, and AUC rates. Nevertheless, MED method at year 2017 dataset and All-Data detected substantial gaps between sensitivity and specificity rates: 20.4% for 2017 and 17.4% for All-Data. This indicates MED's poor performance to distinguish between qualified and unqualified audit opinion for these two datasets.

As shown in Figure 5.17, MED had relatively better calibrated probabilities for All-Data compared to other datasets' reliability diagrams, hugging the diagonal line due to this MED had lower AURD (12.9%) for All-Data. For the year 2017 dataset MED had higher it had higher AURD (26.8%), and the shape of the reliability diagram indicates more over-forecasting compared to other calibrated probabilities shapes. Likewise, Figure 5.16 shows that MED method has a better ROC curve (near to corner 1) over the year 2019 and 2018 datasets than for the year 2017 and All-Data datasets.

Table 5.6: Evaluation test results of MED

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	96.2%	96.5%	96.1%	93.3%
Type II Error	0.7%	0.2%	0.2%	1.4%
Type I Error	9.5%	9.8%	20.6%	18.8%
Specificity	90.5%	90.2%	79.4%	81.2%
Sensitivity	99.3%	99.8%	99.8%	98.6%
AUC	98.5%	98.7%	97.3%	97.4%
F-measure	95.1%	95.2%	90.6%	90.7%
Brier score	3.61%	3%	3.9%	5.17%
AURD	13.99%	20.3%	26.8%	12.9%

Source: Author

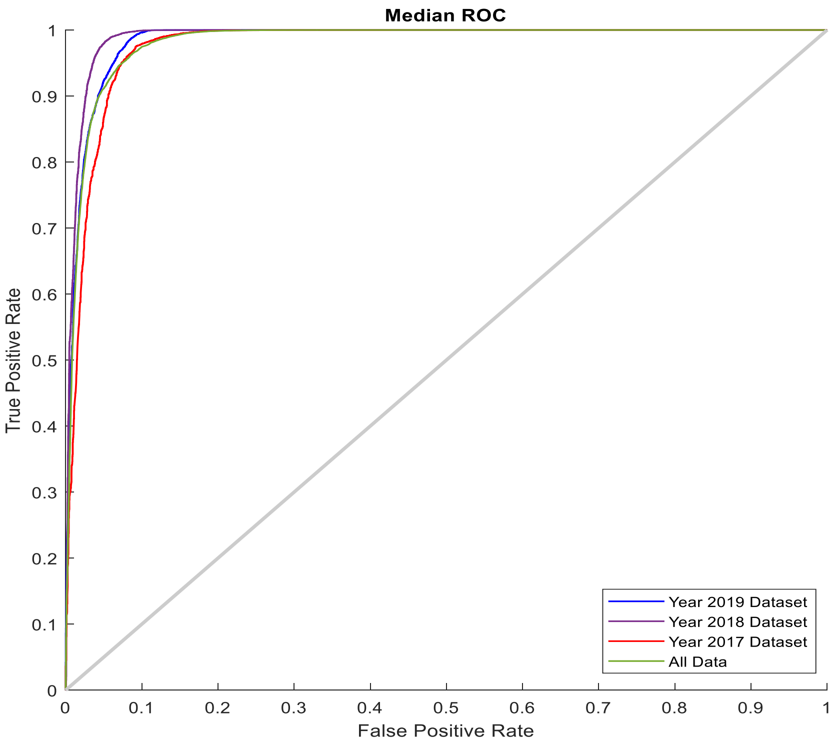


Figure 5.16: MED ROC curves over four datasets

Source: Author

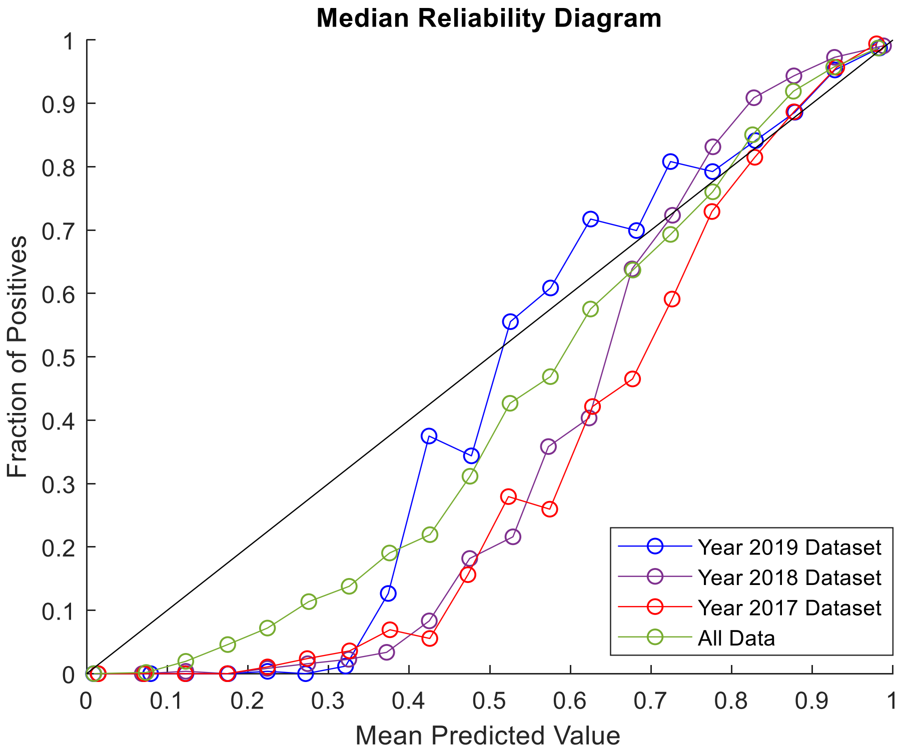


Figure 5.17: MED reliability diagrams over four datasets

Source: Author

5.3.6. Majority Voting Method

The performance evaluation of MajVot method across the four tested datasets illustrated bad performance compared to MED, AVG, WAVG, and CON classifiers. However, for all four datasets classification results, MajVot had good ability to correctly classify qualified opinion (specificity) and unqualified opinion (sensitivity rate), due to smaller gaps between them (Table 5.7).

All four datasets, as shown in Table 5.7, MajVot method achieved average accuracy rates of 90.8-96.5%, AUC rates of 97.7-98.9%, F-measure of 89-95.4%, and Brier scores of 3.2-7.9%. The MajVot ROC curves (Figure 5.18) for the year 2018 and 2019 datasets are near to corner 1, with better curves than for the other datasets, but the All-Data ROC is poor, reflecting larger Type I Error (19.4%) and lower specificity (95.8%) compared to other datasets. For All-Data MajVot shown poor ability to distinguish between qualified and unqualified correctly by given higher gap between sensitivity and specificity at 15.2%. In addition, Figure 5.19 shows that all reliability diagrams for MajVot method across the four tested datasets are under the diagonal line, which means that the MajVot model has over-forecast, achieving AURDs ranging from 17.8-23.8% for the four datasets.

Table 5.7: Evaluation test results of MajVot

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average accuracy	96.1%	96.5%	96%	90.8%
Type II Error	0.6%	0.3%	2.1%	4.2%
Type I Error	10%	9.4%	8.6%	19.4%
Specificity	90%	90.6%	91.4%	80.6%
Sensitivity	99.4%	99.7%	97.9%	95.8%
AUC	98%	98.9%	98.6%	97.7%
F-measure	94.4%	95.4%	94.8%	89%
Brier score	3.8%	3.2%	3.97%	7.9%
AURD	19.6%	20.02%	17.8%	23.8%

Source: Author

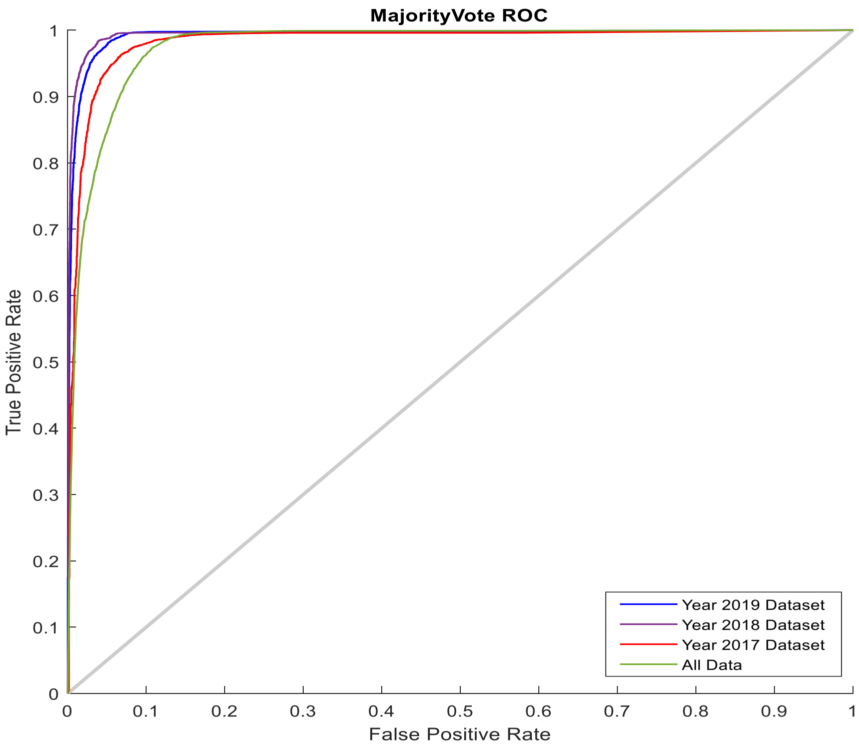


Figure 5.18: MajVot ROC curves over four datasets

Source: Author

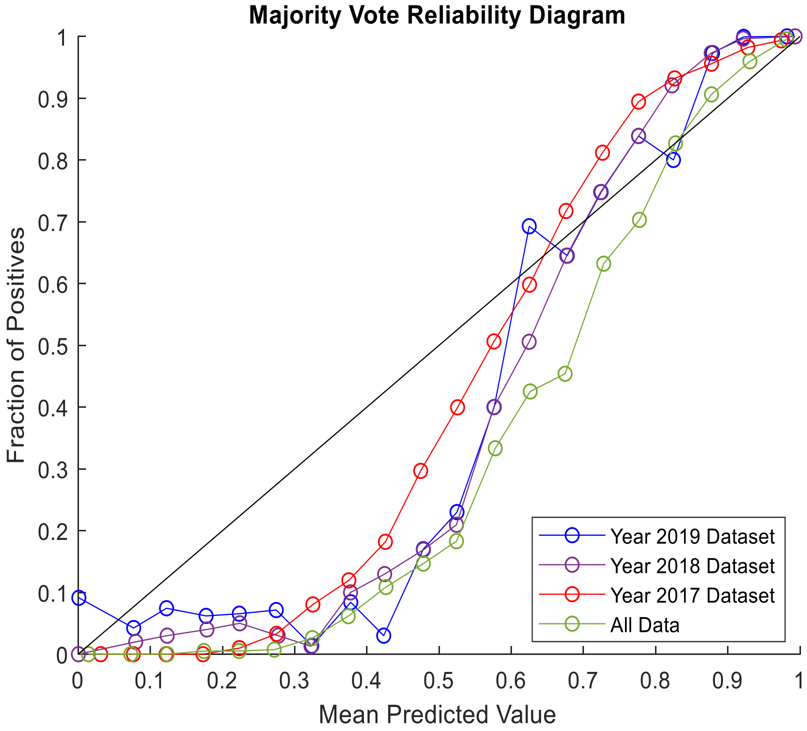


Figure 5.19: MajVot reliability diagrams over four datasets

Source: Author

5.3.7. MIN Method

Table 5.8 presents MIN classifier performance evaluation results across the four tested datasets shown had good performance to classify audit opinion correctly. Overall, as shown in Figure 5.20, MIN classifier had good ROC curves across the four tested datasets (between 0.9 and 1), with AUC above 96.3%. Figure 5.21 indicates that MIN method had a better reliability diagram for the All-Data dataset compared to those for the year 2019, 2018 and 2017 datasets.

Table 5.8 shows that MIN achieved better performance for the year 2018 dataset, indicated by F-measure (95.3%), average accuracy (96.3%), AUC (98.2%), specificity (91.2%), Brier score (3.3%), Type I Error (8.8%), and lower gap between Type I and Type II Error (7.8%). MIN also had its best ROC curve for the year 2018 dataset, but a poorer reliability diagram (Figure 5.21), indicating higher AURD. Reliability diagrams for all four datasets are under the diagonal line, meaning that the MIN classifier has over-forecast (the production values are too large). On the other hand, the MIN shown has poor performance for All-Data dataset compered to other datasets in terms of the evaluated parameters: average accuracy (94.8%), F-measure (92.3%), sensitivity (84.1%), Brier score (4.4%), and AUC (96.8%, indicating a poor ROC), with higher Type I Error (15.9%), and the biggest gap between true positive and true negative rates of 15% compared to the gap achieved in year 2019 and year 2018 datasets.

Table 5.8: Evaluation test results of MIN

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	95.9%	96.3%	94.7%	94.8%
Type II Error	1.2%	1%	0.7%	0.9%
Type I Error	9.4%	8.8%	16.4%	15.9%
Specificity	90.6%	91.2%	83.6%	84.1%
Sensitivity	98.8%	99%	99.3%	99.1%
AUC	97.7%	98.2%	96.4%	96.8%
F-measure	94.5%	95.3%	92%	92.3%
Brier score	3.97%	3.3%	5%	4.4%
AURD	15.9%	20.2%	16.8%	11.8%

Source: Author

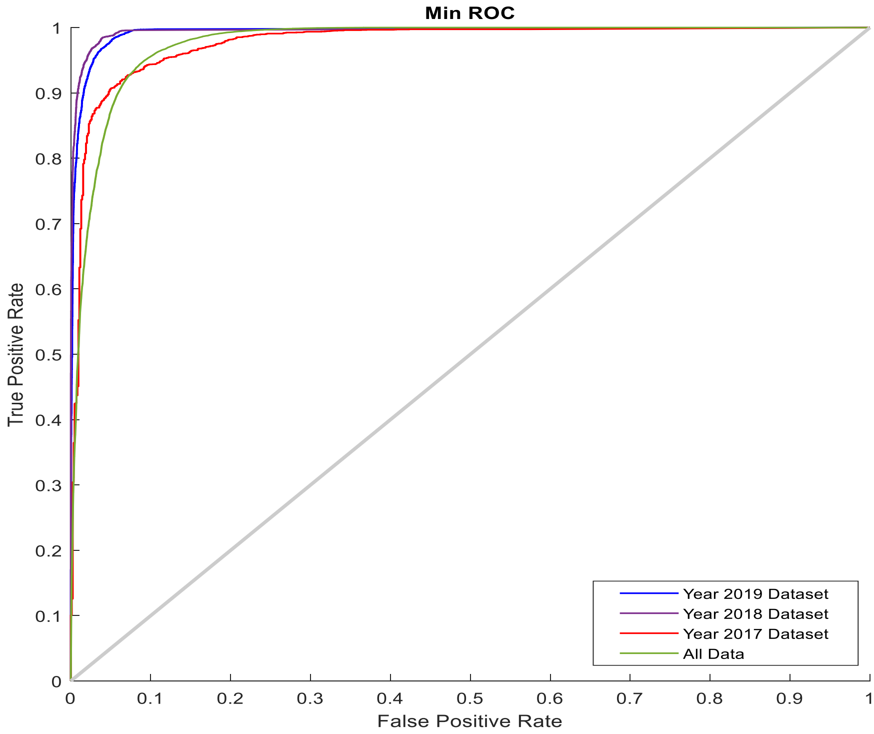


Figure 5.20: MIN ROC curves over four datasets

Source: Author

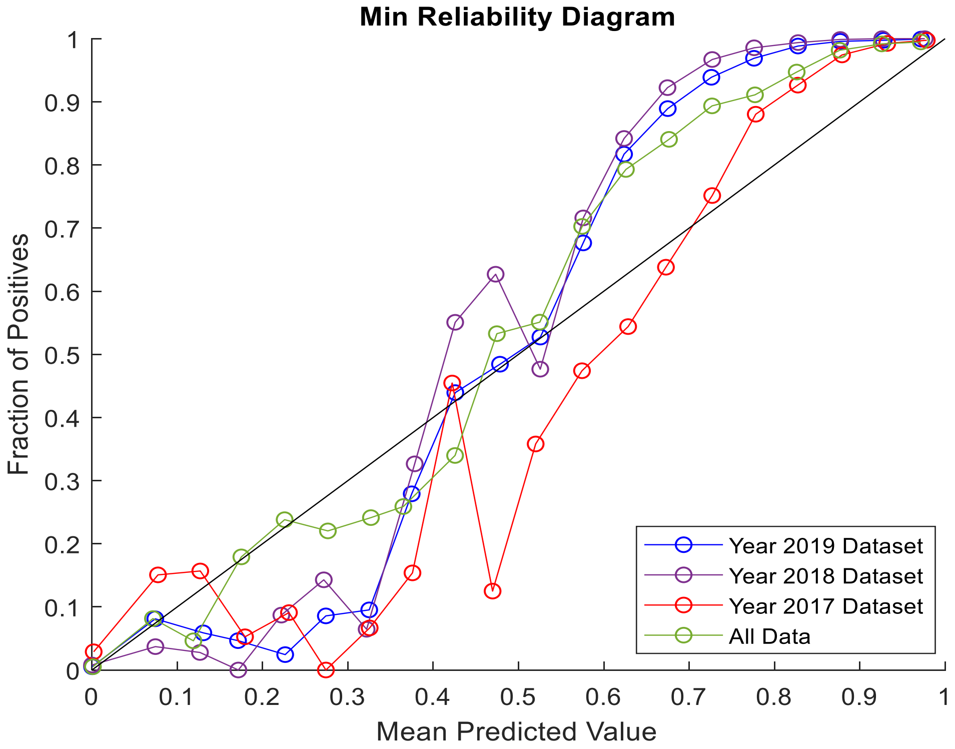


Figure 5.21: MIN reliability diagrams over four datasets

Source: Author

5.3.8. MAX Method

The MAX method performance evaluation (Table 5.9) revealed poorer performance compared to the AVG, WAVG, MED, MajVot, and CON classifiers, although it had superior ability to correctly classify unqualified audit opinion. Across all four datasets, MAX classifier had lower gaps in identifying different types of audit opinion, according to the rates of specificity, sensitivity, and Type I and Type II Error: 5.9% (2019), 6.7% (2018), 21.8% (2017), and 8.4% (All-Data). For all four datasets MAX achieved close results, with small variations in average accuracy, Brier score, sensitivity, Type I Error, and AURD. These results lead to the conclusion that the size of dataset does not significantly affect MAX classifier performance. MAX classifier had good ROC curves for all datasets, with AUC rates ranging from 94.6-98.5% (Figure 5.22). The reliability diagrams (Figure 5.23) are similar, with AURDs ranging from 21.01-25%.

Table 5.9: Evaluation test results of MAX

	Year 2019	Year 2018	Year 2017	All-Data
Average accuracy	95.4%	96.2%	94%	94.4%
Type II Error	3.2%	2%	0.6%	3.1%
Type I Error	9.1%	8.7%	22.4%	11.5%
Specificity	90.9%	91.3%	77.6%	88.5%
Sensitivity	96.8%	98%	99.4%	96.9%
AUC	97.8%	98.5%	94.6%	95.8%
F-measure	94%	94.8%	89.6%	93%
Brier score	4.2%	3.4%	5.53%	4.6%
AURD	25%	24.7%	22%	21.01%

Source: Author

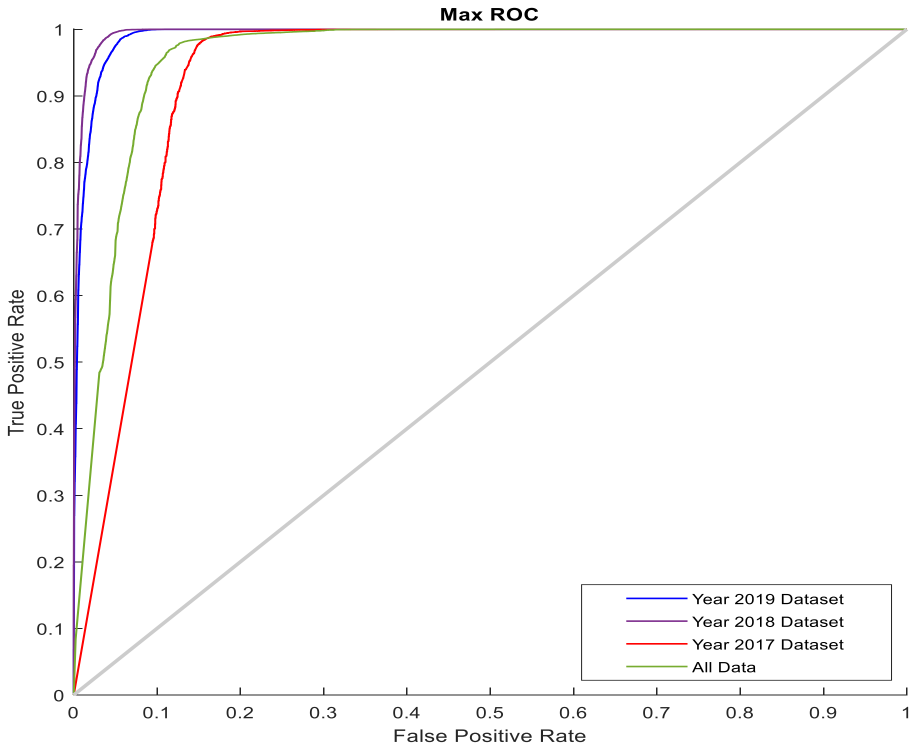


Figure 5.22: MAX ROC curves over four datasets

Source: Author

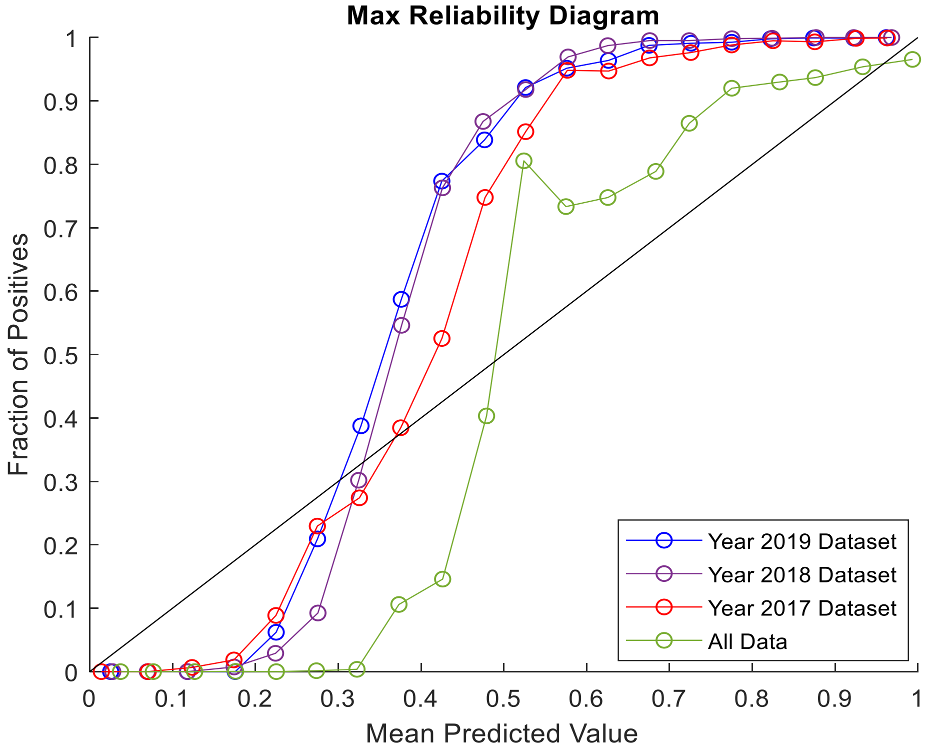


Figure 5.23: MAX reliability diagrams over four datasets

Source: Author

5.4. Comparative Analysis and Discussion

This section compares and discusses the evaluation metric results achieved from the studied classifier methods, to ascertain which of them best classify audit opinions. Tables 5.10-5.13 illustrate the valuation metrics results utilised to compare the performance of all eight combiner methods.

Table 5.10: Comparing evaluation test results of all models with 2019 dataset

	Year 2019 Dataset								
	Aver Acc.	**Type II Error**	**Type I Error**	**Specificity**	**Sensitivity**	**F-measure**	**AUC**	**Brier score**	**AURD**
CON	98.1%	0.1%	3.9%	96.1%	99.9%	98%	98.9%	1.9%	18%
WAVG	96.9%	1.2%	6.5%	93.5%	98.8%	96.2%	98.7%	2.8%	5.4%
FC	96.5%	0.6%	9%	91%	99.4%	95.4%	98.7%	2.9%	10.8%
AVG	96.3%	0.7%	9.1%	90.9%	99.3%	95.3%	98.5%	3.6%	18.26%
MED	96.2%	0.7%	9.5%	90.5%	99.3%	95.1%	98.5%	3.61%	13.99%
MajVot	96.1%	0.6%	10%	90%	99.4%	94.4%	98%	3.8%	19.6%
MIN	95.9%	1.2%	9.4%	90.6%	98.8%	94.5%	97.7%	3.97%	15.9%
MAX	95.4%	3.2%	9.1%	90.9%	96.8%	94%	97.8%	4.2%	25%

Source: Author

Table 5.11: Comparing evaluation test results of all models with 2018 dataset

	Year 2018 Dataset								
	Aver Acc.	**Type II Error**	**Type I Error**	**Specificity**	**Sensitivity**	**F-measure**	**AUC**	**Brier score**	**AURD**
CON	98.6%	0.1%	2.9%	97.1%	99.9%	98.3%	99.5%	1.2%	23%
WAVG	97%	0.2%	6.1%	93.9%	99.8%	96.9%	99.2%	2.3%	6.3%
FC	96.8%	0.2%	6.6%	93.4%	99.8%	96.7%	98.9%	2.48%	6.6%
AVG	96.6%	0.5%	8%	92%	99.5%	96.1%	99.4%	2.9%	23.12%
MED	96.5%	0.2%	9.8%	90.2%	99.8%	95.2%	98.7%	3%	20.3%
MajVot	96.5%	0.3%	9.4%	90.6%	99.7%	95.4%	98.9%	3.2%	20.02%
MIN	96.3%	1%	8.8%	91.2%	99%	95.3%	98.2%	3.3%	20.2%
MAX	96.2%	2%	8.7%	91.3%	98%	94.8%	98.5%	3.4%	24.7%

Source: Author

Table 5.12: Comparing evaluation test results of all models with 2017 dataset

	Year 2017 Dataset								
	Aver Acc.	**Type II Error**	**Type I Error**	**Specificity**	**Sensitivity**	**F-measure**	**AUC**	**Brier score**	**AURD**
CON	97.5%	0.1%	10.7%	89.3%	99.9%	94.9%	98.2%	2.1%	23.8%
WAVG	96.7%	0.2%	14%	86%	99.8%	93.4%	98.5%	2.98%	26.5%
FC	96.4%	0.2%	18.9%	81.1%	99.8%	91.2%	98.5%	2.98%	9.2%
AVG	96.2%	0.2%	20%	80%	99.8%	90.8%	97.8%	3.7%	24%
MED	96.1%	0.2%	20.6%	79.4%	99.8%	90.6%	97.3%	3.9%	26.8%
MajVot	96%	2.1%	8.6%	91.4%	97.9%	94.8%	98.6%	3.97%	17.8%
MIN	94.7%	0.7%	16.4%	83.6%	99.3%	92%	96.4%	5%	16.8%
MAX	94%	0.6%	22.4%	77.6%	99.4%	89.6%	94.6%	5.53%	22%

Source: Author

Table 5.13: Comparing evaluation test results of all models with All-Data

	All-Data								
	Aver Acc.	**Type II Error**	**Type I Error**	**Specificity**	**Sensitivity**	**F-measure**	**AUC**	**Brier score**	**AURD**
CON	95.7%	0.5%	13%	87%	99.5%	93.3%	98.8%	2.3%	16.9%
WAVG	95.5%	0.6%	15%	85%	99.4%	92.7%	98.3%	4.3%	17.2%
MIN	94.8%	0.9%	15.9%	84.1%	99.1%	92.3%	96.8%	4.4%	11.8%
MAX	94.4%	3.1%	11.5%	88.5%	96.9%	93%	95.8%	4.6%	21.01%
AVG	93.8%	1.2%	17.6%	82.4%	98.8%	91.3%	97.4%	4.8%	17.6%
FC	93.5%	1.8%	19%	81%	98.2%	90.4%	96.8%	5.2%	6.5%
MED	93.3%	1.4%	18.8%	81.2%	98.6%	90.7%	97.4%	5.17%	12.9%
MajVot	90.8%	4.2%	19.4%	80.6%	95.8%	89%	97.7%	7.9%	23.8%

Source: Author

The following discussion is based on the evaluation test measurement outputs of all four datasets as presented in Tables 5.10-5.13.

It can be seen from the tables that all eight committee methods had acceptable performance in terms of correctly classifying audit opinion, as indicated in the evaluation parameter ranges: average accuracy (90.8-98.6%), Brier score (1.2-7.9%), F-measure (89-98.3%), and AUC (above 94.6%). Likewise, they all have specificity and sensitivity greater than 77.6%, and their Type I Error and Type II Error classification rates are below 22.4%, 4.2%, respectively.

CON method had superior ability to classify audit opinion correctly, outperforming all other committee methods (AVG, WAVG, MIN, MajVot, MAX, FC, and MED) over all four datasets in terms of obtaining higher accuracy (95.7-98.6%), specificity (87-97.1%), sensitivity (99.9-99.5%), F-measure (93.3-98.3%), and AUC (98.2-99.5%), with lower Brier scores, and Type I and II Error rates (incorrectly flagging items as another class). This means that it has excellent classifier performance, thus CON method is more effective in distinguishing audit opinion than the other methods based on obtaining the most favourable F-measure, ROC, and difference between Type I and Type II Error rates. For the year 2019 dataset, the CON model had a 3.8% gap between Type I and Type II Error rates.

After CON, WAVG had better performance than FC, MAX, MIN, AVG, MajVot, and MED methods, as indicated by its average accuracy (96.9% for 2019, 97% for 2018, 96.7% for 2017, and 95.5% for All-Data), F-measure (92.7-96.9%), and Brier scores (under 4.3%). WAVG had a lower gap than FC, MAX, MIN, AVG, MajVot, and MED methods between the number of incorrectly qualified companies (Type I Error rates) and number of incorrect unqualified companies (Type II Error rates): 5.3% for 2019, 5.9% or 2018, 13.8% for 2017, and 14.4% for All-Data. WAVG is thus a balanced method, able to distinguish between audit opinions correctly. However, for the year 2019 and 2018 datasets, the weighted average had lower area under reliability diagram rates compared to the FC, MIN, MAX, AVG, MajVot, MED, and CON methods.

FC method had lower AURD rates across all four datasets compared to the MIN, MAX, AVG, MajVot, MED, and CON methods, particularly for the year 2017 and All-Data datasets (the predicted values were relatively close to the diagonal line). For the year 2019, 2018, and 2017 datasets, FC had good performance in correct identifying qualified and unqualified companies compared to the MAX, MIN, AVG, MED, and MajVot methods, attaining higher accuracy rates (96.8-96.4%) and F-measure (90.4-96.7%), and lower Brier scores (2.48-2.98%). Except for CON and WAVG, Fuzzy logic had best evaluation measurements for the 2019 and 2018 datasets compared to the other models.

AVG had better performance to indicate correctly identified qualified and unqualified companies than the MIN, MAX, MED, and MajVot methods, with higher average accuracy and AUC rates and lower Brier scores. However, for the All-Data dataset AVG only outperformed the FC, MED, and MajVot methods. AVG and MED methods had similar results. For the year 2018 and 2017 datasets, AVG had average accuracy rates of 96.6-96.2%, F-measure of 96.1-90.8%, ACU rates of 99.4-97.8%, Brier scores of 2.9-3.7%, Type I Error rates of 8-20%, Type II Error rates of 0.5-0.2%, AURD rates of 23.12-24%, specificity rates of 92-80%, and sensitivity rates of 99.5-99.8%. MED method for the same datasets had average accuracy rates of 96.5-96.1%, F-measure of 95.2-90.6%, ACU rates of 98.7-97.3%, Brier scores of 3-3.9%, Type I Error rates of 9.8-20.6%, Type II Error rates of 0.2-0.2%, AURD rates of 20.3-26.8%, specificity rates of 90.2-79.4%, and sensitivity rates of 99.8-99.8%.

For the year 2019, 2018, and 2017 datasets, MajVot outperformed MIN and MAX methods in term of got higher average accuracy, AUC, F-measurement, sensitivity with lower Type II error and Brier score. But it had poorer evaluation results compared to MIN, MAX, MED, FC, AVG, WAVG, and CON methods in terms of lower average accuracy, F-measure, specificity, sensitivity, and AUC, with higher Type I and II Error rates, Brier score, and AURD. For the same datasets, MajVot had similar results to MED. For the year 2018 dataset, MED and MajVot had the same average accuracy (96.5%), Type I Error (0.2-0.1%), Type II Error (8-9.8%), specificity (90.2-90.6%), sensitivity (99.8-99.9%), AUC (98.7-98.9%), F-measure (95.2-95.4%), Brier score (3-3.2%), and AURD (20.3-20.02%).

In general, MIN method outperformed MAX method across all four datasets. MIN used the smallest prediction value of each row of all classifiers, attaining the final ranking level, while the MAX method conversely used the maximum prediction value of each row for all classifiers. After the improvement of the threshold in the MIN and MAX methods, as described in sections 5.2.7 and 5.2.8, they followed the same path to determine correct audit opinion. Due to this, the MIN and MAX methods had similar final evaluation parameter results. For instance, for the year 2019 dataset, MIN method had average accuracy of 95.9%, Type II Error rate of 1.2%, Type I Error rate of 9.4%, specificity of 90.6%, sensitivity of 98.8%, AUC of 97.5%, and Brier score of 3.97%; MAX method had average accuracy of 95.4%, Type II Error rate of 3.2%, Type I Error rate of 9.1%, specificity of 90.9%, sensitivity of 96.8%, AUC of 97.8%, Brier score of 4.6%. Both MIN and MAX method had inferior performance for the year 2019, 2018, and 2017 datasets compared to the other committee methods (AVG, MajVot, MED, FC, WAVG, and CON methods). However, for the All-Data dataset, the MIN and MAX methods outperformed AVG, MED, FC, and MajVot.

5.5. Statistical Significance Testing

This section evaluates the statistical significance of committee models' test results. Friedman statistical test was run to test the performance of the five best-performing models (CON, WAVG, FC, and MED), with post-hoc Bonferroni–Dunn pairwise comparison testing to determine whether the control model (selected as the control classification model) has statistically significant performance outcome differences compared to the other models. CON used as a comparator for the WAVG, FC, AVG, and MED.

Table 5.14 shows the Friedman test results for all classifiers and the four best ones over all four tested datasets. All classifiers reached Friedman outcomes higher than the best classifiers across the four datasets, because the size of the data points for all classifiers are greater than size of the entry data point for best classifiers. The Friedman test ranks the best combiner methods for each dataset separately. The alternative hypothesis is that there was significant difference between model performances; the null hypothesis is that no statistically significant performance variations exist between the best combiner methods, whereby the significance of each model's outputs is random. Tables 5.15-5.18 show that the best classifiers reached Friedman outcome values higher than $X^2_{0.1}(4)$= 7.8 and $X^2_{0.05}(4)$= 9.5 for each dataset, with P-values lower than the significance levels (α = 0.05, α = 0.1). Consequently, the null hypothesis is rejected, and the alternative hypothesis is approved.

Tables 5.15-5.18 present all possible pairwise comparisons for each two models, to evaluate their relative performance. The P-values for pairwise comparison results between the models are not high, which means that the significance of each model is suitable in terms of average accuracy, and there were significant variations between the best committee models.

Table 5.14: Friedman test results

Datasets	**Year 2019**	**Year 2018**	**Year 2017**	**All-Data**
Friedman X_F^2 (All classifiers)	6,535	8,460	8,587	9,474
Friedman X_F^2 (Best classifiers)	963	692	157	589

Source: Author

Table 5.15: Pairwise comparison results for 2019 dataset

FriedmanX_F^2 = 963 P-value= 0.000000243	**WAVG**	**FC**	**AVG**	**MED**	**Accuracy**
CON	0	0	0	0	97.4%
WAVG	-	0	0	0	96.9%
FC	-	-	0	0	96.5%
AVG	-	-	-	0	96.3%
MED	-	-	-	-	96.2%

Source: Author

Table 5.16: Pairwise comparison results for 2018 dataset

Friedman X_F^2= 692 P-value= 0.000007	**WAVG**	**FC**	**AVG**	**MED**	**Accuracy**
CON	1	0.016	0	0	97.6%
WAVG	-	0.008	0	0	97%
FC	-	-	0	0	96.8%
AVG	-	-	-	0	96.6%
MED	-	-	-	-	96.5%

Source: Author

Table 5.17: Pairwise comparison results for 2017 dataset

Friedman X_F^2 = 157 P-value= 0.00000046	**WAVG**	**FC**	**AVG**	**MED**	**Accuracy**
CON	0.001	1	0.239	0.4	97.1%
WAVG	-	0.079	0	0	96.7%
FC	-	-	0.005	0	96.4%
AVG	-	-	-	-	96.2%
MED	-	-	-	-	96.1%

Source: Author

Table 5.18: Pairwise comparison results for All-Data dataset

Friedman X_F^2= 589 P-value= 0	**WAVG**	**FC**	**AVG**	**MED**	**Accuracy**
CON	0.059	0	0	0	95.7%
WAVG	**-**	0	0	0	95.5%
FC	**-**	**-**	0.296	0	93.5%
AVG	**-**	**-**	**-**	0	93.8%
MED	**-**	**-**	**-**	**-**	93.3%

Source: Author

Having proved the alternative hypothesis that all committee models do not have the same performance, the models do not have the same average ranks. Post-hoc for Bonferroni–Dunn pairwise comparison test compares between the five best committee classifiers. The comparison is done through comparing the average rankings achieved from Freidman test for each committee combiner models, with two cut-lines representing the threshold for the better performing model (at significance levels α = 0.05 and α = 0.1). These two cut-lines are calculated by the sum of the CD from Bonferroni-Dunn, at α = 0.05 and α = 0.1, with the lowest rank referring to the best classifier performance. In our case, $CD_{0.05} = 5.3$ where $q_{0.05} = 2.724$, and $CD_{0.1} = 4.85$ where $q_{0.1} = 2.498$. The first cut-line is equal to the sum $CD_{0.1} = 4.85$, with CON rank at 1; and the second cut-line is equal to the sum the $CD_{0.05} = 5.3$, with CON rank at 1.

Figure 5.24 presents that all committee models are below the lower two cut-lines except MIN and MAX models, which are above, reflecting poorer performance to classify audit opinion correctly. However, CON, WAVG, and AVG models have better performance to classify audit opinion compared to the others. CON model presented the best performance, and MAX the worst, obtaining the average rank of 7.8, far from both cut-lines.

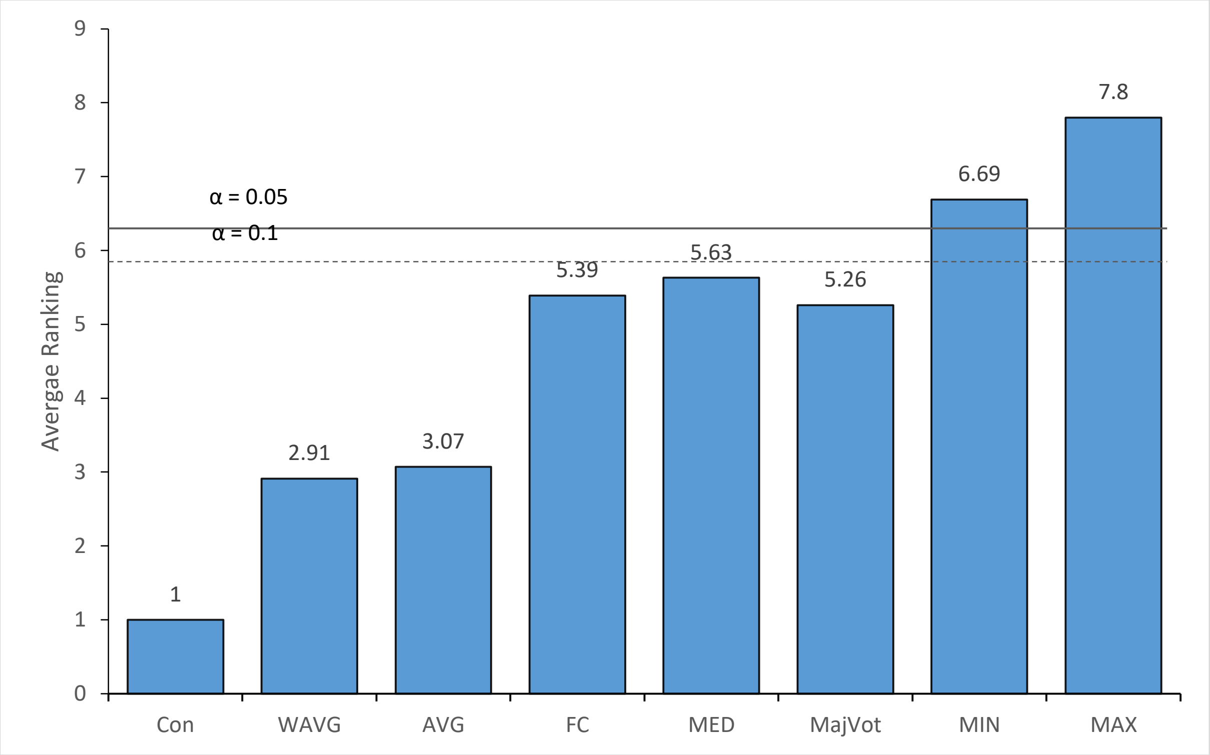


Figure 5.24: Bonferroni-Dunn correction for Committee combiner models, with significance levels.

Source: Author

5.6. Classification Model Training Time

To evaluate the committee combiner model and base single classifiers, the pre-process steps of data imputation, normalisation, and feature selection are applied, as described in Chapter 3. The pre-process steps of this study are conducted using SPSS version on a 25 GB RAM. This process took around 36 minutes. In general, more attributes and larger dataset sizes require longer time to do this phase. Subsequently, the experiments of individual classifiers and committee combiner models are conducted using MATLAB 2019a version on an 8 GB RAM personal computer with 3.4 GHz, Intel CORE i7, and Microsoft Windows 10 operating system. Base individual classifiers can be trained. Table 5.19 displays the training time taken by each individual classifier. Table 5.19 shows that across the four datasets, DPL required a longer training time compared to the other combiner models and single classifiers, because of its greater complexity of construction. LDA and DT completed the training process more quickly than the other single classifiers, taking 5.792 and 6.6 seconds (respectively) for the 2019 dataset, while other single models took more time to process the dataset.

After computing all base classifiers, the final phase is to evaluate combiner model outputs. As seen from Table 5.20, combiner models did not take as much training time as single classifiers, which took longer to compute required data. CON took less time for data training across the four datasets, including 2192.1354 seconds for the year 2019 datasets. Consequently, it was selected rather than other combiner models, as it can expediently offer new, different audit opinions in a fraction of a second.

Table 5.19: Training time for single models in seconds

Dataset	Year 2019	Year 2018	Year 2017	All-Data
DPL	1040	964	793.35	3800
BEC	64.371	58.181	35.86	5027.7
DT	6.5594	7.61	6.6362	582.17
ANN	20.1	23.49	12.4	682.92
SVM	175.3	193.32	292.41	1426.5
LR	20.313	9.9442	9.062	362.46
LDA	5.792	5.28	4.3623	367.41
K-NN	386.99	342.24	323.34	5159.5
NBN	468.11	475.14	485.94	4166.9

Source: Author

Table 5.20: Training time for committee combiner model in seconds

Dataset	Year 2019	Year 2018	Year 2017	All-Data
CON	4.6	4.8	4	6.1
WAVG	75.5	75.3	44.8	151.71
AVG	6.3	7	5.39	7.46
FC	33	34.5	27	80.32
MED	5.1	5.4	4.7	8
MajVot	6.86	6.93	5.27	10.1
MIN	6.71	6.6	5.73	9.34
MAX	6.35	6.64	5.57	9.42

Source: Author

5.7. Summary

This chapter discussed the development of eight committee combiner methods used to combine the predictions for each of the single classifiers presented in Chapter 4 in order to enhance the accuracy of classification audit opinion modelling. It presented the capability of CON and FC, and each of the individual committee methods (AVG, WAVG, MED, MajVot, MIN, and MAX) as classification tools for auditing opinion. The comparatives analysis and discussion of the eight committee methods across the four tested datasets showed that all of the methods had acceptable performance based on correctly classifying audit opinion in terms of average accuracy rates, AUC, Type I and II Error rates, F-measure rates, sensitivity rates, specificity rates, and AURD, and there was no big gap between the performance evaluation results of each committee classifier across all four datasets. Likewise, the experimental results presented that CON outperformed the performance evaluation results for all nine single classifiers across all four datasets. The experimental result shown that CON has best ability to increase individual classifier accuracy camper to other committee combiner models. In particular, weighted average combiner outperformed DT, SVM, ANN, NBN, BEC, LR, LDA and K-NN over all four datasets. Because of this, all the committee machine methods can be utilised in real-life applications, supported by the performance evaluation results for different types of datasets.

In addition, WAVG and CON revealed superior ability in classifying the audit opinion correctly, outperforming the MIN, MAX, MajVot MED, average, and FC models, and they are balanced in distinguishing between audit opinions due to having fewer gaps between unqualified and qualified audit opinions. On the other hand, the MAX method had lower performance across the year 2018, 2019, and 2017 datasets compared to MIN, average, weight average, FC, MED, MajVot and CON. Conversely, MIN method outperformed the MAX, AVG, MED, FC, and MajVot combiners for All-Data, for which dataset majority vote had the worst performance. However, comparative analyses across all four datasets indicated that MIN and MAX methods had essentially similar evaluation result performance, due to the threshold enhancement of the MIN and MAX method processes.

Statistical significance testing presented that CON has the best performance to correctly classify audit opinion, and the ability to enhance auditing opinion model accuracy, but max model has poorer performance compared to committee models.

Chapter 6
Dynamic Modelling

6.1. Introduction

This chapter describes dataset preparation to fit with dynamic model to develop dynamic modelling performance for DPL-LSTM, NAR, and NARX to predict audit opinion in advance of one year. This is based on model changes and inputs to predict the audit opinion, which is basically an early warning system to see if the current performance of the company will end up with auditing problems for financial statements. Dynamic model training to predict audit opinion in year 5 (2019) uses actual datasets for the studied companies from the year 2018 to year 2015. In the next step, each model performance is evaluated using nine evaluation measurements, and these evaluation results are compared with those of the benchmark model (DPL-LSTM classifier), with the actual dataset for the year 2019, in order to examine dynamic modelling predictive capability for correct audit opinion. The evaluation experimental results for DPL-LSTM, NARX, and NAR are evaluated and compared with the benchmark model results to determine dynamic modelling ability to correctly predict audit opinion one year in advance, and which model has the best capability.

6.2. Dataset Preparation

Cluster method is used to make datasets suitable for dynamic modelling, grouping the set of data points into clusters by similarity, supposing that there is an element of resemblance between points of a same cluster, based on function distance measure. This thesis used the algorithm of fuzzy c-means clustering, which quickly and credibly clusters algorithms to elicit beneficial information from large datasets (Suganya and Shanthi, 2012). Fuzzy c-means algorithm is a data clustering method in which the dataset is grouped into number of clusters (M), whereby each point (company) in a dataset belonging to each M cluster has a high degree of membership to that group, and other points of the dataset located far from a cluster centre have a low degree of membership to that cluster. A single data point can have partial membership in two classes (Sreenivasarao and Vidyavathi, 2010). Cluster method results in a number of clusters, each of which has a set of points with varying distances from one another, and with relatively large spaces from other clusters (Suganya and Shanthi, 2012).

6.3. Dynamic Modelling Development and Experimental Results

This section illustrates the development of DPL-LSTM, NAR, and NARX forecasting performance for financial statement audit opinion one year in advance. In addition, it displays evaluation measurement results such as confusion matrix, F-measure, ROC curve, and reliability diagrams, to evaluate the testing models' performances individually, and it compares the evaluation results of each dynamic model with those of the benchmark classifier model.

6.3.1. Deep Learning LSTM (DPL-LSTM)

DPL is a novel dynamic algorithm model that attains good prediction dynamics for forecasting one or more-time steps ahead, and which can resultantly update the network state. DPL algorithm uses deep or multiple layer attributes to extract ingrained features in data with a loud level of representation of characteristic classes or attributes, via minimum levels feature. It can dramatically amplify detection precision, and detect massive amounts of construction in the data. To develop the simulation and computing platform for DPL forecasting of future time step values in the sequence, it is built using LSTM network training layers architecture. In a stratified classifier phase, the dataset is prepared by entering it into a classifier and processing it through the following sequence:

- LSTM cells as an input layer (there are 34 hidden sequence input layers, corresponding to the number of independent variables).
- LSTM layers of 34 units, with output mode specified as a sequence for LSTM layers. The reason for choosing sequence classification is due to recalling only the significant sides of the input sequence.
- One dropout layer at 0.2 probability of defining the next layer input elements to 0 in order for a network not to be sentient to a tiny group of neurons in the layer.
- Connecting all the neurons in a previous layer by a fully connected layer.
- The *softmax* layer normalises the fully connected layer output, and then creates the prediction possibility outcome for each class, consisting of positive figures that sum to one.
- The classification layer utilises the predictions from restoration by the *softmax* layer for each input to specify an input to one of mutually exclusive classes, and calculates the final error class.

After building the training layers, DPL training options are constructed by generating a set of DPL training options utilising Adam optimiser, employing a maximum number of 1000, with a minimum batch of 500 observations for each iteration, and a gradient threshold of 1. The longest sequence length is identified by each mini batch holding the same length for the longest sequence, and the dataset is recalled to hold the same length for the longest sequences. To assure that a dataset stays arranged by sequence length, every epoch data shuffle avoids ignoring the same data in every epoch. In training options, the execution environment is specified as *Auto* (if the GPU is available, training utilises it; otherwise, CPU is used). Training options end by specifying the training progress as plots to display training metrics at every iteration. After building the training LSTM network and options, the DPL network is created by training LSTM network with training options, and input dataset and target data using *trainNetwork*. To predict one step ahead and update the network state at each prediction from DPL classifier model, *predictAndUpdateState* function is applied for the input data with the output of *trainNetwork*. This output is the forecasting of future audit opinion of year 5 (year 2019).

The evaluation measurement results for testing DPL-LSTM performance shown in Table 6.1 indicate that DPL-LSTM has excellent capability to correctly forecast audit opinion in advance, with average accuracy of 96.8%, AUC of 97.2%, F-measurement of 95.9%, Type II Errors (qualified companies being identified as unqualified, false negative) of 0.7%, Type I Errors (unqualified companies being flagged as belonging to the qualified class, false positive) of 7.7%, Brier score of 1.48, true positive rate of 99.3%, true negative rate of 92.3%, and AURD of 6%. Likewise, DPL-LSTM has an ROC curve close to the corner of 1, and a relatively good reliability diagram indicated by proximity to the diagonal line (Figures 6.1-6.2).

In addition, comparison between the DPL-LSTM and benchmark model evaluation results (Table 6.1) shows that DPL-LSTM achieved results near to the benchmark in terms of fewer gaps between average accuracy (0.5%) and AUC (2.8%) rates, Brier scores (0.8%), AURD (0.7%), Type I Error (2.3%), Type II Error (0.5%), and F-measurement (0.87%). This indicates DPL-LSTM's superior ability to forecast future time series and classify audit opinion correctly.

Table 6.1: Evaluation test results of DPL-LSTM model with benchmark model

	DPL-LSTM	**Benchmark**
Average accuracy	96.8%	97.3%
Type II Error	0.7%	1.2%
Type I Error	7.7%	5.4%
Specificity	92.3%	94.6%
Sensitivity	99.3%	98.8%
F-measurement	95.9%	96.8%
AUC	97.2%	98.9%
Brier score	1.48%	0.86%
AURD	6%	6.7%

Source: Author

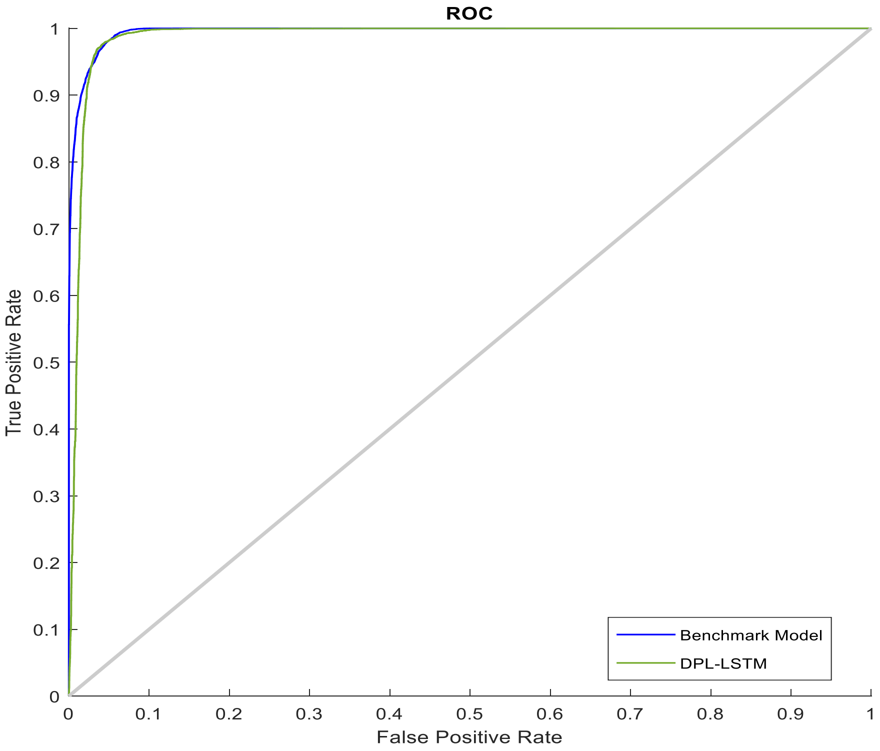


Figure 6.1: ROC curves for DPL-LSTM model with benchmark model

Source: Author

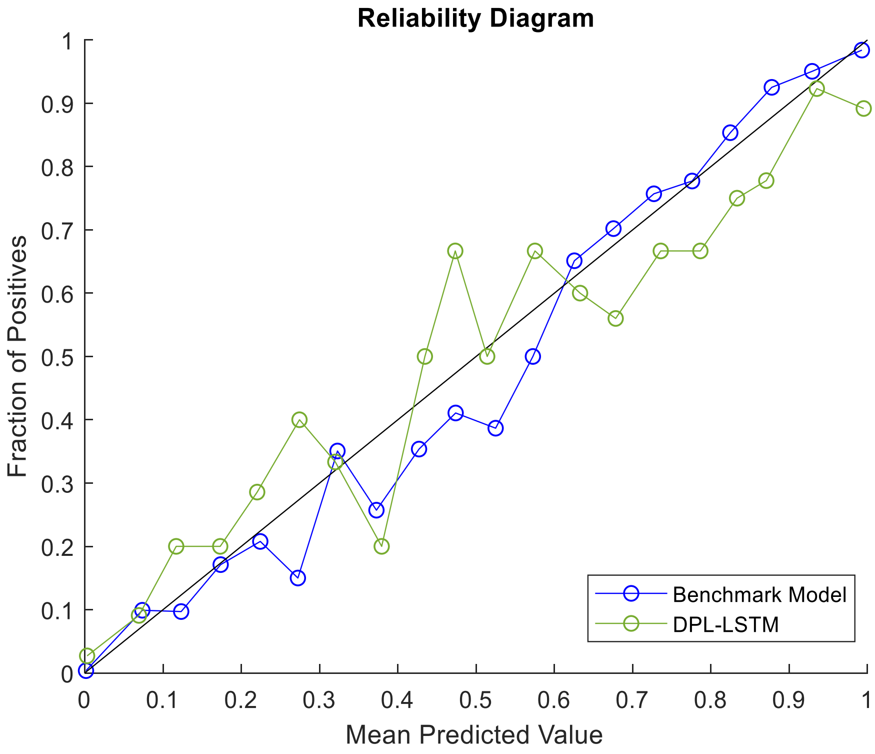


Figure 6.2: Reliability diagrams for DPL-LSTM model and benchmark model

Source: Author

6.3.2. Nonlinear Autoregressive Exogenous (NARX) Neural Network

All dynamic network models have been augmented and are focused networks, NAR has only an input layer or feedforward networks as, not as NARX which has exogenous input with feedback. NARX is an RNN with feedback connections enclosing several layers of a network, with the ability to efficiently lay out the time series with feedback connections enclosing several layers of a network. Due to this, NARX is utilised in time series layout. NARX network inputs have two classes that influence the series of interest: past values of an interest output series, and previous values of driving exogenous inputs series. The NARX NN formula is shown in equation 6.1.

$$Y(T+1) = f\ (Y(T), Y(T-1), \cdots, Y(T-N_Y), X(T), X(T-1), \cdots, X(T-N_X)) \qquad 6.1$$

where T represents the time series, N is the number of exogenously designated input and output delay, $Y(T+1)$ is the NARX output, $(Y(T), Y(T-1), \cdots, Y(T-N_Y))$ refers to the previous values of an exogenous output series of interest, $(X(T), X(T-1), \cdots, X(T-N_X))$ comprises variables driving exogenous input series, and f maps the non-linear NN function.

For training the NARX model, the input time series (X(T)), feedback (output) time series (Y(T)), and setup division function are first built using *dividerand*, with division mode type as *time*, to divide the dataset into three partitions for training, testing, and validation. For the

NARX NN structure (Figure 6.3), one hidden layer was utilised, and its size was selected as 10, based on consideration of model complexity; this size better fits the input in relation to the number of input and feedback delays (1:2). In addition, the training function used to update the input weight and bias input variables are specified to obtain optimal performance output value when training the NARX, using several types of training function (*trainscg*, *trainbr*, and *trainlm*). For this case, the default *trainlm* (Levenberg-Marquardt backpropagation) is used, which means the network training function updates bias and weight values based on Levenberg-Marquardt optimisation. Therefore, the network is structured and trained in open-loop form, as displayed in Figure 6.3. The reason for using open-loop rather than closed-loop is due to the latter being less efficient, and the former authorising the feeding of a training network with correct previous outputs (feedback) in order to produce the correct current output values.

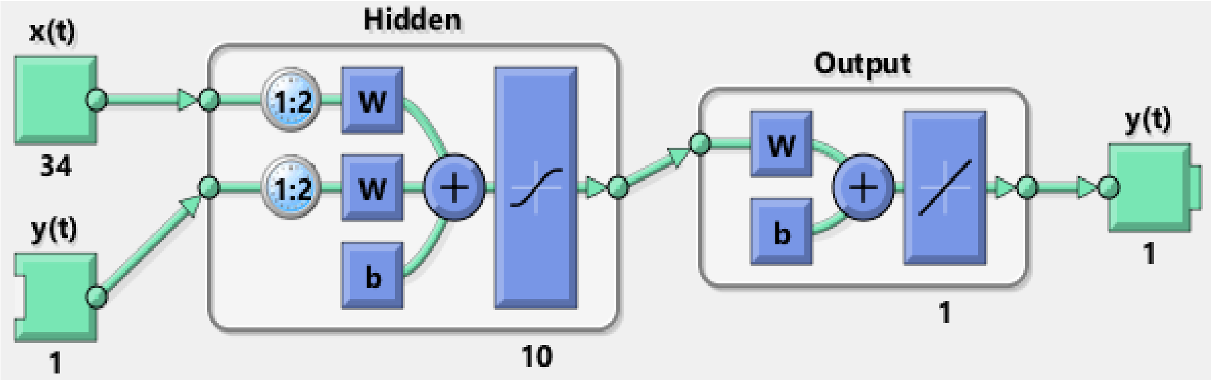


Figure 6.3: NARX view command

Source: Author

All of the training is done in series-parallel configuration inclusive of training, testing, and validation phases. A workflow to fully generate an open-loop training network in series-parallel configuration is used, then open-loop form is converted to closed-loop for multi-step-ahead prediction. For multistep prediction, it is beneficial to simulate the NARX network in open-loop (series-parallel configuration), as the output data is known, after which the training of the NARX network and final cases can be converted from open-loop to closed-loop mode, to make multi-step-ahead prediction with only external inputs (time series) provided. Figure 6.4 shows how to transform NARX network from open- loop to closed-loop form, and run a five time steps ahead prediction.

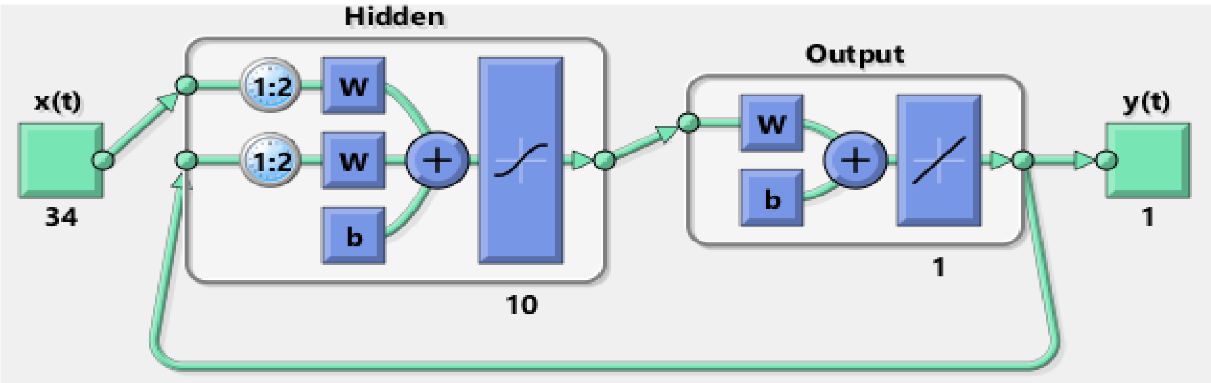


Figure 6.4: NARX closed-loop

Source: Author

Step-ahead prediction network is used to obtain predicted timestep values early. A NARX network can been obtained to return predicted values output Y(T+1) timestep early to utilise these predictions as direct input for the next step through *removedelay* (remove one delay), so the minimal tap delay is now 0:1 instead of 1:2 (Figure 6.5). Then, a new NARX network can be returned to do the same predictions output as an initial network, but output predictions are moved one timestep.

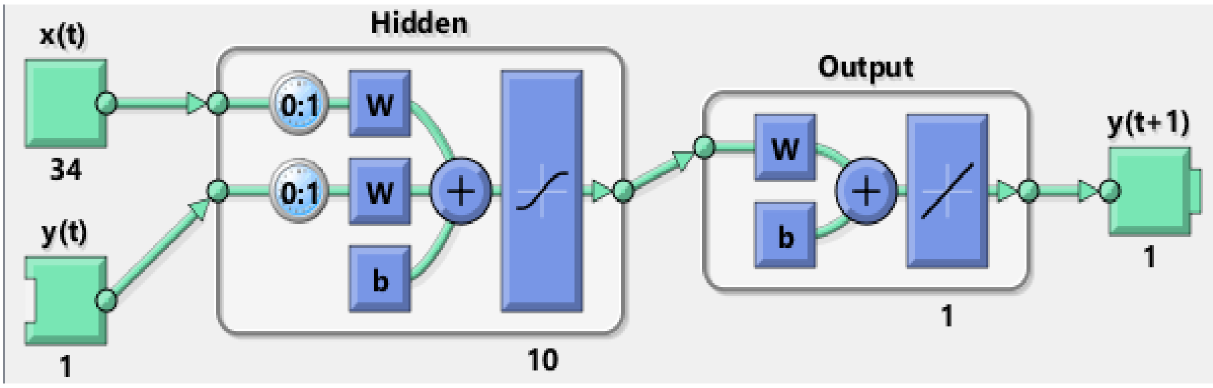


Figure 6.5: NARX predicted one step ahead

Source: Author

The evaluation results shown in Table 6.2 indicate that NARX has good ability to forecast audit opinion in advance, with 91% average accuracy, 95.9% AUC, 5.8% Brier score, 87.8% specificity, 92.9% sensitivity, 12.2% Type I Error, 7.1% Type II Error, 90.5% F-measure, and 19.69% AURD. However, NARX dynamic results for year 5 step ahead forecasting has lower performance compared to the DPL-LSTM benchmark model, due to the gap between the evaluation results: 6.3% for average accuracy, 5.9% for Type II Error, 6.8% for Type I Error, and 12.99% for AURD. Additionally, NARX dynamic has poorer ROC curve compared to the benchmark model, with a higher false positive rate and lower true positive rate. This is reflected in poor alignment with the diagonal for the reliability diagrams (Figures 6.6-6.7), where NARX dynamic has a greater area under diagram compared to the actual dataset for the year 2019.

Table 6.2: Evaluation test results of NARX model with benchmark model

	NARX	**Benchmark**
Average accuracy	91%	97.3%
Type II Error	7.1%	1.2%
Type I Error	12.2%	5.4%
Specificity	87.8%	94.6%
Sensitivity	92.9%	98.8%
AUC	95.9%	98.9%
F-measurement	90.5%	96.8%
Brier score	5.8%	0.86%
AURD	19.69%	6.7%

Source: Author

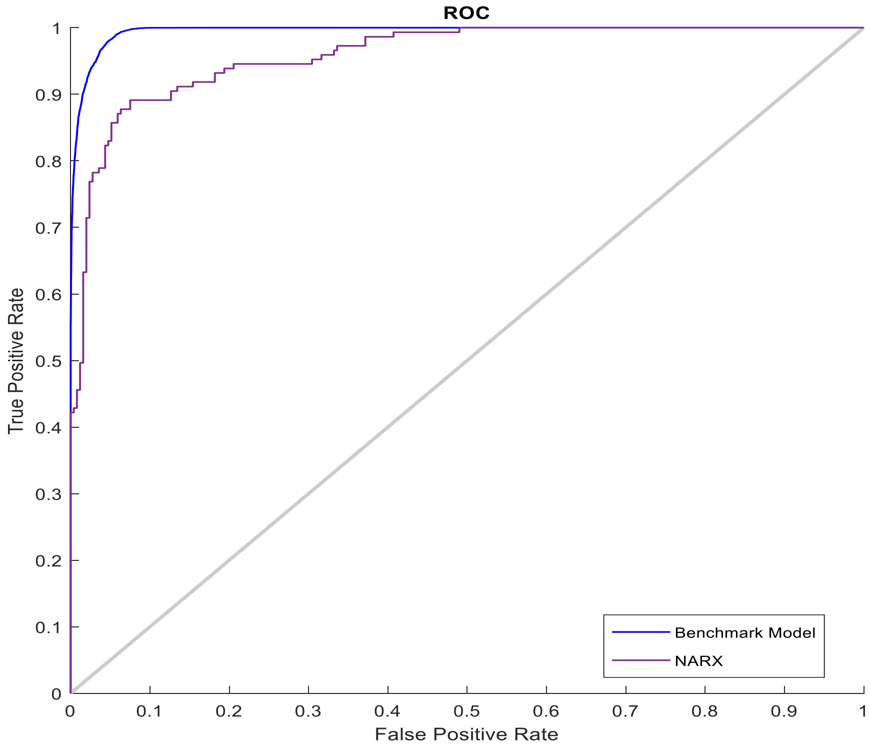


Figure 6.6: ROC curves for NARX model with benchmark model

Source: Author

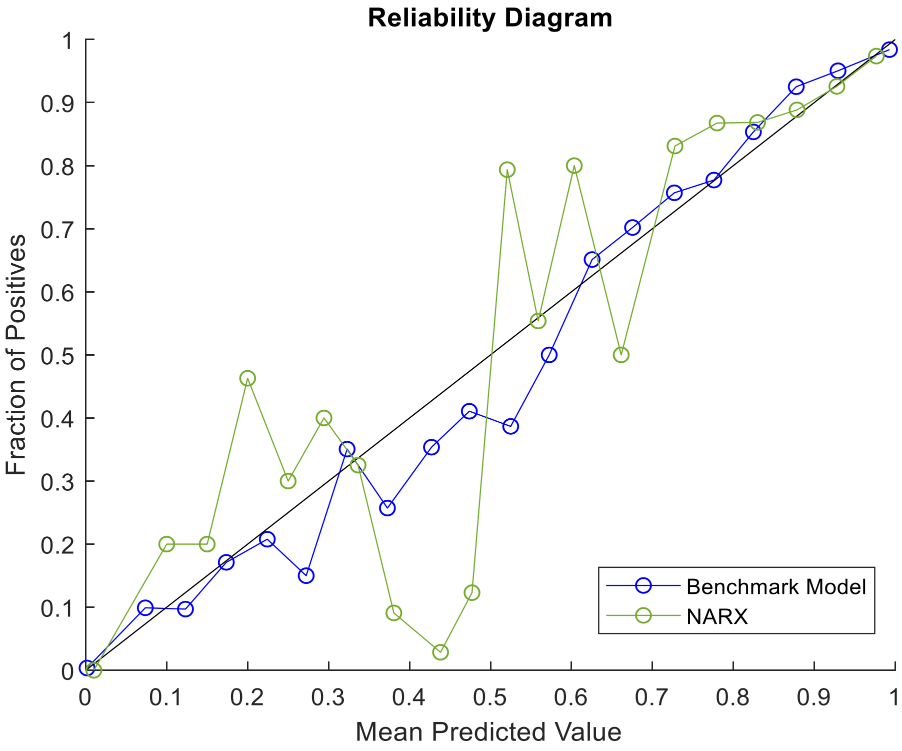


Figure 6.7: Reliability diagram for NARX model and benchmark model

Source: Author

6.3.3. Nonlinear autoregressive (NAR) neural network

NAR forecasts the time series based on previous values of a particular series. The NAR topology is not as complex as that of NARX. The difference between NAR and NARX is in the use of outputs. NAR feeds inputs only to predict the next step, while NARX uses current inputs and outputs to predict the next step output. The NAR is an RNN with feedback links endorsing layers of a network, so a prediction final output is based on a previous value output of time series. Equation 6.2 presents how the NAR network forecasts the value of the data time series Y at T (time series), Y(T) employing $\mathrm{N_Y}$ (the previous value output of time series). The objective of NAR network training is to approximate a function f through enhanced weight of the NAR network and reduced neuron bias.

$$Y(T+1) = f\ (Y(T), Y(T-1), \cdots, Y(T-N_Y)) \qquad 6.2$$

where T represents the time series, N is the number of designated output delays, $Y(T+1)$ is the NAR output, $(Y(T), Y(T-1), \cdots, Y(T-N_Y))$ refers to previous values of output series, and f maps the non-linear NN function.

The NAR network design needs to specify the feedback data time series Y(T) and setup division function as *dividerand* with division mode type in terms of the time to divide the feedback data into three partitions of target timesteps for:

- Training: during NAR network training, error is used to modify the NAR network.
- Testing dataset: this does not affect network training, and is used as an independent evaluation of network performance during and after training.
- Validation: employed to a calculate NAR network popularisation and to stop network training when popularisation ceases enhancing.

As shown in Figure 6.8, the NAR network structure uses one hidden layer whose size (10) was identified based on model complexity, providing better performance to fit the feedback dataset in best path with number of feedback delays at 1:2. Subsequently, the specified training function is utilised to update the input weight and bias input variables in order to obtain an optimal performance output value during training a NAR. While *trainscg*, *trainbr*, and *trainlm* training functions are available, the default *trainlm* (Levenberg-Marquardt backpropagation) is used to train the NAR network, thus the network training function that updates bias and weight values is based on Levenberg-Marquardt optimisation. A NAR network is structured and trained in open-loop form utilising a real target dataset as an echo, to make sure of quality relative to the correct number in the training process. Figure 6.8 summarises the NAR network training view command.

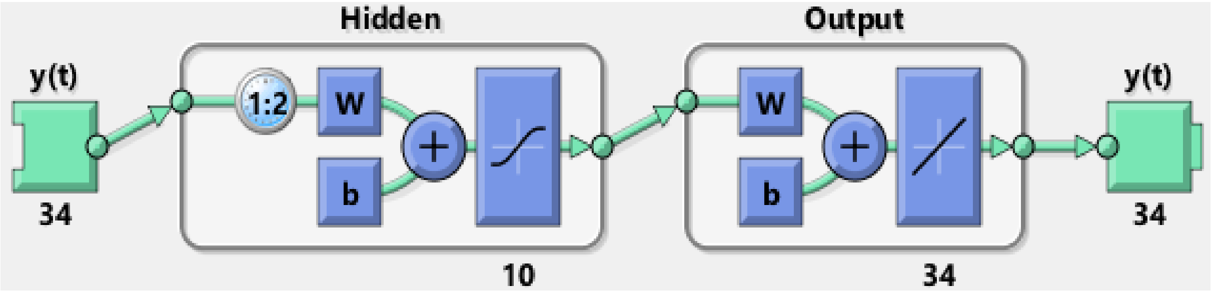


Figure 6.8: NAR topology

Source: Author

After building up and training the NAR network, it is transformed from open- to closed-loop form, and the forecasting values are utilised to provide new NAR network inputs. Figure 6.9 shows how the NAR network is transformed to closed-loop form.

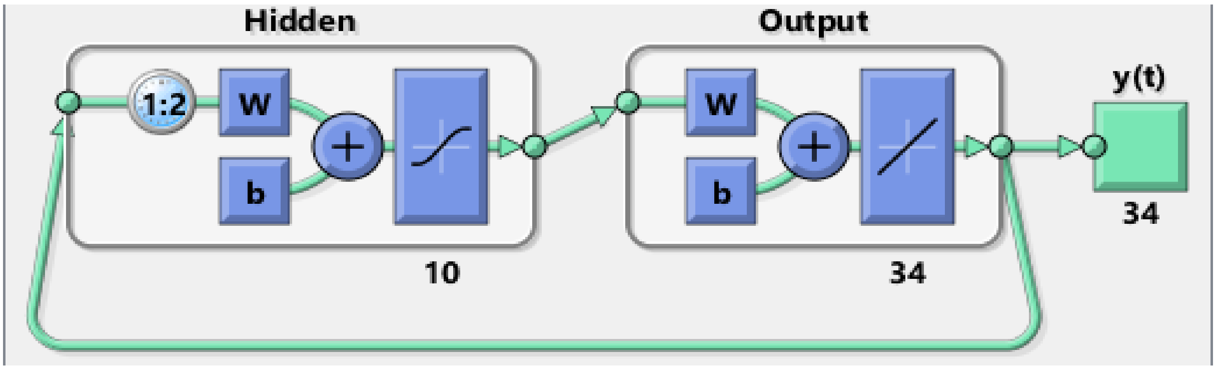


Figure 6.9: NAR topology with closed-loop

Source: Author

The final prediction uses step-ahead prediction network to obtain predicted values, whereby the NAR network can return forecasting value timestep-early outputs, and these predictions can be deployed as direct inputs for a next step through *removedelay* (remove one delay), so that its minimal tap delay changes from 1:2 to 0:1, as shown in Figure 6.10. Then, a new NAR network returns the same output predictions as an initial network, but output predictions are moved left one timestep.

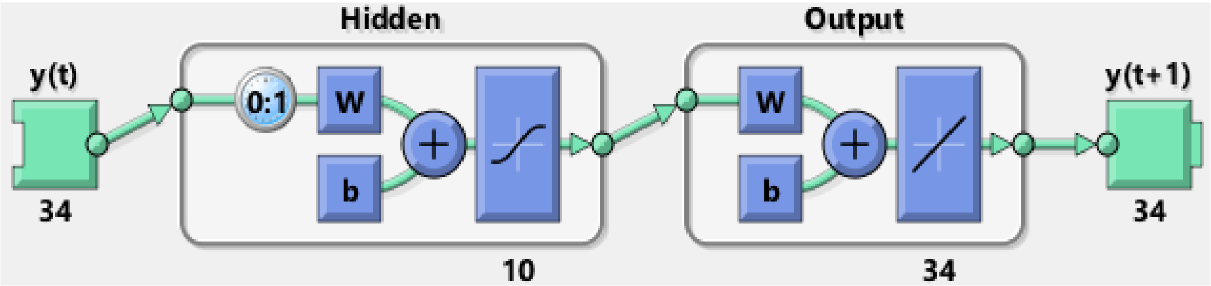


Figure 6.10: NAR topology with one step ahead prediction

Source: Author

Evaluation results of NAR dynamic (Table 6.3) illustrate good performance ability to produce correct audit opinion in advance, with 95% average accuracy rate, 96.3% AUC, 97% sensitivity, 93% specificity, low rates of false positive (7%) and false negative (3%), and area under reliability of 17.31%. NAR dynamic shows few differences with DPL-LSTM (benchmark) in terms of performance evaluation results for the year 2019 dataset: 1.6% between Type I Error rates, 1.8% between Type II Error rates, 1.6% between specificity rates, 1.8% between sensitivity rates, 2.3% between average accuracy rates, 2.6% between AUC percentages, 1.54% between Brier scores, and 1.7% between F-measurements.

On the other hand, the ROC curves displayed in Figure 6.11 illustrate that NAR dynamic has worse performance than DPL-LSTM model, and the reliability diagrams in Figure 6.12 show that NAR dynamic is not as close to the diagonal line as the reliability diagram for DPL-LSTM.

Table 6.3: Evaluation test results of NAR model with benchmark model

	NAR	**Benchmark**
Average accuracy	95%	97.3%
Type II Error	3%	1.2%
Type I Error	7%	5.4%
Specificity	93%	94.6%
Sensitivity	97%	98.8%
AUC	96.3%	98.9%
F-measurement	95.1%	96.8%
Brier score	2.4%	0.86%
AURD	17.31%	6.7%

Source: Author

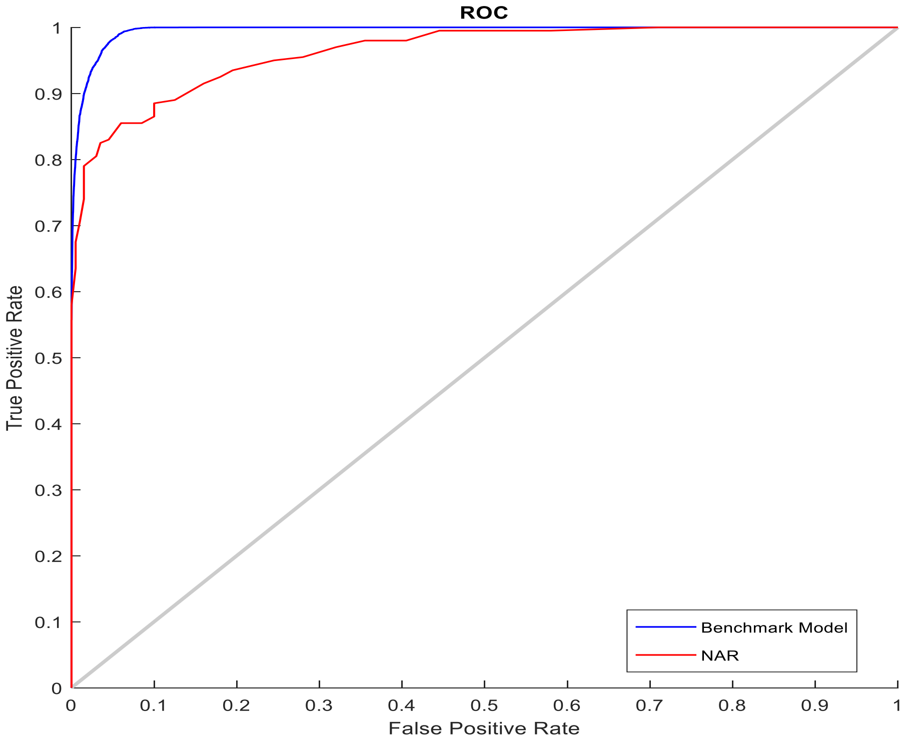


Figure 6.11: ROC curves for NAR model with benchmark model

Source: Author

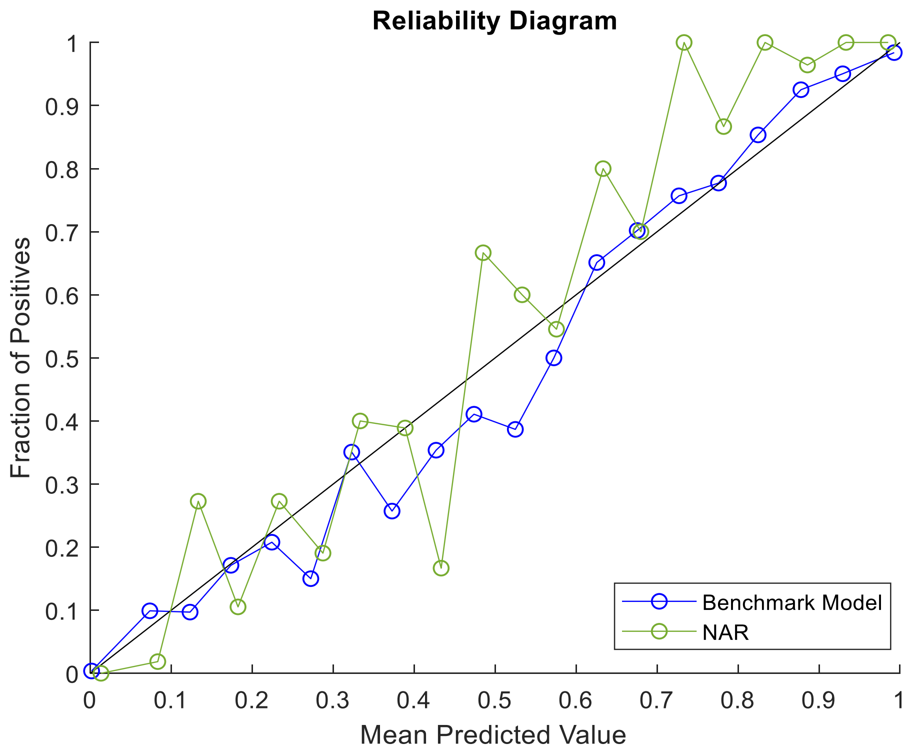


Figure 6.12: Reliability diagrams for NAR model with benchmark model

Source: Author

6.4. Comparative Analysis and Discussion

This section comparatively analyses the experimental results for DPL-LSTM, NAR, and NARX dynamic modelling and the benchmark (DPL-LSTM classification model tested on the original dataset for year 5 (2019) dataset).

Table 6.4: Comparing evaluation test results of dynamic models with benchmark model

	DPL-LSTM	**NARX**	**NAR**	**DPL-LSTM (Benchmark)**
Average accuracy	96.8%	91%	95%	97.3%
Type II Error	0.7%	7.1%	3%	1.2%
Type I Error	7.7%	12.2%	7%	5.4%
Specificity	92.3%	87.8%	93%	94.6%
Sensitivity	99.3%	92.9%	97%	98.8%
AUC	97.2%	95.9%	96.3%	98.9%
F-measurement	95.9%	90.5%	95.1%	96.8%
Brier score	1.48%	5.8%	2.4%	0.86%
AURD	6%	19.69%	17.31%	6.7%

Source: Author

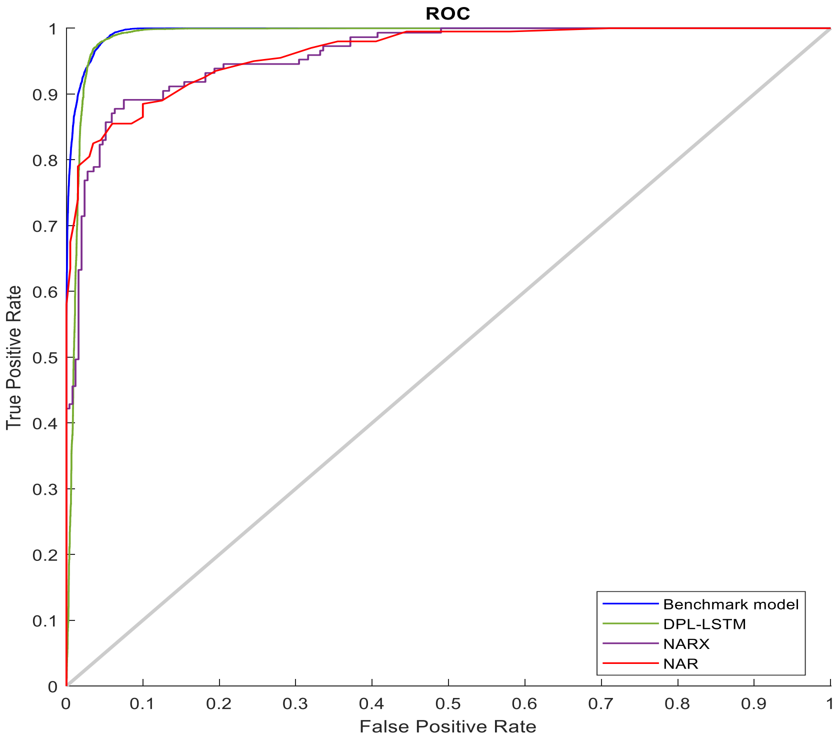


Figure 6.13: ROC curves for three models with benchmark model

Source: Author

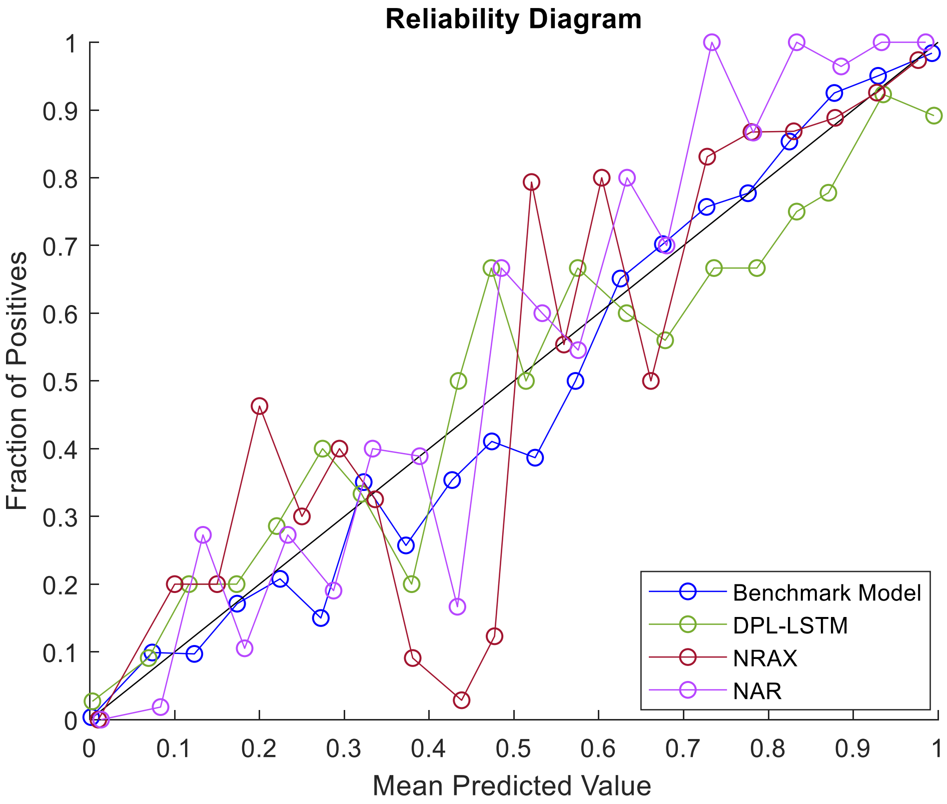


Figure 6.14: Reliability diagrams for three models with benchmark model

Source: Author

The evaluation measurement performance outputs of DPL-LSTM, NAR, and NARX show that these dynamic models have acceptable performance to predict future audit opinion correctly, with average accuracy rates above 90.9%, AUC percentages above 95.8%, specificity rates above 87.7%, F-measurement above 90.4%, sensitivity rates above 92.8%, Brier scores below 5.9%, and less than 19.7% AURD, 12.3% Type I Error rates, and 7.2% Type II Error rates. All three dynamic models have good performance ability to detect predictive year 5 (2019) auditing opinion correctly by reaching results near to performance evaluation results of the DPL-LSTM (benchmark) model, and there was no big gap between performance evaluation results for the three dynamic models and the evaluation results of benchmark model.

Overall, the comparative analysis shown in Table 6.4 for the evaluation measurement results show that the DPL-LSTM dynamic model has more powerful ability to forecast audit opinion in advance than NAR and NARX models, achieving the best evaluation results of 96.8% average accuracy, 7.7% Type I Error, 0.7% Type II Error, 99.3% sensitivity, 97.2% AUC, 6% AURD, 1.48% Brier score, 95.9% F-measurement. Additionally, DPL-LSTM dynamic had the nearest performance evaluation results with those of the benchmark model. Figure 6.13 shows that DPL-LSTM dynamic has the closest ROC curve to the benchmark model, indicating superior capacity to distinguish between classes and to forecast future audit opinion, and it obtained the best reliability diagram (Figure 6.14), closer to the diagonal line than the reliability diagrams for the NAR and NARX models.

The above analysis affirms that the DPL-LSTM dynamic model has superior forecasting ability for advanced audit opinion compared to the NAR and NARX models. This is attributable to the DL classification training using multiple neural network layers, enabling features extraction and transformation for utilisation as inputs into the next hidden layers, and permitting more simple performance of interaction with input data. Also, the DPL time series forecasting development training used the LSTM layer for sequence classification, enabling the creation of various forecasts for each signal time step of series data. Training DPL is extremely computationally intensive, and can training models can generally by accelerated utilising high-performance execution environments (*auto).*

Having acknowledged the superiority of DPL-LSTM dynamic model, Table 6.4 indicates that NAR performance model has better ability to predict audit opinion than NARX model, as indicated in its better performance evaluation results: 95% average accuracy, 96.3% AUC, 95.1% F-measurement, 93% specificity, 97% sensitivity, 2.4% Brier score, 7% Type I Error, 3% Type II Error, and 17.31% AURD. NAR's performance evaluation results are nearer to the benchmark classifier evaluation results than NARX's dynamic results. For example, NAR had smaller gaps with the benchmark models' average accuracy (2.3%), AUC (2.6%), specificity (1.6%), and sensitivity (1.8%). Likewise, Figure 6.13 shows that the NAR model had a better ROC curve than NARX, and Figure 6.14 shows that is reliability diagram was closer to the diagonal line. In addition, NAR is the most balanced dynamic model, with lower difference between Type I and Type II Error (4%) than the DPL-LSTM and NARX models. This means that NAR has excellent performance in forecasting future audit opinion compared to NARX, and it is more effective in distinguishing audit opinion than both the NARX and DPL-LSTM models.

Moreover, NAR dynamic model outperformed NARX model in terms of the former's average accuracy (91%), specificity (87.8%), sensitivity (92.9%), F-measurement (90.5%), AUC (95.9%), Brier score (5.8%), AURD (19.69%), Type I Error (12.2%), and Type II Error (7.1%). In sum, the NARX dynamic model is ineffective in forecasting audit opinion based reached on its performance evaluation results being further from those of the benchmark model than those of the other tested models. Moreover, as shown in Figure 6.13, NARX dynamic achieved worse ROC curves than DPL-LSTM and NAR, because it detected higher false positive rates and lower true positive rates. In addition, Figure 6.14 presents that NARX has the worst reliability diagram shape (and thus poorer ability to forecast), due to the greatest variance from the diagonal line. Consequently, NARX had the worst dynamic performance, and it is the worst model for forecasting time series compared to the DPL-LSTM and NAR models.

6.5. Dynamic Modelling Training Time

This section presents and discusses the computational time for each dynamic modelling process. Feature extraction is the first step in assessing dynamic modelling, selecting all significant ratios by using SPSS version on a 25 GB RAM. This step takes around 40 seconds for each dataset (t-1, t-2, t-3, t-4). Procedure clustering modelling is then built and performed to fit all four datasets for dynamic modelling through using MATLAB 2019a version on an 8 GB RAM personal computer with 3.4 GHz, Intel CORE i7, and Microsoft Windows 10 operating system. This process takes a lot of time, at 681 seconds. In general, the factors (number of hidden layers and dataset size) impact on the time taken for model evaluation. As seen from Table 6.5, the benchmark model evaluation had a higher computational time (134 seconds), because it needs a longer time to train 6000 entries, while DPL-LSTM, NAR, and NARX dynamic models were only trained on 400 data points. Table 6.5 presents the ordering of the dynamic algorithms from slowest to fastest according to training time process. It can be seen that DPL-LSTM evaluation took the longest, at 260.18 seconds, making it the slowest training model; NAR model had the lowest training time at 25.37 seconds, indicating that this is the fastest model for training data. The reason for the egregiously longer training time for DPL-LSTM is that its construction is more complex because of the greater number of hidden layers (NAR and NARX used just one hidden layer). Models need significantly more time to train with multiple hidden layers, which substantially increases the number of epochs necessary to enhance model training performance. For instance, NAR and NARX took nine epochs, while DPL-LTSM required 200.

Table 6.5: Dynamic and benchmark model training times

Dataset	Training time (seconds)
DPL-LSTM	260.18
NARX	26
NAR	25.37

Source: Author

6.6. Summary

This chapter tested the development and ability of dynamic modelling based on DPL-LSTM, NAR, and NARX as advance forecasting tools for auditing opinion has been tested and investigated. The evaluation experimental results performance indicated that output of DPL-LSTM, NAR, and NARX models showed acceptable performance for correct predictive audit opinion in advance, with above 90% average accuracy rates, AUC percentages, and F-measurement; above 87.7% sensitivity and specificity rates; and lower than 19.8% Brier scores, AURD, and Type I and Type II Error rates. Likewise, the three dynamic models had good performance ability to detect year 5 (2019) audit opinion correctly, compared with DPL-LSTM model results for the actual year 2019 audit opinion dataset, though there was no major gap between the performance evaluation results of the three dynamic models and benchmark model.

DPL-LSTM dynamic model was found to have superior capability in productive audit opinion in advance, outperforming the NAR and NARX models, and it is a model able to predict future audit opinion correctly with higher evaluation performance results near to those of the benchmark model. The NAR dynamic model had the next-best ability to predict audit opinion in terms of achieving better results compared to the NARX performance results, and it is a balanced model, able to distinguish between audit opinions through detecting lower difference between the sensitivity rate and true specificity rate compared to the difference values achieved by DPL-LSTM and NARX. However, NARX had poor performance to predict audit opinion, with lower performance evaluation results and a bad ROC curve and average accuracy rate, and it was greatly outperformed by the benchmark model.

Chapter 7
Conclusion and Future Work

7.1. Main Study Outcomes

This thesis aimed to explore potential improvements in the ability of single classifiers and committee combiner models and dynamic modelling as classification and advanced prediction tools for auditing opinion. These aims were reached by constructing several individual classifiers, committee combiner models, and dynamic modelling. Comprehensive comparison of several audit opinion models, starting from simple individual base classifier models to very complex models based on different committee combiner models, achieved the primary objective of this research. Comprehensive comparison of three dynamic modelling advanced predictive tools was also undertaken. This study makes numerous contributions in relation to the research design and empirical results.

7.2. Research Framework

At the framework research design stage, this thesis concentrated on the major phases in the experimental design of auditing opinion model in order to address this identified gap in the literature. It was necessary to carefully consider numerous issues, integrated in several stages:

- Addressing the different sizes of public datasets. This thesis collected four different datasets that were utilised to validate individual classifiers, and the further datasets that were utilised to validate dynamic modelling.
- Using data pre-processing techniques, including the three steps of data imputation, data normalisation, and features selection processing.
- Data splitting after data pre-processing, splitting datasets, which were used to structure and assess the individual classifiers, and committee combiner, into testing and training sets.
- Clustering another four dataset which were used to structure and assess the model dynamic modelling.
- Improving and training nine individual classifiers and three dynamic modelling techniques.
- Combining nine individual classifiers with six improvement traditional committee modelling with two new committee modelling methods.

- Evaluating model performance using seven popular measurement parameters with two evaluation parameters (AURD and Brier score) to measure several aspects of ability for each classifier classification and advanced prediction capability.
- Statistical significance testing.

7.3. Empirical Contributions

The testing results manifest the main empirical contributions of this thesis, as summarised below.

7.3.1. Improved Simulation and Computing Platform

This thesis presents an improved simulation and computing platform for nine individual classifiers that were implemented and tested for the four datasets. The empirical comparison between evaluation measurement performance results and significance testing results for the nine classifiers illustrated that over the four datasets the novel classifier (Deep learning model) had superior classifier performance ability to classify audit opinion correctly compared to all other classifiers, obtaining the highest values for all nine evaluation measurement parameters. For example, for the year 2018 dataset, DPL achieved 97.6% average accuracy, 99.3% AUC, 94.7% specificity, 99.6% sensitivity, 0.4% Type I Error, 5.3% Type II Error, 97.2% F-measure, 1.5% Brier score, and 6.1% AURD. Likewise, DPL model was a balanced model in distinguishing between audit opinion correctly, with the best ability to classify audit opinion correctly, even with large data sizes – other models were fundamentally impaired with larger data sizes. In addition, statistical significance test results approved that DPL has best ability to classify audit opinion, achieving the best Freidman average rank, far below the two cut-lines at α = 0.05 and α = 0.1. After DPL, over the four datasets, the BEC and DTs had the best performance evaluation results and good statistical significance scores compared to the other classifiers' performance results.

7.3.2. Improved Performance Accuracy

This research improved performance accuracy for audit opinion model through experiments based on committee combiner classifiers (combining nine individual classifiers together), utilising traditional combiners (AVG, WAVG, MED, MajVot, MAX, and MIN) and new committee methods such as CON and FC. The performance results clearly presented that committee methods often yield the best results compared to most individual classification models (which are the constituent models of combined committee models). Among traditional committee methods, WAVG had the best ability to enhance the final results, which was attributable to its axiomatic simplicity in determining weight based on the power of classifier performance. Likewise, all the weights of every classifier were commensurate with the certitude of respective outcomes (in terms of how outcomes are relative to actual targets). On the other hand, statistical testing and comparison of experimental evaluation results over the four datasets determined that CON had better performance than traditional committee modelling, FC and nine individual classifiers to improve audit opinion model accuracy and correct classification of audit opinion. This was demonstrated in the best statistical testing results (i.e., the best Friedman average rank, lower than the two cut-lines), and higher evaluation measurement results (e.g. achieved above 98.1% ROC, above 95% average accuracy, and Brier scores lower 2.4%).

7.3.3. Applying Three Dynamic Modelling Techniques

The third contribution was testing the improvement of three dynamic modelling approaches (DPL-LSTM, NAR, and NARX), applied with clustering dataset, whereby outcomes were comparatively analysed with a benchmark model. DPL-LSTM classifier had the best capacity to predict audit opinion in advance correctly, through obtaining all higher performance results, and the closest results to the benchmark model performance compared to the other dynamic modelling techniques: DPL-STML 96.8% average accuracy rate, 1.48% Brier score; NARX 91% average accuracy rate, 5.8% Brier score; NAR 95% average accuracy rate, 2.4% Brier score. Additionally, there was a lower difference (0.5%) between the accuracy rates for DPL-LSTM and the benchmark model.

7.4. Limitations

As with all research studies, a number of limitations were faced in this thesis, which can be classified into literature review and methodological issues.

7.4.1. Literature

Literature searching and collection suffered from the relative dearth of studies investigating DM models for auditing and accounting. While this was part of the rationale for undertaking this study, the lack of initial applications of machine learning models in the auditing area means that there was a lack of effectiveness studies to form a basis for extending research in this field. Very few studies were identified that explored and analysed classification DM tools to understand impacts on auditing decision making and research related to the development of audit opinion models. In addition, no previous studies were found that investigated, explored, used, or improved DM dynamic modelling in auditing in terms of advanced prediction tools. Indeed, previous studies had not even identified the need for research on dynamic modelling, as most still concentrated on improving and investigating classification models, thus they called for more research on individual classifiers and committee methods.

7.4.2. Methodology

At the methodological level, several limitations were faced in relation to carrying out the framework research design of proposed auditing opinion model.

- Collecting datasets was hampered by the lack of companies publishing annual reports containing audit opinions, due to confidentiality and sensitivity.
- Datasets had to be balanced, as differences between the size of qualified and unqualified data affects performance evaluation for models.
- Finding a suitable feature selection method was troublesome, as this has a major impact on the accuracy performance for the DM model.
- Improvements at the committee modelling stage faced two limitations related to processor time needed to adjust each committee modelling parameter to fit the data, and to train and evaluating all individual classifiers.
- Datasets had to been fitted in dynamic modelling and finding the suitable model to make all four datasets fit for dynamic modelling.

7.5. Directions for Future Work

- Improve and test other dynamic modelling techniques and determine how these models can predict audit opinion in advance. Compare between these model performance results with dynamic models such as those tested in this thesis.
- Explore the ability of machine learning tools to classify and predict auditing opinion by comparing performance evaluation results for the models from dataset for companies applying GAAP with their counterpart company datasets using IFRS, in order to see if the type of the accounting standards followed by companies affect DM model performance.
- Explore and apply several data pre-processing techniques for a dataset, for example, other feature selection methods or data-filtering techniques, and accordingly define how these techniques can be reflect on DM model's performance results. For instance, in the filtering method, filtering-condensing method could remove outlier entries, avoiding non-informative entries negatively affecting the training process.
- This thesis compared consensus method with traditional committee modelling (AVG, WAVG, MED, MajVot, MAX and MIN), and concluded that the accuracy obtained with further iterations in consensus method is restricted. Therefore, further preliminary tests can investigate and analyse the impact of such other individual models that not used in this thesis on the accuracy reached via consensus, which would significantly develop the theoretical evidence on how much performance can be obtained.

References

Abdul Rahim, N. F., Jaafar, A. R., Syamsuddin, J. and Sarkawi, M. N. (2017) ‘Internal control system and hazard identification of operational risk in Malaysian conventional banking’, *International Journal of Supply Chain Management*, 6(2), pp. 215-227.

Abellán, J. and Castellano, J. (2017) ‘A comparative study on base classifiers in ensemble methods for credit scoring’, *Expert Systems with Applications*, 73, pp. 1-10.

Acuna, E. and Rodriguez, C. (2004) ‘The treatment of missing values and its effect on classifier accuracy’, in Banks, D., McMorris, F. R., Arabie P. and Gaul W. (eds.) *Classification, clustering, and data mining applications*. Berlin: Springer, pp. 639-647.

Ala’raj, M. and Abbod, M. F. (2016) ‘Classifiers consensus system approach for credit scoring’, *Knowledge-Based Systems*, 104(0950-7051), pp. 89-105.

Alasadi, S. and Bhaya, W. (2017) ‘Review of data preprocessing techniques in data mining’, *Journal of Engineering and Applied Sciences*, 12(16), pp. 4102-4107.

Albashrawi, M. (2016) ‘Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015’, *Journal of Data Science*, 14(3), pp. 553-569.

Alghofaili, Y., Albattah, A. and Rassam, M. (2020) ‘A financial fraud detection model based on LSTM deep learning technique’, *Journal of Applied Security Research*, 15(4), pp. 498-516.

Al-Hiyari, A., Said, N. A. and Hattab, E. (2019) ‘Factors that influence the use of computer assisted audit techniques (CAATs) by internal auditors in Jordan’, *Academy of Accounting and Financial Studies Journal*, 23(3), pp. 1-15.

Alhnaity, B. and Abbod, M. (2020) ‘A new hybrid financial time series prediction model’, *Engineering Applications of Artificial Intelligence*, 95, pp. 1-14.

Alles, M. and Gray, G. L. (2016) ‘Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors’, *International Journal of Accounting Information Systems*, 22, pp. 44-59.

Amani, F. A. and Fadlalla, A. M. (2017) ‘Data mining applications in accounting: A review of the literature and organizing framework’, *International Journal of Accounting Information Systems*, 24, pp. 32-58.

Appelbaum, D. (2016) ‘Securing Big Data provenance for auditors: The Big Data provenance black box as reliable evidence’, *Journal of Emerging Technologies in Accounting*, 13(1), pp. 17-36.

Arens, A. A., Elder, R. J., Beasley, M. S. and Hogan, C. E. (2020) *Auditing and assurance services.* 17 edn. Harlow: Pearson.

Aryal, A., Liao, Y., Nattuthurai, P. and Li, B. (2018) ‘The emerging big data analytics and IoT in supply chain management: A systematic review’, *Supply Chain Management*, 25(2), pp. 141-156*.*

Ashraf, M., Michas, P. N. and Russomanno, D. (2020) ‘The impact of audit committee information technology expertise on the reliability and timeliness of financial reporting’, *The Accounting Review*, 95(5), pp. 23-56.

Assel, M., Sjoberg, D. and Vickers, A. (2017) ‘The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models’, *Diagnostic and Prognostic Research*, 1, 19. Available at: https://diagnprognres.biomedcentral.com/track/pdf/10.1186/s41512-017-0020-3.pdf (Accessed: 9 October, 2021).

BPP Learning Media (2020) *Association of Chartered Certified Accountants Great Britain Audit and Assurance (AA).* London: BPP Learning Media Ltd.

Baharud-din, Z., Shokiyah, A. and Ibrahim, M. S. (2014) ‘Factors affecting the internal audit effectiveness: A survey of the Saudi public sector’, *International Proceedings of Economics Development and Research*, 70, pp. 126-132.

Bahrammirzaee, A. (2010) ‘A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems’, *Neural Computing and Applications*, 19(8), pp. 1165-1195.

Balios, D., Kotsilaras, P., Eriotis, N. and Vasiliou, D. (2020) ‘Big Data, data analytics and external auditing’, *Journal of Modern Accounting and Auditing*, 16(5), pp. 211-219.

Bandyopadhyay, D. and Sen, J. (2011) ‘Internet of Things: Applications and challenges in technology and standardization’, *Wireless Personal Communications*, 58(1), pp. 49-69.

Bao, W., Yue, J. and Rao, Y. (2017) ‘A deep learning framework for financial time series using stacked autoencoders and long-short term memory’, *PloS One*, 12(7), pp. 1-24.

Berger, R. L. (1981) ‘A necessary and sufficient condition for reaching a consensus using DeGroot’s method’, *Journal of the American Statistical Association*, 76(374), pp. 415-418.

Berrar, D. (2017) ‘Confidence curves: An alternative to null hypothesis significance testing for the comparison of classifiers’, *Machine Learning*, 106(6), pp. 911-949.

Bertomeu, J., Cheynel, E., Floyd, E. and Pan, W. (2020) ‘Using machine learning to detect misstatements’, *Review of Accounting Studies*, 26(4), pp. 468-519.

Bhattacharjee, S., Moreno, K. K. and Riley, T. (2012) ‘The interplay of interpersonal affect and source reliability on auditors’ inventory judgments’, *Contemporary Accounting Research*, 29(4), pp. 1087-1108.

Bhattacharyya, T., Chatterjee, B., Singh, P. K., Yoon, J. H., Geem, Z. W. and Sarkar, R (2020) ‘Mayfly in harmony: A new hybrid meta-heuristic feature selection algorithm’, *IEEE Access*, 8, pp. 195929-195945.

Bierstaker, J., Janvrin, D. and Lowe, D. J. (2014) ‘What factors influence auditors’ use of computer-assisted audit techniques?’, *Advances in Accounting*, 30(1), pp. 67-74.

Blake, R. and Mangiameli, P. (2011) ‘The effects and interactions of data quality and problem complexity on classification’, *Journal of Data and Information Quality*, 2(2), pp. 1-28.

Bradford, M., Henderson, D., Baxter, R. J. and Navarro, P. (2020) ‘Using generalized audit software to detect material misstatements, control deficiencies and fraud’, *Managerial Auditing Journal*, 35(4), pp. 521-547.

Brown-Liburd, H., Issa, H. and Lombardi, D. (2015) ‘Behavioral implications of Big Data’s impact on audit judgment and decision making and future research directions’, *Accounting Horizons*, 29(2), pp. 451-468.

Brown-Liburd, H. and Vasarhelyi, M. A. (2015) ‘Big Data and audit evidence’, *Journal of Emerging Technologies in Accounting*, 12(1), pp. 1-16.

Calderon, T. and Cheh, J. (2002) ‘A roadmap for future neural networks research in auditing and risk assessment’, *International Journal of Accounting Information Systems*, 3(4), pp. 203-236.

Cao, M., Chychyla, R. and Stewart, T. (2015) ‘Big Data Analytics in financial statement audits’, *Accounting Horizons*, 29(2), pp. 423-429.

Cao, Q., Leggio, K. and Schniederjans, M. (2005) ‘A comparison between Fama and French’s model and artificial neural networks in predicting the Chinese stock market’, *Computers and Operations Research*, 32(10), pp. 2499-2512.

Carcillo, F. *et al.* (2018) ‘Scarff: A scalable framework for streaming credit card fraud detection with Spark’, *Information Fusion*, 41, pp. 182-194.

Carson, E., Fargher, N. L., Geiger, M. A. and Lennox, C. S. (2013) ‘Audit reporting for going-concern uncertainty: A research synthesis’, *Auditing: A Journal of Practice and Theory*, 32(1), pp. 353-384.

Chan, D. C. V. and Vasarhelyi, M. (2018) *Continuous auditing: Theory and application*. London: Emerald Publishing.

Chang, C. and Chen, C. (2009) ‘Applying decision tree and neural network to increase quality of dermatologic diagnosis’, *Expert Systems with Applications*, 36(2), pp. 4035-4041.

Chen, S. (2019) ‘An effective going-concern prediction model for the sustainability of enterprises and capital market development’, *Applied Economics*, 51(31), pp. 3376-3388.

Chen, Y. J. and Wu, C. H. (2017) ‘On Big Data-based fraud detection method for financial statements of business groups’, *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 986-987.

Chintalapati, S. S. and Jyotsna, G. (2013) ‘Application of data mining techniques for financial accounting fraud detection scheme’, *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(11), pp. 717 - 724.

Cho, S., Miklos, A. V., Ting, S. and Chanyuan, Z. (2020) ‘Learning from machine learning in accounting and assurance’, *Journal of emerging technologies in accounting*, 17(1), pp. 1-11.

Craja, P., Kim, A. and Lessmann, S. (2020) ‘Deep learning for detecting financial statement fraud’, *Decision Support Systems*, 139, pp. 1-13.

Curtis, M. B., Jenkins, J. G., Bedard, J. C. and Deis, D. R. (2009) ‘Auditors’ training and proficiency in information systems: A research synthesis’, *Jou*rnal of Information systems, 23(1), pp. 79-96.

Debreceny, R. S. and Gray, G. L. (2011) ‘Data mining of electronic mail and auditing: A research agenda’, *Journal of Information Systems*, 25(2), pp. 195-226.

Dechow, P., Ge, W., Larson, C. and Sloan, R. (2011) ‘Predicting material accounting misstatements’, *Contemporary Accounting Research*, 28(1), pp. 17-82.

DeGroot, M. H. (1974) ‘Reaching a consensus’, *Journal of the American Statistical Association*, 69(345), pp. 118-121.

Demšar, J. (2006) ‘Statistical comparisons of classifiers over multiple data sets’, *Journal of Machine Learning Research*, 7, pp. 1-30.

Dewiani, Lawi, A., Sarro, M. I. R. and Aziz, F. (2019) 'classification of firm external audit using ensemble support vector machine method', *1st International Conference on Science and Technology, ICOST 2019*. doi: 10.4108/eai.2-5-2019.2284605.

Dong, W., Liao, S. and Zhang, Z. (2018) 'Leveraging financial social media data for corporate fraud detection', *Journal of Management Information Systems*, 35(2), pp. 461-487.

Dopuch, N., Holthausen, R. and Leftwich, R. (1987) 'Predicting audit qualifications with financial', *The Accounting Review*, 62(3), pp. 431-454.

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y. and Liu, S. (2012) 'Multiple classifier system for remote sensing image classification: A review', *Sensors*, 12(4), pp. 4764-4792.

Dutta, I., Dutta, S. and Raahemi, B. (2017) 'Detecting financial restatements using data mining techniques', *Expert Systems with Applications*, 90, pp. 374-393.

Earley, C. E. (2015) 'Data analytics in auditing: Opportunities and challenges', *Business Horizons*, 58(5), pp. 493-500.

Efiong, E., Bassey, B. E., acha Hadrain, A., Charlsie, A. and Golce, B. D. (2017) 'The effects of audit evidence on the audit report of commercial banks in Nigeria', *Asian Journal of Business and Management*, 5(6), pp. 183- 189.

Eilifsen, A. (2010) *Auditing and assurance services.* 2nd edn. London: McGraw-Hill Higher Education.

Eisinga, R., Heskes, T., Pelzer, B. and Te Grotenhuis, M. (2017) 'Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers', *BMC Bioinformatics*, 18(1), pp. 1-18.

Entezari-Maleki, R., Rezaei, A. and Minaei-Bidgoli, B. (2009) 'Comparison of classification methods based on the type of attributes and sample size. *Journal of Convergence Information Technology*, 4(3), pp. 94-102.

Fame, 2021. Fame. Available at: https://fame4.bvdinfo.com/version-20211216/fame/1/Companies/Search. (Accessed 12 January 2021).

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L. and Muller, P. A. (2019) 'Deep learning for time series classification: A review', *Data Mining and Knowledge Discovery*, 33(4), pp. 1-44.

Fernández-Gámez, M. A., García-Lagos, F. and Sánchez-Serrano, J. R. (2016) 'Integrating corporate governance and financial variables for the identification of qualified audit opinions with neural networks', *Neural Computing and Applications*, 27(5), pp. 1427-1444.

Fernández, M. Á., Serrano, J. R. S., Aguilera, D. A. and Casado, G. (2018) ‘Predicting going-concern opinion for hotel industry using classifiers combination’, *International Journal of Scientific Management and Tourism*, 4(1), pp. 91-106.

Francis, J. and Krishnan, J. (1999) ‘Accounting accruals and auditor reporting conservatism’, *Contemporary Accounting Research*, 16(1), pp. 135-165.

Gaganis, C. (2009) ‘Classification techniques for the identification of falsified financial statements: A comparative analysis’, *Intelligent Systems in Accounting, Finance and Management*, 16(3), pp. 207-229.

Gaganis, C., Pasiouras, F. and Doumpos, M. (2007) ‘Probabilistic neural networks for the identification of qualified audit opinions’, *Expert Systems with Applications*, 32(1), pp. 114-124.

Gandomi, A. and Haider, M. (2015) ‘Beyond the hype: Big data concepts, methods, and analytics’, *International Journal of Information Management*, 35(2), pp. 137-144.

García, V., Marqués, A. I. and Sánchez, J. S. (2015) ‘An insight into the experimental design for credit risk and corporate bankruptcy prediction systems’, *Journal of Intelligent Information Systems*, 44(1), pp. 159-189.

George, G., Osinga, E. C., Lavie, D. and Scott, B. A. (2016) ‘Big Data and data science methods for management research’, *Academy of Management Journal*, 59(5), pp. 1493-1507.

Gepp, A., Kumar, K. and Bhattacharya, S. (2010) ‘Business failure prediction using decision trees’, *Journal of Forecasting*, 29(6), pp. 536-555.

Gepp, A., Linnenluecke, M. K., O’neill, T. J. and Smith, T. (2018) ‘Big Data techniques in auditing research and practice: Current trends and future opportunities’, *Journal of Accounting Literature*, 40, pp. 102-115.

Glover, S., Jiambalvo, J. and Kennedy, J. (2000) ‘Analytical procedures and audit planning decisions’, *Aud*iting*: A Journal of Practice and Theory*, 19(2), pp. 27-45.

Gospel, J., Ordu, P., Barigbon, M. and Namapele, A. (2019) ‘Sufficiency and appropriateness of audit evidence for giving an opinion on the true and fair view of financial statements’, *International Journal of Innovative Development and Policy Studies*, 7(3), pp. 36-43.

Gray, G. L. and Debreceny, R. S. (2015) ‘A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits’, *International Journal of Accounting Information Systems*, 15(4), pp. 357-380.

Gweon, H. and Yu, H. (2019) ‘How reliable is your reliability diagram?’, *Pattern Recognition Letters*, 125, pp. 687-693.

Hajek, P. and Roberto, H. (2017) ‘Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods’, *Knowledge-Based Systems*, 128, pp. 139-152.

Hamal, S. and Senvar, O. (2021) ‘Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs’, *International Journal of Computational Intelligence Systems*, 14(1), pp. 769-782.

Hayes, R., Gortemaker, H. and Wallage, P. (2014) *Principles of auditing.* 3rd edn. Harlow: Pearson Education Limited.

Heaton, J., Polson, N. and Witte, J. (2017) ‘DPL for finance: Deep portfolios’, *Applied Stochastic Models in Business and Industry*, 33(1), pp. 3-12.

Hogan, C. E., Rezaee, Z., Riley Jr, R. A. and Velury (2008) ‘Financial statement fraud: Insights from the academic literature’, *Auditing*, 27(2), pp. 231-252.

Holowczak, R., Louton, D. and Saraoglu, H. (2019) ‘Testing market response to auditor change filings: A comparison of machine learning classifiers’, *The Journal of Finance and Data Science*, 5(1), pp. 48-59.

Hooda, N., Bawa, S. and Rana, P. S. (2018) ‘Fraudulent firm classification: A case study of an external audit’, *Applied Artificial Intelligence*, 3(1), pp. 48-64.

Hooda, N., Bawa, S. and Rana, P. S. (2020) ‘Optimizing fraudulent firm prediction using ensemble machine learning: A case study of an external audit’, *Applied Artificial Intelligence*, 34(1), pp. 20-30.

Hoque, F., Islam, M. and Shatabda, S. (2015) ‘A two-tier classification model for financial fraud’, *International Journal of Computer Applications*, 118(19), pp. 1-8.

Hsu, Y. and Lee, W. (2020) ‘Evaluation of the going-concern status for companies: An ensemble framework-based model’, *Journal of Forecasting*, 39(4), pp. 687-706.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y. and Xu, W. (2018) ‘Applications of support vector machine (SVM) learning in cancer genomics’, *Cancer Genomics-Proteomics*, 15(1), pp. 41-51.

Huang, W., Lai, K. K., Nakamori, Y., Wang, S. and Yu, L. (2007) ‘Neural networks in finance and economics forecasting’, *International Journal of Information Technology and Decision Making*, 6(1), pp. 113-140.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K. and Felix, W. F. (2011) ‘Identification of fraudulent financial statements using linguistic credibility analysis’, *Decision Support Systems*, 50(3), pp. 585-594.

Hunton, J. E. and Rose, J. M. (2010) ‘21st century auditing: Advancing decision support systems to achieve continuous auditing’, *Accounting Horizons*, 24(2), pp. 297-312.

Issa, H., Sun, T. and Vasarhelyi, M. A. (2016) ‘Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation’, *Journal of Emerging Technologies in Accounting*, 13(2), pp. 1-20.

Jahani, A. M. and Soofi, F. (2013) ‘Application of hybrid genetic algorithm (GA) and artificial neural networks (ANNs) approach in auditing’, *Advances in Environmental Biology*, 7(10), pp. 2819-2828.

Jan, C.- l. (2018) ‘An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan’, *Sustainability*, 10(2), pp. 1-14.

Jan, C.- L. (2021) ‘Using DPL algorithms for CPAs’ going’, *Information*, 12(2), pp. 1-22.

Janvrin, D., Bierstaker, J. and Lowe, D. J. (2008) ‘An examination of audit information technology use and perceived importance’, *Accounting Horizons*, 22(1), pp. 1-21.

Kaplan, S. E., Pany, K., Samuels, J. and Zhang, J. (2012) ‘An examination of anonymous and nonanonymous fraud reporting channels’, *Advances in Accounting*, 28(1), pp. 88-95.

Kend, M. and Nguyen, L. A. (2020) ‘Big Data Analytics and other emerging technologies: The impact on the Australian audit and assurance profession’, *Australian Accounting Review*, 30(4), pp. 269-282.

Khemakhem, S. and Boujelbene, Y. (2018) ‘Predicting credit risk on the basis of financial and non-financial variables and data mining’, *Review of Accounting and Finance*, 17(3), pp. 316-340.

Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D. and Sharma, M. (2018) ‘Credit card fraud detection using Naïve Bayes model based and K-NN classifier’, *International Journal of Advance Research, Ideas and Innovations in Technology*, 4(3), pp. 44-47.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2008) ‘Support vector machines, decision trees and neural networks for auditor selection’, *Journal of Computational Methods in Sciences and Engineering*, 8(3), pp. 213-224.

Kirkos, E., Spathis, C., Nanopoulos, A. and Manolopoulos, Y. (2007) ‘Identifying qualified auditors’ opinions: A data mining approach’, *Journal of Emerging Technologies in Accounting*, 4(1), pp. 183-197.

Kiziloz, H. E. (2021) ‘Classifier ensemble methods in feature selection’, *Neurocomputing*, 419, pp. 97-107.

Kotsiantis, S., Koumanakos, E., Tzelepis, D. and Tampakas, V. (2006) ‘Forecasting fraudulent financial statements’, *International Journal of Computational Intelligence*, 3(2), pp. 104-110.

Krishnan, J. and Krishnan, J. (1996) ‘The role of economic trade-offs in the audit opinion decision: An empirical analysis’, *Journal of Accounting, Auditing and Finance*, 11(4), pp. 565-586.

Lahmiri, S., Bekiros, S., Giakoumelou, A. and Bezzina, F. (2020) ‘Performance assessment of ensemble learning systems in financial data classification’, *Intelligent Systems in Accounting, Finance and Management*, 27(1), pp. 3-9.

Lessmann, S., Baesens, B., Seow, H. and Thomas, L. (2015) ‘Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research’, *European Journal of Operational Research*, 247(1), pp. 124-136.

Li, H., Dai, J., Gershberg, T. and Vasarhelyi, M. A. (2018) ‘Understanding usage and value of audit analytics for internal auditors: An organizational approach’, *International Journal of Accounting Information Systems*, 28, pp. 59-76.

Lina, K. Z., Fraserb, I. A. and Hatherlyc, D. J. (2003) ‘Auditor analytical review judgement: A performance evaluation’, *The British Accounting Review*, 35, pp. 19-34.

Lin, C. C., Chiu, A. A., Huang, S. Y. and Yen, D.C (2015) ‘Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts’ judgments’, *Knowledge-Based Systems*, 89, pp. 459-470.

Lin, T. H. (2009) ‘A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models’, *Neurocomputing*, 72(16-18), pp. 3507-3516.

Livieris, I., Pintelas, E. and Pintelas, P. (2020) ‘A CNN-LSTM model for gold price time-series forecasting’, *Neural Computing and Applications*, 32(23), pp. 17351-17360.

Lu, N., Lin, H., Lu, J. and Zhang, G. (2012) ‘A customer churn prediction model in telecom industry using boosting’, *IEEE Transactions on Industrial Informatics*, 10(2), pp. 1659-1665.

Mentz, M., Barac, K. and Odendaal, E. (2018) ‘An audit evidence planning model for the public sector’, *Journal of Economic and Financial Sciences*, 11(1), pp. 1-14.

Mikalef, P., Pappas, I., Krogstie, J. and Giannakos, M. (2018) ‘Big Data Analytics capabilities: A systematic literature review and research agenda’, *Information Systems and e-Business Management*, 16(3), pp. 547-578.

Millichamp, A. H. and Taylor, J. R. (2018) *Auditing.* 11th edn. London: Cengage.

Min, J. and Lee, Y. (2005) ‘Bankruptcy prediction using SVM with optimal choice of kernel function parameters’, *Expert Systems with Applications*, 28(4), pp. 603-614.

Mohammadi, M., Yazdani, S., Khanmohammadi, M. and Maham, K. (2020) ‘Financial reporting fraud detection: An analysis of data mining algorithms’, *International Journal of Finance and Managerial Accounting*, 4(16), pp. 1-12.

Mustapha, M. and Lai, S. J. (2017) ‘Information technology in audit processes: An empirical evidence from Malaysian audit firms’, *International Review of Management and Marketing*, 7(2), pp. 53-59.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y. and Sun, X (2011) ‘The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature’, *Decision Support Systems*, 50(3), pp. 559-569.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D. and Joksimovic, I. (2013) ‘The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements’, *Expert Systems with Applications*, 40(15), pp. 5932-5944.

Omidi, M., Min, Q., Moradinaftchali, V. and Piri, M. (2019) ‘The efficacy of predictive methods in financial statement fraud’, *Discrete Dynamics in Nature and Society*, 2019, 4989140. doi: 10.1155/2019/4989140.

Ozdagoglu, G., Ozdagoglu, A., Gumus, Y. and Kurt Gumus, G. (2017) ‘The application of data mining techniques in manipulated financial statement classification: The case of Turkey’, *Journal of AI and Data Mining*, 5(1), pp. 67-77.

Padmavathy, P. and Mohideen, S. (2020) ‘An efficient two-pass classifier system for patient opinion mining to analyze drugs satisfaction’, *Biomedical Signal Processing and Control*, 57, pp. 1-9.

Pai, P., Hsu, M. and Wang, M. (2011) ‘A SVM-based model for detecting top management fraud’, *Knowledge-Based Systems*, 24(2), pp. 314-321.

Papík, M. and Lenka, P. (2020) ‘Detection models for unintentional financial restatements’, *Journal of Business Economics and Management*, 21(1), pp. 64-86.

Papouskova, M. and Hajek, P. (2019) ‘Two-stage consumer credit risk modelling using heterogeneous ensemble learning’, *Decision Support Systems*, 118, pp. 33-45.

Pedro, H., Coimbra, C., David, M. and Lauret, P. (2018) ‘Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts’, *Renewable Energy*, 123, pp. 191-203.

Pohar, M., Blas, M. and Turk, S. (2004) ‘Comparison of logistic regression and linear discriminant analysis: A simulation study’, *Metodoloski Zvezki*, 1(1), pp. 143-161.

Pourheydari, O., Nezamabadi-pour, H. and Aazami, Z. (2012) ‘Identifying qualified audit opinions by artificial neural networks’, *African Journal of Business Management*, 4(66), pp. 11077-11087.

Priyanka, D., Priyanka, J. V. and Rao, S. P. (2020) ‘Statistical analysis of various measures in auditing practices using optimization’, *Science, Technology and Development*, 9(7), pp. 37-87.

Pumsirirat, A. and Yan, L. (2018) ‘Credit card fraud detection using DPL based on auto-encoder and Restricted Boltzmann Machine’, *International Journal of Advanced Computer Science and Applications*, 9(1), pp. 18-25.

PwC (2015) *Data driven: What students need to succeed in a rapidly changing business world.* New York: PwC.

Rafiei, F., Manzari, S. and Bostanian, S. (2011) ‘Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence’, *Expert Systems with Applications*, 38(8), pp. 10210-10217.

Ramlukan, R. (2015) ‘How Big Data and analytics are transforming the audit’, *Financial Executive*, 31(3), pp. 14-19.

Ramona, B., Razvan, B. and Alexandra, F. (2014) ‘Big Data and specific analysis methods for insurance fraud detection’, *Database Systems Journal*, 4, pp. 30-39.

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P. and Nandi, A. K. (2018) ‘Credit card fraud detection using AdaBoost and majority voting’, *IEEE Access*, 6, pp. 14277-14284.

Rashid, C. A. (2017) ‘The importance of audit procedure in collecting audit evidence: Case of Kurdistan Region, Iraq’, *International Journal of Social Sciences and Educational Studies*, 4(2), pp. 15-22.

Ravisankar, P., Ravi, V., Rao, G. and Bose, I. (2011) ‘Detection of financial statement fraud and feature selection using data mining techniques’, *Decision Support Systems*, 50(2), pp. 491-500.

Richins, G., Stapleton, A., Stratopoulos, T. C. and Wong, C. (2017) ‘Big Data Analytics: Opportunity or threat for the accounting profession?’, *Journal of Information Systems*, 31(3), pp. 17-63.

Risteska, S. B. and Trivodaliev, K. (2017) ‘A review of Internet of Things for smart home: Challenges and solutions’, *Journal of Cleaner Production*, 140, pp. 1454-1464.

Rose, A., Rose, J., Suh, I. and Thibodeau, J. (2020) ‘Analytical procedures: Are more good ideas always better for audit quality?’, *Behavioral Research in Accounting*, 32(1), pp. 37-49.

Rostamy-Malkhalifeh, M., Amiri, M. and Mehrkam, M. (2021) ‘Predicting financial statement fraud using fuzzy neural networks’, *Advances in Mathematical Finance and Applications*, 6(1), pp. 137-145.

Saif, S. M., Mehdi, S. and Ebrahimi, F. (2012) ‘Finding rules for audit opinions prediction through data mining methods’, *European Online Journal of Natural and Social Sciences*, 1(2), pp. 28-36.

Saif, S. M., Sarikhani, M. and Ebrahimi, F. (2013) ‘an expert system with neural network and decision tree for predicting audit opinions’, *IAES International Journal of Artificial Intelligence*, 2(4), pp. 151-158.

Salijeni, G., Samsonova-Taddei, A. and Turley, S. (2019) ‘Big Data and changes in audit technology: Contemplating a research agenda’, *Accounting and Business Research*, 49(1), pp. 95-119.

Sánchez-Serrano, J. R., Alaminos, D., García-Lagos, F. and Callejón-Gil, A. M. (2020) ‘Predicting Audit opinion in consolidated financial statements with artificial neural networks’, *Mathematics*, 8(8), pp. 1-14.

Santoso, N. and Wibowo, W. (2018) ‘Financial distress prediction using linear discriminant analysis and support vector machine’, *Journal of Physics: Conference Series*, 979(1), pp. 1-7.

Sathyapriya, M. and Thiagarasu, V. (2017) ‘Big Data Analytics techniques for credit card fraud detection: A review’, *International Journal of Science and Research*, 6(5), pp. 206-2011.

Seddon, J. J. and Currie, W. L. (2017) ‘A model for unpacking big data analytics in high-frequency trading’, *Journal of Business Research*, 70, pp. 300-307.

Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V. and Alonso-Betanzos, A. (2017) ‘Ensemble feature selection: Homogeneous and heterogeneous approaches’, *Knowledge-Based Systems*, 118, pp. 124-139.

Sharma, A. and Panigrahi, P. K. (2012) ‘A review of financial accounting fraud detection based on data mining techniques’, *International Journal of Computer Applications*, 39(1), pp. 37-47.

Singh, D. and Singh, B. (2020) ‘Investigating the impact of data normalization on classification performance’, *Applied Soft Computing*, 97, pp. 1-23.

Singh, R. and Singh, K. (2010) ‘A descriptive classification of causes of data quality problems in data warehousing’, *International Journal of Computer Science Issues*, 7(3), pp. 41-49.

Sivasankar, E., Selvi, C. and Mahalakshmi, S. (2020) ‘Rough set-based feature selection for credit risk prediction using weight-adjusted boosting ensemble method’, *Soft Computing*, 24(6), pp. 3975-3988.

Skurichina, M. and Duin, R. (2002) ‘Bagging, boosting and the random subspace method for linear classifiers’, *Pattern Analysis and Applications*, 5(2), pp. 121-135.

Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T. (2018) ‘Big Data: Deep learning for financial sentiment analysis’, *Journal of Big Data*, 5(1), pp. 1-25.

Song, X., Hu, Z., Du, J. and Sheng, Z. (2014) ‘Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China’, *Journal of Forecasting*, 33, pp. 611-626.

Sreenivasarao, V. and Vidyavathi, S. (2010) ‘Comparative analysis of fuzzy C-mean and modified fuzzy possibilistic C-mean algorithms in data mining’, *International Journal of Computer Science and Technolog*, 1(1), pp. 104-106.

Stanišić, N., Radojević, T. and Stanić, N. (2019) 'Predicting the type of auditor opinion: Statistics, machine learning, or a combination of the two?', *European Journal of Applied Economics*, 16(2), pp. 1-58.

Suganya, R. and Shanthi, R. (2012) 'Fuzzy C-means algorithm: A review', *International Journal of Scientific and Research Publications*, 2(11), pp. 440-442.

Sun, J., Jia, M. and Li, H. (2011) 'AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies', *Expert Systems with Applications*, 38(8), pp. 9305-9312.

Sun, T. (2019) 'Applying deep learning to audit procedures: An illustrative framework', *Accounting Horizons*, 33(3), pp. 89-109.

Sun, T. and Sales, L. J. (2018) 'Predicting public procurement irregularity: an application of neural networks', *Journal of Emerging Technologies in Accounting*, 15(1), pp. 141-154.

Sun, T. and Vasarhelyi, M. A. (2018) 'Embracing textual data analytics in auditing with deep learning', *International Journal of Digital Accounting Research*, 18, pp. 49-67.

Tang, J. J. and Karim, K. E. (2017) 'Big Data in business analytics: Implications for the audit profession', *The CPA Journal*, 87(6), pp. 34-39.

Tang, J. and Karim, K. E. (2018) 'Financial fraud detection and big data analytics: Implications on auditors' use of fraud brainstorming session', *Managerial Auditing Journal*, 34(3), pp. 324-337.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. and Asadpour, M. (2020) 'Boosting methods for multi-class imbalanced data classification: An experimental review', *Journal of Big Data*, 7(1), pp. 1-47.

Tealab, A., Hefny, H. and Badr, A. (2017) 'Forecasting of nonlinear time series using ANN', *Future Computing and Informatics Journal*, 2(1), pp. 39-47.

Titera, W. (2013) 'Updating audit standard: Enabling audit data analysis', *Journal of Information Systems*, 27(1), pp. 325-331.

Tsai, C. (2008) 'Financial decision support using neural networks and support vector machines', *Expert Systems*, 25(4), pp. 380-393.

Tsai, C. F. (2009) 'Feature selection in bankruptcy prediction', *Knowledge-Based Systems*, 22, pp. 120-127.

Tsai, C. F. (2014) ‘Combining cluster analysis with classifier ensembles to predict financial distress’, *Information Fusion*, 16, pp. 46-58.

Tsai, C. F., Hsu, Y.-F. and Yen, D. C. (2014) ‘A comparative study of classifier ensembles for bankruptcy prediction’, *Applied Soft Computing*, 24, pp. 977-984.

Uddin, N., Meah, M. and Hossain, R. (2013) ‘Discriminant analysis as an aid to human resource selection and human resource turnover minimization decisions’, *International journal of Business and management*, 8(17), pp. 153-169.

Vasarhelyi, M. A., Kogan, A. and Tuttle, B. M. (2015) ‘Big Data in accounting: An overview’, *Accounting Horizons*, 29(2), pp. 381-396.

Wang, S. (2010) ‘A comprehensive survey of data mining-based accounting-fraud detection research’, *2010 International Conference on Intelligent Computation Technology and Automation*, pp. 50-53. doi: 10.1109/ICICTA.2010.831.

Warren, J. D. J., Moffitt, K. C. and Byrnes, P. (2015) ‘How Big Data will change accounting’, *Accounting Horizons*, 29(2), pp. 397(11).

Weisheimer, A. and Palmer, T. (2014) ‘On the reliability of seasonal climate forecasts’, *Journal of the Royal Society Interface*, 11(96), pp. 1-10.

Woźniak, M., Grana, M. and Corchado, E. (2014) ‘A survey of multiple classifier systems as hybrid systems’, *Information Fusion*, 16, pp. 3-17.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014) ‘Data mining with Big Data’, *IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp. 97-107.

Xiao, Z., Wang, Y., Fu, K. and Wu, F. (2017) ‘Identifying different transportation modes from trajectory data using tree-based ensemble classifiers’, *ISPRS International Journal of Geo-Information*, 6(2), pp. 1-22.

Yao, J., Pan, Y., Yang, S., Chen, Y. and Li, Y. (2019) ‘Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: A multi-analytic approach’, *Sustainability*, 11(6), pp. 1-17.

Yaşar, A., Yakut, E. and Gutnu, M. M. (2015) ‘Predicting qualified audit opinions using financial ratios: Evidence from the Istanbul Stock Exchange’, *International Journal of Business and Social Science*, 6(85), pp. 57-67.

Yeo, A. C. M. and Carter, S. (2017) ‘Segregate the wheat from the chaff enabler: Will Big Data and data analytics enhance the perceived competencies of accountants/auditors in Malaysia?’, *Journal of Self-Governance and Management Economics*, 5(3), pp. 28-51.

Yijing, L., Haixiang, G., Xiao, L., Yanan, L. and Jinling, L. (2016) ‘Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data’, *Knowledge-Based Systems*, 94, pp. 88-104.

Yoon, K., Hoogduin, L. and Zhang, L. (2015) ‘Big Data as complementary audit evidence’, *Accounting Horizons*, 29(2), pp. 431-438.

Yu, Z., Wang, Z., You, J., Zhang, J., Liu, J., Wong, H. S. and Han, G. (2017) ‘A new kind of nonparametric test for statistical comparison of multiple classifiers over multiple datasets’, *IEEE Transactions on Cybernetics*, 47(12), pp. 4418-4431.

Zareapoor, M. and Shamsolmoali, P. (2015) ‘Application of credit card fraud detection: Based on bagging ensemble classifier’, *Procedia Computer Science*, 48, pp. 679-685.

Zerbino, P., Aloini, D., Dulmin, R. and Mininno, V. (2018) ‘Process-mining-enabled audit of information systems: Methodology and an application’, *Expert Systems with Applications*, 110, pp. 80-92.

Zhang, J., Yang, X. and Appelbaum, D. (2015) ‘Toward effective Big Data analysis in continuous auditing’, *Accounting Horizons*, 29(2), pp. 469-476.

Zhang, Q., Yang, L. T., Chen, Z. and Li, P. (2018) ‘A survey on deep learning for Big Data’, *Information Fusion*, 42, pp. 146-157.

Zhou, L. (2013) ‘Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods’, *Knowledge-Based Systems*, 41, pp. 16-25.

Zuca, S. (2015) ‘Audit evidence-necessity to qualify a pertinent opinion’, *Procedia Economics and Finance*, 20, pp. 700-704.

Zuh, D. (2010) ‘A hybrid approach for efficient ensembles’, *Decision Support Systems*, 48(3), pp. 480-487.