PREDICTING BUSINESS FAILURE USING ARTIFICIAL INTELLIGENCE SYSTEM

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

by

Yaser Allozi

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

December 2021

Abstract

Predicting business insolvency is considered one of the main supportive sources of information for decision making for financial institutions, investors, creditors, and other participants in the business market. Financial reporting systems provide relevant information that can be used to assess the financial position of firms. It is crucial to have classification and prediction models that can analyse this financial information and provide accurate assurance for users about business health. Recent studies have explored the use of machine learning tools as substitute for traditional statistical methods to develop classification models to classify firm insolvency according to financial statement information. However, these models have no ideal classifier, since each provides a certain percentage of wrong outputs, which is a crucial consideration; every percentage of wrong response can mean massive financial losses for stakeholders. Therefore, this study proposes new insolvency classification and perdition models based on machine learning modelling techniques to develop an improved classifier.

Individual modelling techniques using statistical methods and machine learning were used to develop the classification model of business insolvency. The results showed that machine learning method outperformed statistical methods. Deep Learning (DPL) achieved the highest performance based on all performance measurements used in the study, and it was the best individual classifier, with average accuracy of 97.2% using all-years dataset. Ensemble-Boosted Decision Tree classifier ranked second, followed by Decision Tree classifier. Thus, it has been proven that DPL modelling approach is useful for business insolvency classification. A key contribution in enhancing individual classifier outputs is the use of traditional combining methods with two new aggregation methods in business insolvency (Fuzzy Logic and Consensus Approach). The Consensus Approach showed the best improvement in the results of all individual classifiers with average accuracy of 97.7%, and it is considered the best classification method not only in comparison with individual classifiers, but also with traditional combiners.

This study pioneers the development of a time series business insolvency prediction model with Big Data for UK businesses. The aim of the model is to provide early prediction about a business health. Three prediction models were developed based on Nonlinear Autoregressive with Exogenous Input models (NARX), Nonlinear Autoregressive Neural Network (NAR), and Deep Learning Time-series model (DPL-SA) and achieved average accuracy rates of 83.6%, 89.5%, and 91.35%, respectively. The results show relatively high performance in comparison with the best individual classifier (deep learning).

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.

Yaser Allozi

December 2021

London, UK

Dedication

This thesis is dedicated to my late father, Mohammed, for his continuous support of all types when he was alive. May ALLAH grant him Al-Jannah. Next in line is my mother, Ryya, who has been always supporting me in my life, and her endless love and encouragements were the main support to achieve my dream. This work would not have been possible without her support and her permanent support. May ALLAH continue to guide, guard, bless, and preserve her and give her long life and prosperity, Ameen.

I also dedicate this work to my dear brothers and sisters, Saud, Abdelrahman, Mustafa, Basma, Sumia, Sana, and Eman, for their moral support. My success is entirely thanks to my family.

Acknowledgements

First and foremost, I am grateful and thankful to almighty ALLAH who has given me the strength and patience to complete this work and who remains to bless me every day.

I am very grateful and thankful to my thesis supervisor, Dr Maysam Abbod, for his substantial guidance and enduring support to the completion and succession of this work. I appreciate his endless and valuable efforts to keep my work progress in the right pathway and being there when I have needed him. His encouragements and comments have made a huge contribution my personal and academic development.

I want to use this opportunity to express my heartfelt appreciation and gratitude to everyone who has assisted and supported me throughout my PhD journey. Special thanks go to my friends Dr Mohammed Radi and Dr Ziad Hunaiti for their endless support to me, and for being true friends during this journey.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

ACC	Average Accuracy Rate
AUC	Area Under Curve
AVG	Average Rule
Cons	Consensus Combiner
DT	Decision Tree
DPL	Deep Learning
DPL-SA	Deep Learning-Step Ahead
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MajVot	Majority Voting Rule
MAX	Maximum Rule Combiner
MIN	Minimum Rule Combiner
NAR	Nonlinear Autoregressive
NARX	Nonlinear Autoregressive with Exogenous Input
NB	Naïve Bayes
NN	Neural Networks
ROC	Receiver Operating Characteristics
SVM	Support Vector Machines
WAVG	Weighted Average Rule

Chapter 1
Introduction

1.1. Background

The increasingly interconnected globalised economy has increased competition among business institutions to remain competitive and retain or increase market share worldwide, which has become a universal need among all business stakeholders. This puts more pressure on companies, and increases uncertainty about their survival, compounded by advanced technologies in robotics and the Internet of Things that are poised to revolutionise industries and supply chains. Therefore, companies face several critical challenges in their marketing, human resources, financial supply, and innovative systems dimensions, which can eventually lead to business failure if not managed effectively (Li and Sun, 2011). Business failure is defined as the state in which the company is not capable of honouring its commitment (Smiti and Soui, 2020). Many factors, such as weak corporate governance, lack of innovation, and weak performance of business management in a market full of uncertainty may hinder the achievement of organisational goals and can eventually lead to business failure (Gordini, 2014).

There are numerous factors that lead to companies failing, the most prominent of which among major corporations are high interest rates, recessionary profits, and high debt burdens (Charitou *et al.*, 2004). In addition, specific industrial features can have profound impacts, such as government regulation and the nature of operations, which could contribute to the financial distress of companies and lead to failure (du Jardin, 2015). Studies of business failure patterns in Australia, Canada, the UK, and the US have shown that small, private and newly formed companies with inefficient control procedures and poor cash flow planning are more vulnerable than larger public companies to financial distress (Charitou *et al.*, 2004). For any size of institution there are significant economic costs of business failure, which extend far beyond the failed institutions themselves (e.g., ancillary impacts of unemployment, interrelated with socio-economic development and local and national economic growth). Evidence shows that before their final collapse, the market value of distressed companies decreases significantly; thus, business failures severely affect capital suppliers, investors, and creditors, as well as internal stakeholders (management, employees, and supply chain partners). Auditors may also face the

risk of prosecution when firms fail, as they are mandated to provide qualified audit opinions that provide early warning signals about potential failure of companies (Lisic *et al.*, 2015).

Business failure prediction is of great importance to help business stakeholders make rational economic decisions. Enterprise condition concerns the business community components of its participants as well as policy makers and the global economy. Hence, the economic and social costs resulting from businesses' inability to survive have triggered the attention of researchers to better asses the financial condition of enterprises and their future long-term operation prospects using their financial results (Jabeur *et al.*, 2021). Companies' own financial reporting systems (which are subject to state regulation) are considered to be a major indicator used by stakeholders to get information about business status (Williams, 2016).

Business failure analysis faces the issue of how to classify firm status (as either active or failed). The aim of classification models is to help users to make better economic decisions on whether to invest in businesses based on their financial indicators. After the pioneering study of Altman (1968), the field of business failure modelling has been widely researched and developed by scholars deploying statistical approaches such as Linear Regression and Multiple Discriminant Analysis (MDA), to build and develop classification models that have been used later as a baseline benchmark model in the field. More recent technological developments have enabled the use of artificial intelligence (AI) modelling techniques such as Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM) as substitutes for traditional statistical methods of analysis, for the purpose of developing more robust business failure models (Heo and Yang, 2014; Tsai *et al.*, 2014; du Jardin, 2015; Iturriaga and Sanz, 2015; Barboza *et al.*, 2017; Jing and Fang, 2018; Smiti and Soui, 2020).

The large amount of information on firms' financial data has attracted the attention of many researchers to apply more data mining techniques in order to improve classification accuracy in evaluating the situation of firms. It has been concluded from the literature that machine learning techniques outperform statistical methods in terms of classification accuracy of business insolvency. However, despite of the impressive results achieved by researchers in the field, and the ability of data mining tools to classify business failure using financial data, there is still a call for larger sets and more complex modelling techniques to develop models and increase classification accuracy (De Andrés *et al.*, 2011; Lee and Choi, 2013; Zhou, 2013; Gordini, 2014).

The use of various techniques in the construction of business failure models has evolved with time. Researchers have adjusted the design of the individual business failure models using more complex modelling methods, such as ensemble and combining methods, in an attempt to improve model classification capabilities and to enhance performance (Lee and Choi, 2013; Wang *et al.*, 2014; Barboza *et al.*, 2017; Fan *et al.*, 2017; Choi *et al.*, 2018; du Jardin, 2021). The basic idea of utilising this method is to achieve better classification results, using a group of single classifiers that have been trained and tested in classifying business status individually. However, computational costs accompanied with increased modelling complexity can be a drawback when adopting ensemble and combining methods in business failure studies. Nevertheless, it is believed that using more complex modelling methods could lead to better business failure models, which is the main aim of this thesis (Lee and Choi, 2013; Wang *et al.*, 2014; Barboza *et al.*, 2017; Fan *et al.*, 2017; Choi *et al.*, 2018; du Jardin, 2021).

The purpose of this study is to clarify the potential capability of more complex data mining as a classification tool in accounting and financial analysis to classify and to predict business failure. This study attempts to develop an accurate classification model and another early prediction model using different using data mining techniques. This study is the first to use Big Data of financial information to develop a model capable to classify and to predict business failure in the UK.

1.2. Research Motivation

Recently, the research in the area of business failure prediction has focused on adopting individual classification models. However, complex ensemble methods have also been adopted based on combined pools of diversified individual classifiers for building prediction models. There are two types of combination methods that have been used in business failure studies: homogenous and heterogeneous classifier ensembles. The first method combines different classifiers that have the same algorithm, such as ensemble boosting DTs, while the second combines classifiers that have different algorithms. Both methods aim to enhance classifiers prediction performance through offsetting the weaknesses of individual classifiers.

Recent business failure prediction studies have shown that ensemble classifiers can enhance prediction results and produce better performance than single AI classifiers. Most of these studies have focused on homogenous combining methods, such as weighted average, majority voting, reliability-based methods, and fuzzy rules. However, few studies have focused on employing heterogeneous combining methods. Ensemble learning uses n heuristic algorithm

to combine all classifiers' decisions, after they are trained independently, to produce one final classification decision.

1.3. Aim and Objectives

The main aim of this study is to discover and explore a new combining technique and new individual classifiers that can enhance business failure prediction for the UK firms by developing a new combination method and classifiers. This study addresses the question of whether new classification techniques and complex combining methods can enhance classification performance and produce more reliable results. The process in developing the proposed models starts with simple individual classifiers using different individual data mining techniques, followed by adopting and implementing various combination techniques to achieve the main goal of this study. Firstly, deep learning classifier is used as an individual classifier, to answer the question of whether it can outperform other individual classifiers. Then, a new combination method (Cons combiner and Fuzzy logic) is developed whereby individual classifiers work in tandem, interacting and cooperating to solve the same problem. Another aim of this study is to explore new time-series classifiers, such as Nonlinear Autoregressive Neural Network (NAR), NARX, and DPL-SA, to predict business failure one year earlier. To achieve the main aim, this study seeks to address the following objectives:

- Collect a balanced dataset of the UK businesses.
- Clean the datasets.
- Implement nine individual classifiers, including DPL classifier and LR as a benchmark for performance comparisons.
- Improve individual classifiers' performance by implementing Cons and Fuzzy logic combiners and comparing the outcomes with traditional combining methods.
- Implement time-series modelling methods (NAR, NARX, and DPL-SA).
- Measure classifiers' performance based on different performance measurements.
- Using significant test, Friedman, and Bonferroni to validate classifier performance and demonstrate that Cons has better classification performance than other individual or combined classifiers tested in this work.

1.4. Contributions to Knowledge

This study uses various modelling techniques to build and develop business failure prediction models in an attempt to improve classifier performance. The main contributions of this work are as follows:

- A comprehensive literature review on different modelling approaches used in the field of business failure prediction.
- Using a balanced Big Dataset of UK firms.
- Applying a new individual classifier using DPL techniques and comparing it with all available individual classifiers that have been used in studies of business failure classification.
- Improving individual classifier performance through two new combining methods (Cons and Fuzzy Logic) and comparing the results with different traditional combining models.
- Implementing a time-series prediction classifier to predict business failure in advance.
- Introducing new performance measurement to validate classifier performance.
- Using significant tests to explore which classifier has the most reliable prediction performance.
- Allozi, Y., Abbod, M., (2021). Predicting Business Failure using Neural Network: An Empirical Comparison with Statistical Methods and Data Mining Methods. The 3rd International Conference on Deep Learning, Artificial Intelligence and Robotics, (ICDLAIR). University of Salerno, Italy. (Accepted).

1.5. Structure of the Thesis

This thesis consists of seven individual chapters, structured as follows.

Chapter 2 presents the theoretical background in the field of business failure and its related issues in the literature, focusing on the historical development of the field accompanied, with a focus on the source of business failure indicators that used to develop classification and prediction models. Secondly, it focuses on the application of data mining techniques in other financial areas, such as fraud prediction and on-going concern prediction. The next section presents a systemic review of related works on business failure that have utilised data mining techniques for developing their proposed models, followed by a comprehensive critical analysis of its findings to finally highlight the gaps in the reviewed works.

Chapter 3 presents the methodological experimental design adopted on this thesis, explaining each step in developing the classification and prediction models. The chapter explains data collection and pre-processing techniques used to structure the proposed classifiers. Classification methods and all performance measurements are explained in their theoretical context.

Chapter 4 presents all individual classifiers used to develop business failure model (ANN, LR, linear discriminant analysis [LDA], KNN, SVM, DT, NB, ENS-DT, and DPL. All individual classifiers' results were presented, followed by a comparison in terms of their performance measurements. Significance testing were used to determine the best classifier that can classify business failure most accurately.

Chapter 5 presents two new combining methods (Cons and Fuzzy logic) and another six traditional committee combiners (AVG, WAVG, Median, MAX, MIN, and Majority Voting), which were tested to improve the individual classifier performance (presented in Chapter 4). This chapter explains the theoretical background of each combining method. Their performance results after tested are presented and discussed. Finally, a decision is made to determine the best combiner method that has the best classification improvement in term of correctly classifying firm status.

Chapter 6 presents the time-series modelling techniques adopting in this study to predict business failure in advance (including NAR, NARX, and DPL-SA). The prediction results of each model are then evaluated and presented based on all performance measurements, and are compared with the performance results of the individual classifier DPL for the Year 2019 benchmark dataset. In the final step, the statistical significance test is applied to determine which method has the best prediction capability for business failure.

Chapter 7 highlights the main conclusions drawn from the experimental results of all classifiers' performance, noting the salient outcomes, It also identifies the study limitations and directions for future research in this field.

Chapter 2
Literature Review

2.1. Introduction

This chapter presents a comprehensive review of literature relating to business failure prediction modelling approaches. It begins with the theoretical background of business failure and its definition, the importance of financial reporting and financial analysis as a main indicator of business failure to assess firms' health, and its on-going concern on the market. This is followed by the history of business failure prediction modelling techniques, from statistical to more advanced machine-learning methods. Modelling techniques are then described, and their prediction performance is discussed in detail, to better understand their predictive capabilities. To date, there are a large number of studies in the literature which have been undertaken by researchers and scholars to enhance business failure classification using a variety of efficient modelling approaches. Only those which are more related and most relevant to business failure that used quantitative modelling methods were selected and collected for analysis in this study, in order to achieve its aim. Finally, all findings are demonstrated and summarised, along with justifications of the importance of the research trends of classification and business failure.

2.2. Business Failure Conceptualisation

The integration of worldwide markets and economic globalisation has increased competition among businesses in all industries and areas. This places more pressure on businesses in an increasingly competitive environment and poses more uncertainty about survival. Consequently, enterprises are more vulnerable to potential disruptions, crises, or inefficiencies in any operational processes, including marketing, human recourses, financial supply, innovation systems and so on, any of which can potentially lead to business failure (Li and Sun, 2011). Ineffective management and innovation to address defects in business institutions can result in losses and business failure. In most cases, business failure is a result of a combination of various critical factors within the enterprise, which are ultimately reflected in the financial position of the firm (i.e., economic losses). The accumulation of sustained losses accompanied with poor management undermines the prospects for firm development and survival, and for larger firms – or clusters of interdependent firms in a supply chain network or

market segment – can potentially cause major macroeconomic impacts, such as financial crises in the capital market (Lin, 2008).

An increasing number of failed firms in general is associated with substantial losses in the economy, thus the need to classify and predict their failure early in advance has become a very important topic. Therefore, both the academic and the industrial fields have become increasingly concerned about how to effectively and correctly predict the survival of the firms using their financial performance across the years, in order to protect the national macroeconomic environment as well as to optimise short- and long-term individual investment prospects.

Business failure can be caused by numerous factors, including legal, economic, financial, strategic, organisational, and managerial circumstances. Numerous studies over the years have attached varying importance to particular indicators of business failure. Altman (1983) highlighted new firm formation, economic growth, and credit policy. Wruck (1990) found that financial distress was chiefly related to inadequate cash flow, and that information asymmetries makes it hard for firms to restructure and renegotiate with creditors when they are facing financial difficulties. More recently, Huynh *et al.* (2020) demonstrated that information asymmetry in Vietnam has a major negative effect on firm value. Several studies have been carried out to identify the default risk of corporations by conducting financial analysis on accounting information, highlighting the informative capabilities of such information in classifying and predicting business failure (du Jardin, 2015; Barboza *et al.*, 2017).

Enterprise default predictions are an important informative tool used in various fields across the economy, helping corporations to establish and adjust their strategies based on their current survival status, as indicated by various prediction models. Executive management can be more effective and avoid possible failures using key indicators of default risk. Such data provides investors with substantial information about companies' on-going survival, reducing uncertainty in the market, which allows them to better manage their portfolios and avoid the likelihood of firms' defaults. Moreover, business failure prediction can help governments to revise and impose macroprudential policies to improve the economy through more appropriate financial regulations (Alaka *et al.*, 2018). As a result of all these beneficial aspects of firms' failure prediction models, the financial system can be designed and improved to avoid financial crisis and market instability. The 2008 global financial crisis and increasing demand to reduce uncertainty in world markets highlighted the importance of this field. Therefore, employing

and developing prediction models using data mining and machine learning tools to determine firms' status has become the cutting edge of advanced financial engineering in recent studies.

The quality of business failure classification plays a major informative role for many firm stakeholders, such as investors, creditors, employees, and suppliers, etc. From an investor perspective, it is crucial to minimise uncertainty about the businesses they are working with or investing in, to minimise any potential losses that might accrue from any failure of the business. Moreover, other stakeholders who rely on business survival and profitability of the entity that they have business with need to understand the level of entity on-going operation and its financial health to reduce risk and protect their operation income (Priego *et al.*, 2014). In relation to the high-risk environments facing businesses these days, adopting business failure prediction models can reduce the cost of financial analysis, the uncertainty of the market in general and of the business in particular, and speed-up the business evaluation, allowing better observation of business accounts and health. Another benefit of business failure prediction is for auditing firms. A substantial role of the auditor is to assess the on-going concern related to audited businesses, and failure risk analyses can be an excellent tool for auditing (Kuruppu *et al.*, 2003).

In the UK, the term 'bankruptcy' in the business field refers only to individuals, who are governed by the Insolvency Rules 1986, and part nine of the Insolvency Act 1986. The *de facto* 'bankruptcy' of a limited company is referred to as 'insolvency', which is governed by UK Insolvency Law, by which a business can be compelled to undergo compulsory liquidation relatively easily (Gov.uk, 2019). However, business insolvency cannot be easily defined in a single definition. According to Watson and Everett (1993), it is simply defined when one of four circumstances occur:

- Sale of the business to avoid more losses.
- Termination of trading and loss of credit.
- Ending the business for any other reasons.
- Not successfully starting a (new) business.

In this study, insolvent businesses are those that have gone through bankruptcy based on receivership or liquidation under UK insolvency law, or which have merged with more stable firms. Studied firms are those with available financial data, which can be extracted for analysis. Figure 2.1 displays total company liquidations in England and Wales by sector as of Q4 2019.

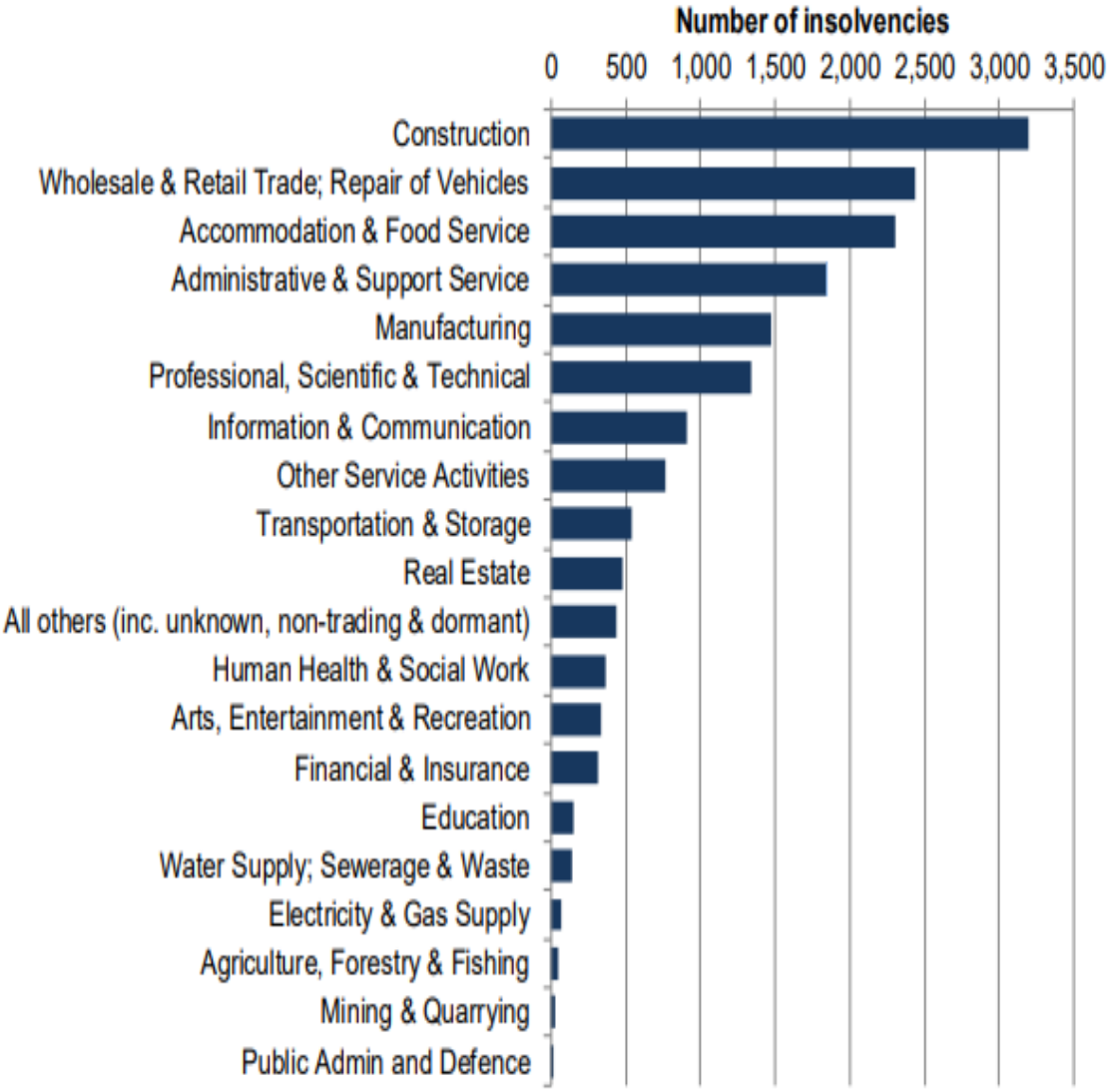

Figure 2.1: Total company liquidations in England and Wales by broad industry sector, year ending Q4 2019

Source: Gov.uk (2019)

2.3. Financial Reporting System

Financial reporting is a communication language used by management to represent financial information to firm stakeholders, such as existing and potential investors, creditors, suppliers, government agencies, and other external users (Ball, 2006). The reporting system consists of the firm's financial department preparing financial statements, with the collaboration of other departments within the firm, including narrative and numerical representation of company operations, interests, managerial assumptions, and financial positions on a quarterly and annual basis (Ikpefan and Akande, 2012). Narrative disclosures highlight the accounting and financial

policies the firm complies with to reflect its managerial assumptions and estimations used to report its resources, operations, and liabilities. Numerical disclosures form the core of the financial reporting system and the applicable regulatory framework, quantifying the financial position of firms through transparent and verifiable representation of business models, resources, sources of finance, and operating activity. The purpose of financial disclosures is to provide material high-quality information to users that can be effectively used for monitoring on-going operations (Ball, 2006). Hence, business failure models rely heavily on this information to be able to classify and predict the continuity of business. The reporting system consists of five essential statements identified by Mackenzie *et al.* (2012) and explained below.

- Statement of financial position

The financial position statement is the most important, providing summary information about three main elements: the assets that the firm owns, and the liabilities and owners' equities used to finance these assets. Assets are divided into current and non-current assets, thus the liabilities and the total amount of assets should equal the summation of both the liabilities and owner equity.

- Income statement

The income statement consists of the firm's total revenue, the subtraction of all types of expenses and costs, and the consequent net income.

- Cash flow statement

The cash flow statement consists of operating, investing, and financing cash flow used by the firm during the accounting period.

- Change in equity statement

The change in equity statement provides information about shareholders' equity in the firm and how earnings are divided (retained earnings and share dividends).

- Statement of comprehensive income

The statement of comprehensive income is a general summary of the firm's earnings relative to the above factors.

All of these financial statements should provide relevant, reliable, and faithful representation of financial information to users, to support their decision-making process, and publicly listed firms are legally obliged to disclose certain types of information. According to Barth *et al.*

(2008), financial statements make fundamental differences for stakeholder decision-making, which is the reason they are so important. Faithful representation means that the financial information represents the real economic situation of the firm in the market, with confirmatory and predictive value reflecting reality. Naturally, firm stakeholders (particularly management) have an interest to present performance favourably, while the board of directors is entrusted to oversee managerial activities on behalf of shareholders, to avoid information asymmetry between principals and agents. Many researchers have tested the value relevance of financial information, and they normally define it as value relevant (Barth *et al.*, 2008; Clarkson *et al.*, 2011; Yip and Young, 2012).

According to Barth *et al.* (2008) there are two primary qualitative characteristics that financial statements should deliver to users to inform their decision-making with useful information: relevance and faithful representation. Relevance is defined as the capability of the delivered financial information to make a difference to users' financial decisions. It means that the financial information has confirmatory and predictive value, which is reflected in the real market. Faithful representation means that the financial information represents the real economic situation in the market. Moreover, the financial statements have a complementary enhancing qualitative characteristic alongside the primary characteristics, which are verifiability, timeliness, comparability, and understandably. The purpose of these enhancing characteristics is to make financial information more useful for readers and more reliable for decision making.

In the UK, the financial reporting system went through a major development regarding the emergence of different regimes used by different firms' classes for reporting. Small and medium size companies were first defined and permitted by the Companies Act 1981 to report their financial information using abbreviated accounts (Iatridis, 2010a). Later in 1994 these companies were exempted from auditing their reports. Another amendment to their reporting regulations occurred in 1997, when they were allowed to report under simpler accounting regulations according to FRSSE. Moreover, they were eligible to report their financial performance according to UK GAAP. From January 2005 onwards, in compliance with EU requirements, all listed companies in the UK started complying with International Financial Reporting Standards (IFRS) in all published consolidated accounting statements, while unlisted medium-sized and large firms continued to use the UK's own Financial Reporting Standards (FRS).

The IFRS are the accountability rules issued by an independent London-based organisation, the International Accounting Standards Board (IASB). They offer a set of rules which they claim should ideally apply equally to all public companies' financial reporting worldwide. According to regulatory bodies such as the Financial Accounting Standards Board (FASB) and the IASB, financial statement reporting provides relevant and reliable information to users that is useful for the decision-making process. Therefore, many researchers have tested the value relevance of financial information under the IFRS, and they normally define it as value relevant (Ball, 2006; Barth *et al.*, 2008; Clarkson *et al.*, 2011; Yip and Young, 2012).

IFRS adoption is considered a substantial step toward the harmonisation and globalisation of the financial reporting standards (Leuz and Wysocki, 2016). Another advantage of the IFRS is the improvements it achieves in financial statements' qualitative characteristics, such as transparency and comparability, which in return provide more informative information for financial statement users (Ball *et al.*, 2015). These reports help financial analysis and other financial statement users to better understand firm performance through extracting different financial ratios as an indicator of profitability, and to measure the survival prospects, to make better, more informed decisions (Iatridis, 2010b).

2.4. Financial Statements Analysis

It is obvious from the previous section that financial reports are an indispensable source of information for business failure prediction. With this noted, this section illustrates the analytical process of financial reports that provide the final informative features of business failure models. Aside from its implicit meaning, financial analysis comprises the process of understanding and interpreting the financial information provided in financial statements and reports (Rashid, 2018). The analysis of financial statements is at the core of developing reliable business failure prediction models. Financial statements analysis is defined as the process of converting financial information into more meaningful ratios that can be useful tool to serve managements and other financial statements users to analyse firms' historical performance and allow comparison with other peers (Rashid, 2018). Financial statements illustrate firms' operations and profitability during a particular time period.

Despite their historical nature, financial ratios can provide useful information for users in terms of financial matters, but interpretation is still ultimately subjective, relating to user perspectives in the context of personal goals and concerns, and expected market developments in the future (Garrison *et al.*, 2010). Traditionally, firms were evaluated mainly in terms of on-going

operational efficiency by management, future ability to meet obligations by creditors and suppliers, and future profitability and dividends by shareholders and potential investors; based on these dimensions, several types of financial ratios can be calculated according to different priorities (Altman, 1968). Financial analysis using financial ratios plays a major role in examining the firms' future predictions and business failure, which are the primary objectives for all users (Garrison *et al.*, 2010).

Business failures are generally a result of financial distress that companies experience and cannot sustain or overcome. Analysing company financial statements is an informative source of indicators used to assess the level of the financial difficulties a business might experience that could lead to bankruptcy and failure (du Jardin, 2015). Financial statements analysis is a useful tool to understand the financial condition of a company, providing an illustration of its performance and profitability. Through financial statements analysis, projections about a company's future can be drawn using various financial aspects, which can also be used as an indicator of possible bankruptcy. Therefore, predicting the continuity or going concern is an important aspect of the financial analysis, whereby prediction models can be constructed using the output of the analysis to avoid losses resulting from firm failure.

In an investigation of the importance of financial analysis as a determination of the companies' condition, Bhargava *et al.* (2017) conducted a study on the telecommunication industry and concluded that there is a high demand for providing financial measurements to monitor the economic performance of businesses for improved understanding and decision making by investors and other stakeholders. Therefore, financial statement analysis plays a major role in assessing firms' financial conditions and worthiness. Other studies have proven the importance of financial ratios and their usefulness in predicting firms' market and share values (Lewellen, 2004). However, although such analyses offer useful information on financial reports and provide an in-depth understanding of firm condition, precaution has to be taken when interpreting them (Abraham, 2004).

Mesak (2019) analysed the Indonesian stock exchange to explore the impact of financial ratios in identifying financial distress conditions. Ratios related to liquidity, profitability, financial leverage, and operating cash flow were extracted from companies' annual financial reports and were used as a predictive variable. The model was developed based on logistic regression applying 5% significance level. The results showed negative relationships between companies' financial distress and ratios related to liquidity and profitability, and a positive relationship

with financial leverage. However, cash flow ratios showed no effect on the condition of firms. It can be concluded that companies' management can use these analyses as an early sign of possible failure and make and adjust new policies to prevent the firm descending into bankruptcy.

Kulustayeva *et al.* (2020) used financial ratios calculated from the publicly available financial statements of insurance companies of the Republic of Kazakhstan as indicators of their profitability and stability. Profitability, leverage, and liquidity ratios were selected as independent variables in the model. The empirical testing showed that financial leverage has the greatest impact on companies' profitability and stability, with a positive relationship. However, according to Barua *et al.* (2018), financial leverage ratios have no impact in the short term on the profitability of insurance companies, and they have a negative impact over the long term. These studies revealed the importance of the analysis of reported financial data and its superior informative information for users such as investors, creditors, and other stakeholders to assess and predict firm stability, and thus to make better decisions.

Kanapickienė and Grundienė (2015) studied 'The Model of Fraud Detection in Financial Statements by Means of Financial Ratios on the International Scientific Conference Economics and Management - 2015 (ICEM-2015)', to identify the best financial ratios indicative of financial statement fraud. They developed a fraud detection model based on best financial ratio fraud indicators, applied to financial ratios extracted from the financial statements of 40 fraudulent and 125 non-fraudulent Lithuanian firms. Logistic regression method was used for model creation. The independent variables of the model included 51 financial ratios, including profitability, liquidity, solvency, activity, and structure ratios. Their findings indicated that 32 financial ratios were suitable for use in classifying fraudulent Lithuanian companies.

Samman (2015) studied the industrial sector of Oman to measure the financial determination of firms' profitability. With a model consisting of seven ratios, they concluded that financial analysis has a significant interpretation performance in determining firm's profitability and stability. Mbona and Yusheng (2019) investigated the best financial ratios to measure the performance of the Chinese telecom industry. They analysed financial statements using a multiple regression model consisting of 18 ratios as independent variables to determine firm performance. The aim of the study was to help different stakeholders picking the most significant ratios that reflect companies' performance.

Eng *et al.* (2018) conducted a fundamental analysis of a sample of 2,164 Chinese companies using nine financial ratios extracted from firms' financial statements to investigate their relationship with excess returns. The results show an association between these ratios and firms' excess returns, whereby five ratios had a negative relationship, and the other four had a positive relationship. Hence, high quality financial statement figures provide valuable firm-specific information. However, an earlier study by Bai *et al.* (2006) on the Chinese market found a weak relationship between financial ratios and excess return; they emphasised that financial ratios' explanatory power varies with time.

Many research studies investigated the purpose of financial ratios as an indicator of firms' profitability and liquidity in various countries. Bolek and Wolski (2012) found that investors in the Warsaw Stock Market are more concerned with investing in companies with a higher level of liquidity which maintain a high level of cash to meet their obligations. Investor and other financial statement users' perspectives support maintaining high levels of profitability and liquidity, represented in financial ratios, in order to assess on-going operations and to avoid any potential bankruptcy (Behn *et al.*, 2001).

Financial ratios provide a useful tool to serve firm management and investors in assessing the financial situation of the enterprise through the process of analysing and comparing their historical financial performance. This allows them to illustrate what has occurred in a certain time to better assess the risk a company faces. However, according to Noreen *et al.* (2011), most financial statement users are more concerned about what will occur in the future. For illustration, Altman (1968) indicated that creditors and lenders are more concerned with the future capability of the firm to meet all its obligations, while stockholders and investors are more concerned about dividends and profitability. In order to assess these matters, financial ratios for different purposes can be extracted and calculated from the firm's income statement, balance sheet, change in equity, and cash flow statements, and be used for business failure model development.

2.5. The History of Financial Insolvency Modelling

The main purpose when developing a business failure model is to establish the best classification technique that can accurately discriminate between healthy and failing companies, and accordingly classify and predict business status. Business failure classification and prediction have been applied widely in the area of finance (Zavgren, 1985; Watson and Everett, 1993; Boritz and Kennedy, 1995; Youn and Gu, 2010; Li and Sun, 2011; Priego *et al.*,

2014; De Bock, 2017; Alaka *et al.*, 2018). Historically, a wide range of classification techniques have been used by researchers in the field, varying from statistical (e.g., LR and LDA) to machine learning (e.g., NN, DT, and SVM). The most significant difference between statistical and machine learning techniques is in classification superiority, since statistical methods rely on assumptions to study the relationship between the dataset features in order to predict outcomes, whereas machine learning does not require any assumptions about dataset features, and relies directly on available data to develop a classification system (Guang-Bin Huang *et al.*, 2004).

According to Beaver (1968), the first study to identify the risk of business failure was conducted in 1908 by Rosedale, who used financial information related to companies' current assets as an indicator of firm failure. Successive researchers in the field focused on using firm's financial information and financial ratios analysis to assess the business insolvency risk more accurately (Fitzpatrick, 1932; Smith, 1935; Beaver, 1966), applying univariate discriminant analysis to multiple financial measures to predict the companies' health status.

Altman (1968) was the first study to propose a foundation model for predicting firms' default in accordance with Beaver's (1966) recommendations. The model, called Z-score, was constructed using five financial ratios representing the financial condition of the firm as the predictor of bankruptcy. Z-score pioneered the use of MDA, which subsequently became the most commonly used statistical modelling tool. It is capable of generating an ordinal ranking of firms' failure risk, called the credit score. Altman Z-score is considered to be the first prediction model that consists of several financial ratios as independent variables that can best determine whether a firm is default-based on a linear discriminant function of these variables. A firm can be classified as default if the score is above a certain threshold, and as normal if the score is below that threshold. MDA was subsequently popular in accounting and financial literature (Taffler, 1982), and numerous later studies were simply used by finance professionals without considering the assumptions that are to be met for MDA model to be valid. This has led to improper application, whereby some assumptions restrict the generalisation of the discriminant analysis model, such as the proviso that the independent variables should follow multivariate normal distributions (Joy, 1975; Richardson, Davidson, 1984; Zavgren, 1985).

The next generation of predicting business failure used binary response models to classify firms' health status. Binary classification models classify firms based on their activity status as '1' for active firms and '0' for inactive (bankrupt, insolvent, and failed) ones. In most cases,

explanatory variables such as financial ratios are used for model construction to estimate firms' failure probability based on logistic function, such as pro-bit modelling. An example is the O-score model proposed by Ohlson (1980), which applies logistic function using financial ratios to predict firm bankruptcy. This transformation to a binary classification has several advantages over the previous discriminant analysis approach. First, it does not require a specific distribution of the independent variables or any assumptions about the probability of the firm's status. Second, it can allow to verify the explanatory power of each of the independent variables used in the model. Lastly, it can be deployed to predict the probability of firms' failure in advance. Moreover, Kukuk and Rönnberg (2013) developed a mixed logit model as an extension of a binary classification that allows non-linearity and stochastic parameters in the prediction variables.

The third generation of predicting firm's failure using statistical models is called hazard models. According to Shumway (2001), this approach deploys duration analysis to build a model that is capable to predict firm defaults over time better than traditional single-period models. The hazard model is defined as an early trigger about firm defaults or as a survival analysis used to calculate the probability of a firm failing over time. It uses Cox's (1972) hazard regression model as a prediction methodology for binary classification of firm status, where firm status is longer classified once failure occurs (Whitehead, 1980). Many later studies developed the model further to enhance its prediction performance. Chava and Jarrow (2004) confirmed its superior prediction performance, and Nam *et al.* (2008) used time varying covariates to include temporal and macroeconomic predictive factors in an extension of Shumway's (2001) analytical methodology. Dakovic *et al.* (2010) conducted a study on Norwegian companies and proposed a discrete hazard model using a mixed linear model, which they proved outperformed conventional models using Altman's (1968) variables. Tian *et al.* (2015) improved the prediction performance of the hazard model using variable selection methodology, and Traczynski (2017) applied a Bayesian model averaging methodology on hazard model, whose results showed improved performance in correctly predicting default firms compared to typical models.

Due to the increasing availability of advanced technology from the early 1990s, researchers were able to develop and use more sophisticated techniques as alternatives to traditional statistical tools, enabling them to handle larger datasets. To better understand and predict the risk of companies' bankruptcy, non-parametric modelling methods such as ANN began to be used (Odom and Sharda, 1990; Coats and Fant, 1991; Tam and Kiang, 1992; Wilson and

Sharda, 1994; Serrano-Cinca, 1997). As neural network modelling techniques have shown higher prediction performance over traditional statistical approaches, the race to find more effective methods using computing applications continued. Recent years have seen a new category of methods resulting from the implementation of more advance computational modelling techniques, such as genetic algorithm (Shin and Lee, 2002), SVM (Härdle *et al.*, 2012), and colony algorithm (Zhang and Wu, 2011). These methods, in comparison with traditional linear multivariate analysis or logit and probit analysis, enable more capabilities to cope with modelling imprecisely defined problems, large data, and data with incomplete features. While a variety of techniques have been historically used for building business failure prediction models, the main focus recently has been to develop models using machine learning techniques.

Machine learning, pioneered by Samuel (1959), is a tool used by computers to learn without a clear program. These tools have demonstrably superior outperformance capabilities in predicting firm's bankruptcy compared to traditional statistical models (Barboza *et al.*, 2017). Among computerised AI techniques, the most commonly used for bankruptcy prediction are ANN and NN (Aziz and Dar, 2006; Tseng and Hu, 2010), simply because they are the most popular architecture. The ANN back-propagation algorithm for bankruptcy was arbitrarily used in many studies (Wilson and Sharda, 1994; Tam and Kiang, 1992; Odom and Sharda, 1990; Boritz *et al.*, 1995). In addition, ANN's prediction model for a relatively small sample size was developed by Fletcher and Goss (1993), who noted that ANNs generally need large samples for maximum performance (Tam and Kiang, 1992; Wilson and Sharda, 1994).

The SVM classification modelling algorithm has been widely used to solve classification problems, including firm failure prediction. SVM models can classify firms as one of two classes (failed or active) using a separating hyperplane, each of which has number of features. Shin *et al.* (2005) applied SVM algorithm to classify bankrupt firms and showed that it performs better for predicting firm bankruptcy in comparison with back-propagation ANN algorithm. Chiu *et al.* (2011) compared bankruptcy prediction performance of traditional statistical models with the different intelligent modelling techniques and found that SVM achieved the highest performance for both short-term and long-term failure predictions. Liang *et al.* (2016) studied bankruptcy prediction using an SVM model based on financial ratios and corporate governance indicators as input variables, demonstrating improved prediction performance in comparison with other modelling methods.

DT is a classification methodology to solve binary classification problems such as firm failure. Olson *et al.* (2012) compared DT algorithm with other machine learning algorithms and concluded that it is more understandable and provides better performance for bankruptcy predictions. Zięba *et al.* (2016) conducted a study to predict firm's bankruptcy using ensembled boosted DT accompanied by synthetic feature generation where the results of the experiment showed an outperformance of the proposed model in comparison with other classification methods. Tsai *et al.* (2014) developed bankruptcy prediction models using ensemble DT, SVM, and neural network, and the results showed the superior performance of ensemble DT.

ANN is a learning algorithm used to solve problems in processing systems based on the operation of the human brain system. It uses a simple structure mimicking the structure of the brain, which enables it to solve complex problems such as bankruptcy prediction. Yang *et al.* (1999) explored the prediction capabilities of NNs to predict bankruptcy and claimed that it had the best prediction performance in comparison with other statistical approaches. Azayite and Achchab (2016) enhanced the prediction capabilities of an NN algorithm by incorporating discriminant variables. Geng *et al.* (2015) conducted a study to predict financial distress on Chinese companies and found that NN models outperformed other data mining classification methods.

Prusak (2018) reviewed literature on the importance and application of enterprise bankruptcy prediction in central and eastern European countries. The aim of the study was to review the usage of prediction models to assess bankruptcy risk in studied countries and to determine the level of advancement achieved in the field, where In the Czech Republic, Karas and Režňáková (2014) adopted linear multidimensional discriminant analysis and boosted DT methods to build and compare the performance of bankruptcy predictions. They concluded that non-parametric modelling methods have outperformed traditional statistical methods and are more efficient. In Hungary, Bozsik (2010) constructed two prediction models based on LDMA and ANN techniques and compared their prediction efficiency, finding that NN techniques had superior performance in correctly predicting insolvent firms. Moreover, many researchers in Poland, Latvia, Lithuania, Romania, Slovakia, and Ukraine have conducted similar studies to investigate in the application of traditional statistical methods or non-parametric methods, comparing their prediction performance based on their national datasets. It was concluded that foreign prediction models developed in developed countries were mostly used in these studies, which may not be transposable to developing or different national contexts, which calls for the development of more nation-specific models (Prusak, 2018).

Salehi *et al.* (2016) conduct a study to predict corporate financial distress using four data mining classifiers: ANN, KNN, SVM, and NB. The data represented five financial variables related to 117 companies listed in the Tehran Stock Exchange for the period from 2011 to 2014, with distressed companies categorised according to Article 141 of the Iranian Commercial Codes (e.g., companies whose accumulated losses exceed half of their equity). The data was divided into two group samples: a training dataset consisting of 75 firms (43 healthy and 32 distressed), and a control sample consisting of 42 firms (20 healthy and 22 distressed). Three performance measurements were used to compare models' predictions capability: Type I and Type II Error, and average accuracy rate. The analysis of the experimental results showed that ANN outperformed other classifiers with an average accuracy rate of 88.1% in the year before financial distress, and 97.62% for data two years before financial distress. In terms of efficiency, SVM ranked second, and NB fourth.

Li and Miu (2010) compared the classification performance of both traditional statistical methods as a benchmark and the top 10 commonly used data mining tools with classification and regression tree (CART) for the purpose of predicting business failure. The two baseline benchmark methods were MDA and logistic regression. SVM and KNN were selected as the two most commonly used modelling techniques among the top 10 data mining algorithms in the field of business failure prediction modelling. They used data consisting of 30 financial ratios related to 135 pairs of companies (failure and healthy) from the Shenzhen Stock Exchange and Shanghai Stock Exchange. Only four financial ratios were selected (activity, liability, and growth ratios, and per share items and fields) using stepwise MDA method to develop the models. The experimental results show the superior performance of CART over the two statistical methods at the level of 5%, and over SVM and KNN at the level of 10%, with an average accuracy rate of 90.3%. However, the concluded that the employment of MDA as a feature selection method did not enhance CART performance, which indicates that using all features would be more suitable to produce higher classification accuracy.

Jabeur *et al.* (2021) constructed a classification model to predict corporate failure using CatBoost modelling technique. The model uses Ordered Boosting, which overcomes the problem of target leakage. The advantages of the new gradient boosting algorithm include its lower information loss and ability to successfully work with categorical features. Moreover, it is considered useful for small datasets. The new modelling approach was compared with two statistical tools (MDA and LR) and six reference machine learning algorithms (NN, SVM, RF, Gradient Boosting Machine, Deep NN, and Extreme Gradient Boosting). The model was

trained and tested with 18 financial variables extracted from the Orbis database related to the financial statements of French firms in order to predict failure one, two, and three years before failure. The dataset was divided into 70% for training and 30% for testing. The proposed model using CatBoost outperformed other classifiers in terms of average accuracy rate and area under the ROC curve. However, according to the results, it only outperformed other classifiers on data for data one year before failure; XGboost showed higher average accuracy for year two before failure, and RF and NN showed higher average accuracy rates for year three before failure.

2.6. Data Mining Applications in the Financial Analysis

2.6.1. Financial Statement Auditing and Fraud Detection

A financial audit is an objective evaluation of an organisation's financial reports and reporting processes, conducted by an independent auditor. The primary responsibility of a financial auditor is to provide confidence to stakeholders such as regulators, investors, directors, and managers that financial statements are accurate, complete, and free of error or fraud. Although they provide a reasonable level of assurance, they do not give the users of financial statements absolute assurance. Since fraud detection has to account for a number of unknown and inconsistent factors, it has become an exceptionally difficult endeavour that requires both skill and technological innovations. Using business analytics tools for analysing and detecting fraud in financial statements by auditors is considered very beneficial (Holsapple *et al.*, 2014). Detection and prediction analytics modelling is the next trend of analysing data, and it is considered as a useful tool to predict what will happen in the future (Bertsimas, Kallus, 2014).

Chen (2016) constructed a financial statement fraud detection model to analyse Taiwan's listed and OTC companies from various industries, but they excluded the financial industry, since its financial statements and ratios are not comparable to those of other industries. The data contained companies known to have issued fraudulent and non-fraudulent financial statements, with a matching sample design of one fraudulent to three normal companies. Hybrid methods of data mining were used to build up the model, and DT, BBN, SVM, and ANN were used for variable selection. The model consisted of 30 independent variables, including 23 financial variables and seven non-financial variables, to predict and classify fraudulent and normal companies. The results showed that the DT CHAID, combined with CART, provided high accuracy in detecting financial statement fraud.

Amani and Fadlalla (2017) explored the applications of data mining in the accounting field through a systematic review of the topic in previous literature. Their research methodology consisted of seven steps to capture most related literature from different data resources. 209 data mining and accounting related papers were selected to be studied in order to be presented in a structurally logical and thematically coherent approach. Their results showed that 82% of reviewed studies used data mining applications for predictive goals, and 11%, and 7% for descriptive and prescriptive goals, respectively. In terms of research topics, 64% of the research was related to accounting assurance and compliance; managerial accounting consisted of 25%; and financial accounting and accounting information system topics comprised 11% of the sample. The review noted that data mining techniques are mainly used for classification purposes, followed by estimation.

Gray and Debreceny (2014) sought to provide a transparent taxonomy of data mining for the detection and prediction of financial statement fraud. The research focused on the applications of data mining in the auditing field. The authors explained auditing process phases and fraud detection patterns and demonstrated the degree of the applicability of data mining in each phase and pattern. They concluded that there is a growing general awareness of data mining capabilities on auditing and fraud detection by financial statement regulators, standard setters, and accounting firms. Moreover, they implied that text mining such as mining the text in firm's emails, annual reports, and MD&A could be used by financial analysts and auditors to enhance fraud detection and prediction accuracy. The paper provided useful guidance for future research on the application of data mining to fraud detection in financial statement audits.

Hajek and Henriques (2017) developed an early financial statement fraud detection model using data from annual financial statement and MD&A notes from a dataset of 622 US firms from different industries. Machine learning methods (NN, DT, SVM, and ensemble classifiers) were used to construct a model consisting of 32 financial variables (firm size, profitability ratios, activity ratios, business situation, liquidity ratios, leverage ratios, and share related ratios), and eight linguistic ones (qualitative information from managerial discussions and analysis of firm performance, such as frequency count of positive, negative, tone, and uncertainty). The research findings indicate that ensemble DTs achieved higher accuracy in determining true fraudulent reporting firms (true positive). However, Bayesian Belief Network (BBN) outperformed other remaining methods in detecting true non-fraudulent firms (true negative), which could provide potential decision support for auditors.

Dalnial *et al.* (2014) investigated the capabilities of financial ratios to predict fraudulent financial reports, using financial ratios related to financial leverage, profitability, asset composition, liquidity, capital turnover, and Altman Z-score model. Multiple regression was used as a statistical analysis method to analyse a dataset of the financial statements of 130 Malaysian publicly listed firms (65 fraudulent and 65 non-fraudulent). The research results indicated that the means of profitability ratios exhibited no differences between fraudulent and non-fraudulent firms, whereas the other variables showed significant differences. However, based on multiple regression result, total debt/total equity, receivable/revenue, and Z- score ratios can be an effective instrument for constructing a detection financial fraud model.

Alles and Gray (2015) researched the application of Big Data Analytics (BDA) tools and their advantages in auditing, finding that such solutions can provide auditors with superior analytical capabilities, such as strong predictive power, fraud investigation, and the ability to develop predictive models for going concern decision. Moreover, they could also be used to reduce financial statement fraud and increase 'red flag' discoveries, since the substantial Big Data content and the concept of 100% sampling it provides prevents fraudsters from manipulating all data elements.

Benyoussef and Khan (2017) conducted a research study on financial statement fraud detection using the concept of information manipulation theory. The aim of the study was to analyse the relationship between quantity, quality, manner, and timing of financial restatement and committing financial statement fraud. They used the Audit Analytics database of 254 US firms which announce restatements during 2009 and 2010 and developed a regression model using 18 variables related to four components (quantity, quality, manner, and timing) as independent variables, and the occurrence of fraud as the dependent variable. The results revealed four significant variables: (1) date of discovering the material error, (2) the issue being discussed with the audit committee, (3) the presence of item 4.02, and (4) accuracy magnitude.

Albashrawi (2016) conducted a systematic literature review of research on data mining techniques used for the detection of different types of financial fraud for the period 2005 to 2014. The sample comprised 65 articles (58 from journals and seven conference papers) and summarised them based on fraud type, dataset used, data mining technique employed, and best-performing technique (highest accuracy). Data mining tools were used for financial statement and bank fraud detection in 63% (n = 41) of the articles. Logistic regression modelling technique was found to be the most commonly employed tool for detecting fraud across

different financial applications, followed by DT, NN, then SVM. Moreover, most articles were conducted with data from the US (23 articles), followed by Taiwan and China (with eight and seven articles, respectively). Only three of the 65 articles were conducted on UK data.

2.6.2. Going-Concern Prediction

'Going concern' is a professional term in auditing whereby auditors investigate whether the firm will be continuing their normal business operation and sustainable development in the future. The main task of auditors is to investigate the available financial information from the firm's financial statements to evaluate the probability of the enterprise continuing in business or facing financial distress, which could lead to discontinuity. Moreover, companies prepare their financial statements based on the going-concern basis unless the management declares its intention to liquidate or terminate business through a general-purpose financial statement. Hence, it is the auditor's responsibility to provide financial statement users and other stockholders with assurance that these financial statements are free from material misstatement and are prepared on a going-concern basis.

To cater to the advent of the application of data mining techniques in the accounting and auditing field, many researchers adopted data mining tools in their studies to develop skilled models to classify going-concern firms. According to Martens *et al.* (2008), the first going-concern research studies used MDA to construct classification models to support auditor decision judgments (Levitan and Knoblett, 1985; Mutchler, 1985). Later studies mainly used logistic regression as a modelling technique to test for the explanatory power of predictor variables (Menon and Schwartz, 1987; Bell and Tabor, 1991; Chen and Church, 1992; Raghunandan and Rama, 1995; Mutchler *et al.*, 1997; Behn *et al.*, 2001; Gaeremynck and Willekens, 2003). However, a key limitation of statistical traditional classification methods is that they have to be in accordance to the required specific assumptions in the data. Therefore, Martens *et al.* (2008) empirically investigated the sampling methodology of previous researchers and introduced a more advanced data mining technique with an SVM-based classifying model and rule-based classifiers to predict going-concern doubt. Their proposed work provides a decision table allowing auditors and other users easy consultation in everyday audit business practices.

Salehi and Fard (2013) developed a going-concern prediction model using data mining approach which they applied to a balanced data consist of 146 Iranian manufacturing companies listed in the Tehran Stock Exchange for the period from 2011 to 2011. Using

stepwise discriminant analysis for variable selection, only 10 financial ratios were selected out of 42 that were extracted from the financial statements of these companies. The prediction model was developed using Classification and Regression Tree (CART) and Naïve Bayes Bayesian Network (NBBN) to develop the going-concern prediction model, and their results showed; the CART model achieved 98.62% and the NBBN model achieved 75.55%.

Yeh *et al.* (2014) proposed a going-concern prediction model using hybrid random forests and rough set approach in an attempt to enhance the prediction accuracy of going-concern models in the literature. The study used 27 variables related to 220 Taiwanese companies listed on the Taiwan Economic Journal for the period 2004 to 2008. The results showed an increase in accuracy and fewer type 1 and type 2 errors in predicting company continuity.

Goo *et al.* (2016) extended going-concern research by attempting to improve prediction performance for Taiwanese listed companies. They used 22 financial ratios extracted from the financial statements of listed companies from 2002 to 2013. It was the first study to use the least absolute shrinking and selection operator (LASSO) to select important predictive variables before applying data mining techniques. The prediction model was constructed based on NN, SVM, and CART using four selected variables. According to the empirical results, SVM model outperformed NN and CART models based on prediction accuracy of 89.79% and type 1 error measurements of 10%.

Jan (2021) constructed an ongoing concern model using data mining tools for certified public auditors to make correct judgments about firms' going-concern decisions. Deep neural networks (DNN) and recurrent neural networks (RNN) were used as modelling methods, with CART to select important variables. The dataset extracted from the Taiwan Stock Exchange and the Taipei Exchange consisted of 352 companies in total, including 88 companies with going-concern doubt, during the period from 2002 to 2019. Using 16 financial variables and three non-financial variables to construct the optimal model, the RNN model achieved an average accuracy rate of 93.92%.

2.7. Big Data Analytics for Business Failure Prediction

Big Data was directly addressed by Mashey in Silicon Graphics project 'Big Data and the Next Wave of InfraStress' (Diebold, 2012). big Data is defined by its three major characteristics, known as the three V's: velocity, volume, and variety (Zikopoulos and Eaton, 2011). Volume is used to categorise the size of data, velocity represents the generation speed and the type of

analysis required to deal of such data, and variety refers to the variability of data. Another two V's (veracity and value) were later added (Hitzler and Janowicz, 2013). Although the widespread perception about Big Data is that it has to do with unstructured data, structured data still can be classified big Data as long as it has the necessary characteristics (Zikopoulos and Eaton, 2011).

With the advanced analytical capabilities of data mining techniques, Big Data gained a substantial interest among researchers to solve predictive, perspective, descriptive, and inferential analytics' problems (Ohlhorst, 2012; Talia, 2013). The type of analytics used to solve business failure problems is predictive analytics, which is concerned with using past happenings within a dataset to make predictions about future trends, patterns, and probabilities of events. The term 'classification' is usually used in the business failure literature to describe predictive analytics problems, but in some cases inferential analytics has been used as a subordinate of predictive analytics, helping explain the interaction of independent variables with the dependent variable in the dataset (LaValle *et al.*, 2011). The aim of this study focuses on using predictive analytics to answer the question of what will happen, and inferential analytics to select explanatory variables.

The volume characteristic of big Data datasets entails the use of advanced technological tools for storing and processing data (Suthaharan, 2014). Big Data volume requires clusters of computers running in parallel mode, in order to analyse the data and unmask potential patterns (Fan and Bifet, 2013). Although size is defining feature of Big Data, the type of analysis used to analyse the data play a major contribution to the process. Jacobs (2009) conducted a study on a demographic dataset consisting of the world population in a table of ten columns (including gender, ethnicity, material status, and religion, among others) and 7 billion rows. The dataset was stored in a 100-gigabyte hard disk. They attempted to load the data on an enterprise-grade database system using a super performance computer, but the experiment had to be aborted six hours later due to an unsuccessful upload. Hence, it is clear that it can take several days to perform a serious analysis on vast data sizes, and big Data classification can be accorded based mainly on analysis requirements.

The data of thousands of businesses in the UK over some years can be qualified as Big Data. A Microsoft Excel file containing financial input data in rows and columns of thousands of firms might not be considered as 'Big', but it would be onerous on any computer to perform a complex analysis such as business failure classification for a large number of cases using

machine learning tools with iterative classification analysis. For instance, du Jardin (2010) used a relatively large dataset of 500 companies in modelling to predict business failure using ANN algorithm. The results indicated a very good model with a computational duration of five days on 30 running PCs. In contrast, with recent modern technologies such as BDA, these computations can be performed in seconds. In this study a large dataset consisting of thousands of UK firms is used for modelling using BDA, alongside high-performance AI tools.

Richins *et al.* (2017) provided a conceptual framework aiming to provide insight into whether BDA could offer opportunities or threats for the accounting profession (i.e., in that BDA applications could help the automation of many of the tasks hitherto performed by accountants and auditors). In order for accountancy to remain relevant in the BDA context, they proposed a conceptual framework with a data type and analysis approach. Data type was segregated into structured and unstructured data, and the analysis approach was broken down into problem-driven and exploratory analysis. Structured data includes highly organised data generated from firms' systems (sales, inventory, and customer/supplier management systems) which can be easily included in a traditional analysis. Unstructured data are extracted in a variety of forms (text, audio, and video) from variety of sources, such as firm websites, social media accounts, and financial websites – a BDA tool is required to extract features and patterns from such rich data. The authors implied that problem-driven analysis of structured data was used before the era of BDA, but problem-driven analysis on unstructured data and all exploratory analysis of structured and unstructured data are necessary in the Big Data context. They discussed the implications of BDA in financial accounting, management accounting, and auditing, and explained how accountants could enhance their analysis skills and add value to their firms. They emphasised that accountants should start learning BDA techniques and understand principles of programming if they want to be able to communicate with data in the future.

2.8. Business Failure Studies

The field of business failure prediction has been widely investigated in the literature for more than five decades. Although many methods have been utilised in this regard, from statistical to machine learning, the latter has shown more capabilities in implementing single classifier, ensemble, and hybrid models than statistical modelling methods, resulting in more reliable and efficient models. Hence, machine-learning techniques are believed to overcome the shortcomings of statistical methods. Therefore, this thesis focuses on quantitative approaches

of machine-learning methods used in developing business failure prediction models for a database of UK firms.

In practice, real historical financial performance datasets used to develop business failure models differ in size, categories, and characteristics. Since single machine-learning classifiers are not capable of triggering the relationships among these data, some researchers have employed hybrid modelling methods to better capture the classification strength of data and to exploit the potential relationships between them. Moreover, some researchers have deployed the ensemble methods that allow the enhancement of single classifiers' ability to learn from different parts of data and develop higher performance prediction models. The results of these studies using hybrid and ensemble methods have shown their superiority compared to single or individual classifiers. For this reason, this thesis focuses on exploring and applying hybrid and ensemble methods in the field of business failure prediction.

The following subsections explain the mechanism used in this thesis for collecting related studies from the literature.

2.8.1. Literature Review Collection Process

The large number of studies related to business failure prediction modelling using machine-learning techniques is indicative of the importance of the topic in financial studies. In this thesis, the collection process started with searching for the keywords 'business failure', 'bankruptcy prediction', 'business distress', 'machine-learning', and 'data mining' in the relevant fields using four academic science databases: IEEE, Springer, Science Direct, and Google Scholar. The intended search focuses on papers published from 2010 to 2021, in order to include the latest research. Initial searching yielded a huge number of resources, including journal papers, articles, conference papers, and books.

Since journal papers and articles generally provide more in-depth and cutting-edge information about modelling techniques used and data processing more than conference papers and books, only these sources were only included in the further research as the main source of the literature review content. Papers that aimed at using single classifiers, hybrid, and ensemble methods containing business failure prediction were selected and organised in sequential order from 2010 to 2021. All findings were summarised comprehensively and discussed thoroughly based on database nature, modelling methods used, feature selection processes, performance measurements, and type of variables used.

The following subsection summarises the main characteristics of the relevant studies to outline the history of business failure prediction using machine-learning techniques.

2.8.2. Literature Discussion and Analysis

In this section, 37 papers were selected from various scientific journals focusing on machine learning approaches used in developing business failure models. Table 2.1 summarises all valuable information related to methods used, data size, performance measures, and the salient findings extracted from the papers that could lead to reliable conclusions about business failure models using different approaches. Table 2.1 includes all the key information contained within the related studies, taking into consideration most important aspects of building and developing business failure model. The number of firms used in training, validating, and testing the model accompanied with the number of features used for building the model are shown, considering data splitting ratios between failed and active firms in the dataset. Moreover, the data pre-processing using feature selection methods used to train and test models are considered to be an important factor in improving prediction performance. The number of classification techniques used in each study is an essential consideration as it allows the assessment and comparison of how different classifiers perform in correctly predicting business failure. Moreover, hybrid and ensemble approaches are considered, as they represent studies' approaches to enhance model prediction performance. Another aspect is the significance test used to test model prediction performance reliability and robustness.

Table 2.1: Comparison of related studies

No.	Study	No. of firms	Active/Failed	Number of variables	Variable selection method	Classification techniques	Ensemble approach	Performance measures	Significance test
1	Cho *et al.* (2010)	1000	50%/50%	15 financial variables	Yes	ANN, DT, CBR[1], LR	-	ACC	-
2	Yoon and Kwon (2010)	10000	50%/50%	24 credit sales variables	Yes	SVM, ANN, MDA[2], LR	-	ACC	Yes
3	du Jardin (2010)	1020	50%/50%	41 financial variables	Yes	ANN, MDA, LR	-	ACC	-
4	De Andrés *et al.* (2011)	59474	99.77%/0.23%	22 financial variables	Yes	ANN, MDA, LR	-	Total error, Type I & Type II Error	-
5	Chen *et al.* (2011)	1200	50%/50%	31 financial variables	Yes	SVM, ANN, GA[3]	-	Total error, Type I & Type II Error	-
6	Chen *et al.* (2011)	244	53.3%/46.7%	30 financial variables	-	SVM, ANN, GA	-	ACC, AUC, Type I & Type II Error	Yes
7	Li (2011)	370	50%/50%		Yes	RF[4], MDA, LR	-	-	-
8	du Jardin and Séverin (2012)	2360	50%/50%	41 financial variables	Yes	ANN, GA	-	ACC	Yes
9	Jeong *et al.* (2012)	2542	50%/50%	27 financial variables	Yes	ANN	-	Total error	-
10	Tsai and Cheng (2012)	653	45.3%/54.7%	-	-	SVM, ANN, DT, LR	-	ACC, type 1 & Type II Error	-

Table 2.1: Comparison of related studies (cont.)

No.	Study	No. of firms	Active/Failed	Number of variables	Variable selection method	Classification techniques	Ensemble approach	Performance measures	Significance test
11	Kristóf and Virág (2012)	504	86.7%/13.3%	31 financial variables	-	ANN, DT, LR	-	ACC, AUC, ROC	-
12	Lee and Choi (2013)	1775	66.2%/33.8%	21 financial variables	Yes	ANN, MDA	-	ACC	-
13	Zhou (2013)	2010	50%/50%	27 financial ratios	Yes	ANN, DT, MDA, LR	-	ACC, sensitivity, specificity	-
14	Kasgari *et al.* (2013)	135	52.5%/47.5%	49 financial variables	Yes	ANN, GA, LR	-	ACC, sensitivity, specificity, positive predictivity	-
15	Arieshanti *et al.* (2013)	240	53.3%/46.7%		Yes	SVM, ANN	-		-
16	Zhou *et al.* (2014)	2010	50%/50%	27 financial ratios	Yes	ANN, DT, MDA, LR	-	ACC, sensitivity, specificity	-
17	Gordini (2014)	3100	51.6%/48.4%		Yes	SVM, GA, LR	-		-
18	Heo and Yang (2014)	2762	50%/50%		-	SVM, ANN, DT, MDA	-		-
19	Tsai *et al.* (2014)	690	44.5%/55.5%		-	SVM, ANN, DT	-		-
20	Gordini (2014)	3100	52%/48%	18 financial variables	Yes	GA, SVM, LR	-	ACC	Yes

Table 2.1: Comparison of related studies (cont.)

No.	Study	No. of firms	Active/Failed	Number of variables	Variable selection method	Classification techniques	Ensemble approach	Performance measures	Significance test
21	Wang *et al.* (2014)	240	50%/505	30 financial variables	Yes	LR, NB, DT NN, SVM,	Bagging, Boosting	ACC, type 1 & Type II Error	-
22	du Jardin (2015)	16880	50%/50%		-	ANN, MDA, LR	-		-
23	Iturriaga and Sanz (2015)	772	50%/50%		Yes	SVM, ANN, MDA, LR	-		-
24	Barboza *et al.* (2017)	13433	98%/2%	11 financial variables	-	LDA, LR, NN, SVM, RF	Boosting, Bagging	AUC, ACC, Type I and Type II Error	-
25	Jones (2017)	35939	87%/13%	21 financial variables	-	LR	XGBoost[5]	AUC, ACC	-
26	Fan *et al.* (2017)	626	50%/50%	16 financial Ratios	-	LR, SVM, ANN, RF	Gradient Boosting-DT	ACC, Type I & Type II Error	-
27	Choi *et al.* (2018)	385	87%/13%	21 financial variables	-	SVM, DT, NB, LR, KNN	Voting-based ensemble	ROC, AUC	-
28	Jing and Fang (2018)	293	-	18 financial ratios	Yes	LR, NN, SVM, KNN	-	AUC, ACC,	-
29	Bešlić Obradović *et al.* (2018)	126	65%/35%	24 financial variables	Yes	LR	-	ACC, specificity, sensitivity	Yes
30	Veganzones and Séverin (2018)	1500	95%/5%	50 financial variables	Yes	LDA, LR, NN, SVM, RF	-	Sensitivity, AUC	-

Table 2.1: Comparison of related studies (cont.)

No.	Study	No. of firms	Active/Failed	Number of Variables	Variable selection method	Classification techniques	Ensemble approach	Performance measures	Significance test
31	Huang and Yen (2019)	64	50%/50%	16 financial variables	Yes	SVM	XGBoost,	Type I & Type II error, ACC,	-
32	Matin *et al.* (2019)	278047	97%/3%	50 financial variables	Yes	LR, NN,	XGBoost,	AUC	-
33	Son *et al.* (2019)	977940	23137	9 financial variables	Yes	LR, NN, RF	XGBoost,	AUC, ACC	-
34	Uthayakumar *et al.* (2020)	43405	95%/5%	64 financial variables	Yes	-	-	ACC, sensitivity, specificity, FPR, FNR, error rate	-
35	Bateni and Asghari (2020)	174	50%/50%	23 financial ratios	-	GA, LR	-	ACC	-
36	Smiti and Soui (2020)	5910	93.06%/6.94%	64 financial ratios	-	KNN, DT, SVM, ANN, DT, DPL	-	ACC, AUC, sensitivity, specificity	Yes
37	Jabeur *et al.* (2021)	133	50%/50%	18 financial ratios	Yes	LDA, LR, SVM, ANN, RF, GBM[6], DNN[7],	XGBoost, CatBoost[8]	ACC, area under the ROC	Yes

[1]CBR: Case Based Reasoning, [2]MDA: Multiple Discriminant Analysis, [3]GA: Genetic Algorithm, [4]RF: Random Forest, [5]XGBoost: Extreme Gradient Boosting, [6]GBM: Gradient Boosting Machine, [7]DNN: Deep Neural Network, [8]CATBoost: Categorical Boosting Machine

The summary of the studies' findings allows assessment of what has been achieved in the field of business failure prediction, including methods and data sizes used for designing and modelling, and the extent to which these results can help to investigate and apply new techniques that are not fully covered in the area of business failure. As a result, this study is guided in developing a classification and prediction model based on new different aspects, based on adopting new modelling approaches that have never been used in the field. It should be noted that this thesis focuses on proposing *new* classification techniques in the field of business failure prediction in the UK, rather than comparing results with other related studies.

Based on information from Table 2.1, different findings and conclusions can be derived about business failure models. The first concerns the data class distribution used for developing the model. A total of 17 studies out of 38 used balanced dataset with the same number of failed firms as active (e.g., Li, 2011; Heo and Yang, 2014; Huang *et al.*, 2014; Zhou *et al.*, 2014; Jabeur *et al.*, 2021; du Jardin, 2021). Other studies that used imbalanced datasets could suffer from imbalanced classification in their results, which could lead to bias in model classification results, because of the skewed class distribution on the data. This is exemplified when classifying binary two classes, whereby most data related to a specific class represents a normal case in the domain, and only a few other classes represent an abnormal case. As the distribution of classes is not balanced, most machine learning algorithms perform poorly and need modification to avoid predicting the majority class in all cases. Another disadvantage is that model performance measurements could lose their meaning, and alternate evaluation metrics are required such as ROC area under curve.

A substantial step in model development is data pre-processing and the feature selection process used to select the best prediction variables for the model. All of the studies used financial ratios variables as explanatory features or attributes of business failure models. However, it is crucial to select the most relevant features amongst these datasets, and it is important to clean the dataset by removing outliers, noisy, and irrelevant or redundant features in order to improve model classification performance. According to Tsai (2009), cleaning data and filtering it from irrelevant and redundant information can increase the performance of the model, even though this might be time- and cost-consuming. The most commonly used feature selection method used in the literature was stepwise method, which was deployed in 10 out of 37 articles.

The second finding regards the data partitioning or splitting method used to train and validate or test the model. Two main techniques were used to split the datasets: hold-out splitting and K-fold cross-validation. The hold-out splitting method requires dividing the dataset into two parts, one part for training the model and the other for testing and validating. For example, the data can be divided into 70% for training and 30% for testing and validating. On the other hand, the K-fold cross-validation splitting techniques divide the dataset into K number of subsets, also called the number of folds, whereby each fold contains an equal size of data, and K cannot exceed the size of the dataset. Moreover, there are other methods of data splitting techniques, such as leave-one-out and repeated hold-out methods (Garcia *et al.*, 2015). However, according to Garcia (2015), the researcher's preferences determine the selection of the most suitable splitting technique.

The third finding and the most important stage in developing prediction models is the type of classifiers used to build the model. As can be noticed from Table 2.1, the number of business failure classifiers using machine-learning techniques varies among studies. The purpose of deploying different classifiers is that all studies aimed to introduce new modelling techniques while comparing them with other methods within the same study and previous works in the area of business failure. In general, the aim is to prove a model's validation and superiority over other methods or classifiers, to achieve the highest performance of correctly predicting business health status. However, there is no general superiority of one model over another, since each study used different datasets, data-splitting methods, and performance measurements, which must be taken into consideration when comparing modelling results from different studies.

Hybrid models to enhance prediction performance are believed to be superior in achieving better classification results when applied in the business failure field, as discussed previously. The modelling design of these techniques is by integrating different classifying methods in accordance with deploying different feature selection and data filtering methods in order to enhance and exploit single classifiers' strengths while mitigating their performance weaknesses.

Some studies proposed business failure models using ensemble methods as an experimental modelling approach. Using ensemble learning relies on combining several classifiers, such as using different DT structures as ensemble members with different types or parameters in order to train the dataset. The resulting model consists of pooled members after the ensemble

strategies applied to get the final output. According to Wang *et al.* (2011), it is crucial to build an ensemble model that is diverse and accurate.

A substantial step after developing business failure models is the evaluation of its performance, by assessing the classification capability of the model when applied on new data. Most of the studies used average accuracy rates as an indicator of how accurate the model in classifying business health is, accompanied with Type I and Type II Error and sensitivity and specificity ratios. Another performance measure is the AUC, which calculates model performance at various threshold settings, whereby higher AUC values represent better model performance in distinguishing between failed and healthy firms.

The final step after developing business failure model's performance is to investigate the reliability and the robustness of the model. Using statistical tests on models' results is an important step to determine whether model outputs are causative, and not coincidental. However, only seven of the 37 studies employed statistical significance testing. It is worth mentioning that the type of test depends on the number of classifiers needing to be compared.

As can be inferred from Table 2.1, the main steps used to develop a business failure model are as follows:

- Collecting the dataset: most studies collected financial information related to each business presented, mostly with financial ratios.
- Splitting techniques: choosing the best splitting technique for data and considering the size and the distribution of the dataset (e.g., majority and minority classes).
- Modelling techniques: these depend on the manner in which the developer or researcher tries to solve the problem at hand. The main objective is to create an effective model with reliable results. Others try to develop new ideas using either hybrid models or ensembles. Therefore, this is generally determined from the perspective of the researcher.
- Selecting appropriate performance measures: among the many performance measures available, the researcher should select the most suitable measures to reflect every angle of the model performance.
- Statistical analysis: to reach a reliable conclusion, a developed model should be validated statistically, using an appropriate test.

2.9. Summary

This chapter reviewed related literature to present the theoretical background of business failure and its related issues, based on modelling approaches used in previous studies. The first part revealed the concept of business failure prediction in terms of definition, importance, and the history of business failure model development and implementations. It discussed the benefits of early predictions of business failure for the economy, and how it become an important topic for firm stakeholders. Moreover, it discussed the importance of the topic for the machine-learning community, and how it become an important area to investigate by researchers in the field.

The source of information used for predicting business failure, such as financial statement data and financial analysis tools, have been explained. The financial data about firms' operations provided by management and delivered to stakeholders such as investors, creditors, suppliers, government agencies and other external users is considered to be an informative communication language about business health. The reporting system includes the preparation, quarterly and annually, of the financial department's financial statements, together with other departments within the company, including the quarterly and annually representatives of the business, interests of the company, management assumptions and financial positions. Moreover, the accounts show the accounting and financial policies of the company complying with its management assumptions and estimates used to report its resources, transactions, and liabilities. As a core financial reporting system, the numerical disclosures quantify the firm's financial position by ensuring that its business models, resources, source of financial support, and activity are transparent and verifiable. Therefore, the aim of financial disclosures is to provide users with materials of high-quality information that can be used effectively to monitor ongoing companies, which can thus be used for business failure modelling, which relies heavily on this data to predict business continuity.

Subsequently, the chapter focused on the quantitative tools used by researchers to develop business failure prediction models. An overview of the several methods used for modelling ranging from traditional statistical methods to machine learning was highlighted. After considering the different modelling techniques used in business failure studies, a substantial concern is established on the superiority of machine learning methods over traditional modelling that in most cases, machine-learning models achieve higher performance. Moreover, model performance evaluation was explained in relation to checking model reliability and robustness.

The last part of the chapter reviewed literature collected from recent key studies that deployed machine-learning techniques for model development. A systematic analysis followed with a summary of the studies was conducted, focusing on several important factors used for modelling, such as the size of datasets, number of variables used, data-splitting, type of modelling technique, and performance measurements used to evaluate the model performance. Findings and conclusions were revealed during the analysis, highlighting the literature gap of the dearth of studies considering classifier combinations, ensemble selection, and incorporating new classifiers.

Chapter 3
Research Methodology for Proposed Business Failure Prediction Model

3.1. Introduction

This chapter explains and discusses the main steps used in constructing the proposed business failure prediction model, from the data collection to analysis stage.

Firstly, it presents an overview of the data collection process in terms of data types and resources, and the commonly used data pre-processing method adopted in this thesis. Furthermore, data splitting techniques are explained and discussed, with justification for the selection of such tools.

Secondly, the chapter explains the modelling techniques using statistical and machine learning methods, including the theoretical background behind each method.

Thirdly, the performance measurements used to validate and evaluate model's performance are presented.

Finally, extra validation using statistical significance testing is explained.

3.2. Data Collection and Processing

3.2.1. Data Collection

In the field of business failure prediction, data collection is the most fundamental step to execute developed models. All previous studies extracted financial information from firms' financial statements in the form of ratios, which play a major role as indicators of business health, and which have been ubiquitously deployed in business failure modelling. Hence, this study follows in the steps of previous studies and collects financial ratios related to different companies from all industries in the UK.

The performance and reliability of any classifier in binary classification problems are dramatically affected by the data balance ratio, and bias toward any class in the classification must be avoided, as explained previously. For instance, when using extreme imbalance datasets, the classifier will always predict the class of the majority, at the expense of the ignored minority class. This has significant implications for accuracy and performance for minority

class prediction. For example, for an imbalanced dataset consisting of 90% active and 10% of failed firms, the developed classifier can still achieve an average accuracy rate of 90%, due to correctly classifying all active firms, and misclassified all failed firms in the dataset. In other words, the model achieves high accuracy at the expense of reliability (Aljawazneh *et al.*, 2021).

In this thesis, the data were extracted from the FAME website, which is a financial information database of 7 million companies in the UK, updated on a daily basis, with up to 10 years of history (FAME, 2019). It provides detailed firm information, including financial statement data and pre-calculated financial ratios related to firms' financial performance, as shown in Table 3.1.

Table 3.1: Number of firms included in the datasets

	Datasets		
	No. of Firms	Active Firms	Failed Firms
Year 2019 Dataset	20,000	10,000	10,000
Year 2018 Dataset	20,000	10,000	10,000
Year 2017 Dataset	20,000	10,000	10,000
All-Datasets	60,000	30,000	30,000
Dynamic Modelling Dataset for 5 Years	20,000	10,000	10,000

For business failure classification problems, three separate datasets were collected, including all businesses in the database that failed during the years 2017, 2018, and 2019, with a matching number of businesses that still operating during the same years. To achieve the best classification results, both active and failed firms were selected and matched based on the same number of firms from the same industry, with relatively similar market capitalisation (business size).

On the other hand, for developing dynamic time series one step ahead prediction models, data were gathered over five consecutive years for the same businesses entities, to render financial variables as insensitive as possible to any short-term variations that may occur within the company's economic and financial environment. Hence, we used the business in the dataset of the year 2019 and collected five datasets in a time manner series, representing the financial ratios for the studied years:

- t - 4 (2015)

- t - 3 (2016)
- t - 2 (2017)
- t – 1 (2018)
- t year (2019)

This collection process allows the development and design of prediction models based on a time series manner.

3.2.2. Data Pre-Processing

A critical point in developing a prediction model is having a dataset with high quality, to allow model generalisation. This essentially relies on the importance of model attributes and the freedom of data from outliers and missing values. Accordingly, data pre-processing plays a major role in solving business failure classification problems (Alasadi and Bhaya, 2017). Datasets collected for a large number of firms from a real-world database may involve completely raw data, containing noisy and missing values. Therefore, it is an essential step in model development to have a dataset free from irrelevant, redundant, unreliable, or noisy attributes before any further analyses or procedures, as this makes knowledge discovery and predictions easier and more reliable. This can be done using several methods, such as data imputation, feature selection, data normalisation and deleting data that contains outliers. Once data pre-processing is performed, a new dataset is ready to train the proposed models. The following subsections present the pre-processing methods used to make data ready for model development in this study.

3.2.2.1. Data Imputation

When collecting businesses financial ratios datasets, there are missing and incomplete values in some attributes that should be taken into consideration when training the model. In order to overcome this problem, there are two approaches adopted in this study. The first step is to delete instances that contain a large volume of missing feature values. The other way is by adopting an imputation method, such as replacing the missing values with new values based on some estimations. In this study using datasets collected from UK firms that contain some missing values, both approaches were used to make datasets ready for training the classifiers.

3.2.2.2. Data Normalisation

Data normalisation is used for model development containing values out of the range of 0 to 1, which for some classifiers could be an issue when training the model. In this case, data

normalisation becomes an important step to avoid any bias in the data, and accordingly feeding the classifiers with the right attribute values. For example, NN and SVM modelling techniques require input variables to be in the range of 0 to 1; in order to achieve that, attributes can be normalised and transformed into values in the required range using an appropriate normalisation method (Alasadi and Bhaya, 2017). Moreover, data normalisation improves classifier learning performance from the data by removing any outliers in the data, and removing the presence of any controlling variables (Singh and Singh, 2020). For our datasets, min-max normalisation technique was adopted, which requires taking the highest value of an attribute and giving it a value of 1, while the lowest value is given 0, and other values of the attribute are computed with the following equation:

$$new\ value = ((original - min)/(max - min)) \times \left(\underset{new}{max} - \underset{new}{min}\right) + \underset{new}{max} \qquad (3.1)$$

3.2.3. Feature Selection

Selecting a group of attributions that have higher prediction information is extremely important in a wide range of research disciplines, including in business failure modelling. Reducing the amount of unnecessary or redundant features reduces the working time of a learning algorithm dramatically and provides a more general approach. The possibilities of feature selection facilitate the viewing and comprehension of data, reduce measure and storage requirements, reduce training and usage durations, and defy dimensionalities to increase prediction performance (Falangis and Glen, 2010). This also helps in understanding the concept underlying classification in the real-world.

The selection process is usually carried out before models are trained. Feature selection has the advantage of reducing overfitting by removing unnecessary data from the model, allowing it to only focus on the relevant features, and not getting bogged down on irrelevant features (Guyon and Elisseeff, 2003). Removing irrelevant information improves the model's predictions by reducing errors, and reducing the time required to produce the model. The interpretation of a model is easier when there are fewer features. Feature selection is thus vital to achieve accurate prediction for any value, as well as efficient processing.

Adopting a selection method provides a starting point when there is no intuition about the dataset and what features are important for the model. It also allows effective selection of the most significant features from a big size of data. However, a disadvantage that could result from adopting this method is that it does not run through every single combination of features,

which prevents optimal model outcomes. Another disadvantage is models with high multicollinearity between features, due to potential relationships among these features, which can negatively affect the model classification and prediction accuracy (Guyon and Elisseeff, 2003).

Wrapping methods are used to evaluate the importance of each feature to be included in a certain subset that allow the prediction model to achieve higher prediction performance. The process includes iterating and trying different features and subsets until the optimal subset is reached. However, two disadvantages can be taken into consideration when adopting this method: it consumes large computation time when the number of features is high, and it might overfit the model when the size of data points is low. There are three main wrapper methods used for feature selection: forward selection, backward selection, and stepwise selection method.

Forward selection method starts the first model with zero features, and then for each single feature it builds up a model and determines the p-value associated with the t-test. After calculating the p-value for each feature, it selects the one with lowest p-value measurement and adds that feature to the working model. The next step is running the model with the selected first feature and added another feature that has the lowest p-value. New features with the lowest p-value continue to be added to the model until all features with significant p-values are added. The final model thus contains all of the most significant features, and any features with insignificant p-values are excluded by default.

In contrast, in backward selection method, the model starts running with all available features in the dataset, then it computes the p-values associated with the t-test or F-test for each feature of the model and removes features with the most insignificant p-values. The iterations continue until all insignificant features have been removed from the model.

Stepwise selection is a hybrid selection method consisting of using both the forward and backward selection methods. Similar to forward selection method, it being by selecting a model with zero features, and starts adding the first feature based on the rule described above. It then adds the next feature with the lowest significant p-value to the model. When performing the third iteration, a third feature is added to the model with the lowest p-value, while any feature with an insignificant p-value is removed from the model, and so on for the rest of the dataset features. This results in a model which includes all significant features, which comprises the final feature subset. In this thesis, stepwise selection method was selected to determine the

most influential and relevant attributes to construct the proposed classification and prediction models (Table 3.2).

Table 3.2: Selected variable using stepwise method

Independent Variables	Ratio number
Liquidity ratio	5
Operational ratio	7
Profitability ratio	8
Solvency ratio	3
Non-financial ratio	1

3.3. Data Splitting

After all data have been pre-processed, the dataset is ready for training and testing the classification models. In this stage, data splitting is used to partition the dataset into training and testing datasets; the former is used to train the model, while the latter is used to evaluate and validate model performance. This process is considered to be a fundamental step in model development and evaluation. After training the model using the training dataset, the testing dataset is used to evaluate how well the model will perform with real-world datasets. An important aspect when splitting dataset is the size of each data-split. Therefore, more data instances in the training dataset results in a more fitted model, and more data in the testing dataset results in better accuracy estimation, which enhances model reliability.

Another important issue that has a great effect on model performance when splitting data is to have data splits that fairly represent each class of the dataset for both training and testing. Balanced data distribution plays a major role when training the model with different classes, and results in a good generalisation over the testing dataset. For solving business failure models, there are two main data splitting techniques used by researchers in the field: K-fold and hold-out techniques.

3.3.1. K-fold

Using this technique requires partitioning the original dataset into K-subsets, also referred to as folds, which are used to train and test the model performance. For example, a dataset can be divided into equal number of partitions in the form of S_1, S_2, S_3, …., S_n, where n is the number

of subsets. Thereafter, each subset is individually trained and tested using the classification model applied. Table 3.3 illustrates the K-fold cross-validation process.

Table 3.3: K-fold cross-validation

Partitions/Folds	**Training set**				**Testing set**
1	S_2	S_3	S_4	S_k	S_1
2	S_1	S_3	S_4	S_k	S_2
3	S_1	S_2	S_3	S_k	S_3
4	S_1	S_2	S_3	S_k	S_4
5	S_1	S_2	S_3	S_4	S_k

The process of K-fold cross-validation relies on using one partition of the dataset for testing and the rest for training the model. The process is iterated until all of the subset is trained and tested. To calculate the model performance, the final prediction accuracy is estimated by taking the average performance of all folds used for testing. An advantage of using this approach is that it ensures the use of all data available, thereby preventing overlapping. Moreover, due to having many data for training and testing, K-fold method allows for multiple repetition of the process, which results in more robust and efficient model performance.

An important issue that could arise when adopting this approach is the optimal number of folds or partitions to put data in. For example, a high number of folds could result in better performance of the model in terms of accuracy classification, but on the other hand, it results in high variance, while a small number of folds could enhance variance measurements, but the model performance might be biased. According to García *et al.* (2015), the optimal number of folds depends on the size of the dataset, whereby 5 or 10 folds can work in favour of models with differently sized data. Another important issue is the repetition of the process, to minimise variances as much as possible, by using different subsets of training and testing data. In this thesis, K-fold cross-validation was adopted and applied to the dataset to validate and test the model's performance with 10 repetitions using 5-fold cross-validation, resulting in a total of 50 tests to achieve reliable and robust conclusions about the model's performance.

3.3.2. Holdout Technique

Holdout technique divides the dataset into two separate parts: one part is used to train the model, and the other to test and validate its performance. Due to its simplicity, this technique has been widely adopted by researchers in the field of business failure prediction. The most

frequent way to apply this method is by dividing the dataset into an 80% subset for training the model, and the remaining 20% for testing. Another ratio that can be used is 70% training subset and 30% for testing. In this study (80%, 20%) technique is selected. However, the holdout method could result in biased model outcomes, and both training and testing subsets might be unrepresentative, which reduces the scope of practical application (Bischl *et al.*, 2012).

In order to overcome this issue, multiple repetitions of holdout technique can be performed to avoid any bias on the representable dataset. Applying this can reduce the probability of getting a favourable testing subset, and lead to more robust model outcomes.

3.4. Modelling Techniques

This section explains the modelling approach used to develop the business failure classification and prediction models. The approach began with individual classifiers that have been used widely in the literature due to their classification outperformance over statistical methods (as explained previously). To justify this experimentally, individual classifiers were first built using several heterogeneous classifiers, trained and tested using financial data of UK firms adopted in this thesis. Different hybrid and ensemble classifiers were then developed using data extracted from already trained individual classifiers, to enhance their output performance. The final stage was to develop a dynamic classifier that can predict data and business status in advance to solve business failure problems. The following subsections explain the background behind every modelling technique.

3.4.1. Individual Classifiers

The state-of-the art of each individual modelling approach in the field of business failure, from statistical to machine-learning modelling methods, are overviewed below. Many techniques have been proposed by researchers to build well-performing classification and prediction models. According to Heneley (1997), in order to select the best classifiers, the size of the dataset and the type of variables have to be considered in terms of the optimum model fit and performance. Following the steps of the studies in the literature, statistical methods such as LD and LR are used as benchmark classifiers, and their prediction performance is used to compare other machine-learning modelling performance.

3.4.2. Linear Discriminant Analysis

This classification method was first used by Fisher (1936) as a parametric statistical modelling approach to classify between two classes in the dataset population. It has been widely used in

the field of business failure prediction to classify failed companies (Laitinen, 2007). To solve business failure classification problems, assume there is a dataset of n companies, where each company has a certain L number of attributes or variables in the form of (x_1, x_2, x_3, …, x_L) that are used to classify firm's activity status y in a binary classification system (active/1 or failed/0). The main objective of the model is to estimate the probability of a firm to be categorised as either 'fail' or 'active' p(y/x), based on all variables on the dataset. By linearly combining all variables in the dataset, firm status can be classified in its appropriate class as expressed in the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3.2}$$

where y represents the discriminate score of firm status, β_0 is the model intercept, and β_i represents the coefficient related to each variable x of the model. To better explain how the above discriminate model works, the values of the variable's coefficient are adjusted based on the covariance and the mean values of both classes on the dataset. After training the model, the best coefficients values are assigned to each feature of the dataset, whereby the final discriminate score can be calculated for each company (Rafiei *et al.*, 2011). Finally, the calculated score is compared to a threshold in order to classify the firm status as failed or active. However, this modelling technique is considered to have some classification limitations when solving problems that have non-linear relationships between dataset variables (Veganzones and Séverin, 2018).

3.4.3. Logistic Regression

Logistic regression is the most frequently used statistical method to solve business failure problems. Unlike LDA and linear regression models that give continuous output values, LR classification method was developed and used in business failure studies to solve binary classification, in which the final output can be characterised by 0 if the company is failed or 1 if it still active (Vuran, 2009). Logistic regression requires less restrictive statistical assumptions to ensure that all the problems discussed with regard to discriminant analysis are essentially avoided with a logit analysis.

For solving business failure classification problems, the probability of a firm to be classified as failed or not is the result of the relationship of the independent variables and the firm status based on a logistic curve. The result is an s-shape curve where all of the output values are

between 0 and 1, representing the relationship between the independent variables and the output binary classification, by adopting a non-linear function as in the following equation:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{3.3}$$

where p is the probability of target firm status, β_0 is the intercept term, and β_1 represents the coefficient related to the features *X*. For solving business failure prediction problems, if the probability of failed company is p, then the probability of active company is (*1-p*). This concept is referred as odds, calculated as the ratio of probability of having failed firm relative to the probability of having an active firm. Odds can be expressed as follows:

$$Odds = p/(1 - p) \tag{3.4}$$

Thereafter, the odds ratio can be used to calculate the curve equation:

$$p/(1 - p) = exp^{(\beta_0 + \beta_1 x)} \tag{3.5}$$

However, in the above equation, the left-hand side can take values between 0 and 1 while the right-hand side can take any value. This distraction can be solved by taking the natural logarithm of both sides of the equation as in the following:

$$ln\,[p/(1 - p)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{3.6}$$

Once the odds of the logit function are known, the final step is to find the probability in a range between 0 and 1 as follows:

$$p = odds\ /\ (1 + odds) \tag{3.7}$$

In contrast to LDA classification method, in LR method, data do not necessarily have multivariate normal distribution. However, a drawback of this method is it reliance on a full relationship between the predictable variables in relation to the logit of the target variable (Lee and Chen, 2005).

Despite its classification disadvantage, LR classifier has been widely used in the literature in the field of business failure, where in some cases it is been used as a benchmark classifier (Gordini, 2014; Barboza *et al.*, 2017; Jing and Fang, 2018; Veganzones and Séverin, 2018; Matin *et al.*, 2019; Son *et al.*, 2019; du Jardin, 2021).

3.4.4. Artificial Neural Network

ANN is as a computational or mathematical modelling tool of non-linear data, constructed based on an emulation of the biological neural system (human brain function). The concept of ANN mimics the neural system, whereby a group of artificial neurons are interconnected to process information using a connectionist approach to computation, and to learn and adapt form historic data (Veganzones and Séverin, 2018). ANN has been used in modelling business insolvency and financial distress as an alternative to traditional statistical methods (Matin *et al.*, 2019; Son *et al.*, 2019; Smiti and Soui, 2020; du Jardin, 2021).

An ANN has to be configured such that it achieves the desired set of outputs from the application of a set of inputs. As shown in Figure 3.1, ANN topology consists of three layers: input, hidden, and output layers. The structure of the business insolvency model starts by feeding financial variables, as an attribute of each firm, to the input layer for processing them, after which they are processed further in the hidden layer. Finally, after thorough processing of the data, the values are sent to the output layer to give an answer on whether a firm is healthy or is going to fail. Weights are assigned to each attribute to calculate the output based on their relative importance; these weights can be adjusted based on supervised learning rule, by which input and output data are fed to the model to be used in training. All of the weights are summed together using transfer function (i.e., *sigmoid* or *tansig*) in order to predict output. The process of adjusting weights is repeated iteratively to minimise the error between predicted and actual output.

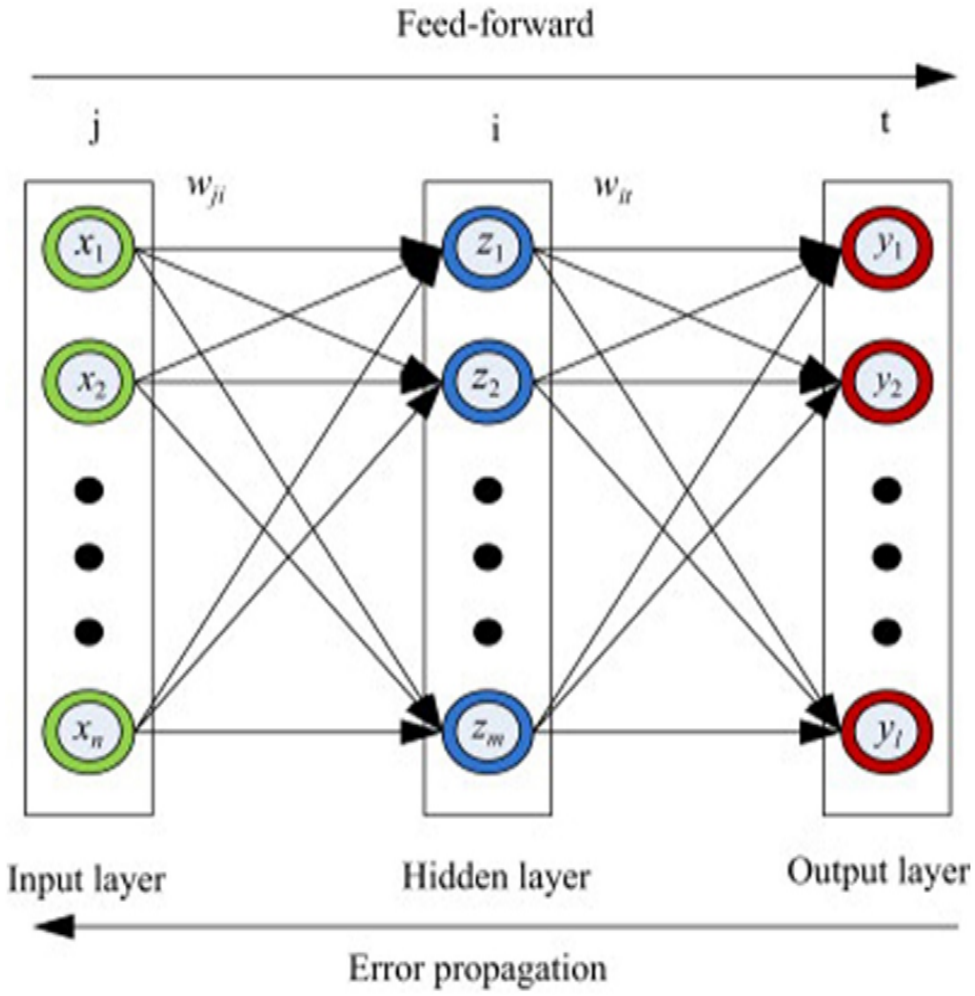

Figure 3.1: NN feed-forward back propagation topology

Source: Alhunaity and Abbod (2020)

To explain how the model works to solve business failure prediction problems, assume there is a training set consist of number of attributes x = {x1, x2,..., xn}; the feed-forward back propagation starts by feeding these data to the input layer. A random initial weight is given to each attribute, which is then fed to the hidden layer, then an activation function is applied to process all data inputs. After processing the data and assigning all attributes new weights, the weighted inputs are linked to the output layer, where a further activation function is applied to lead to the final output. All neurons in the network hidden layer are companioned using the following function:

$$n_i = \int H\left(\sum_{j=1}^{N} x_j . w_{ji}\right), i = 1,2,\ldots,S \tag{3.8}$$

where n_i is the output of the hidden layer, f H is the activation function used, and the most activation function used is the sigmoid function. Thereafter, the new attributes with adjusted weights are passed to the output hidden layer to calculate the final output values:

$$y_t = f\,t\left(\sum_{i=1}^{s} n_i . w_{it}\right), t = 1,2,\ldots\ldots,L \tag{3.9}$$

where y is the final output representing the model's final decision about the data. An important stage of the training process is that if the difference between the final output values and the actual target values is significant, the weights computation process is repeated again, and all weights between the input and hidden layer and the weights between the hidden and output layer are updated until the differences are minimised.

The core advantage of ANN is its capability to find relationships between variables and deal with nonlinearity, choosing among the most important predictable attributes in the model (Messier and Hansen, 1988). Moreover, it has the capability to model using incomplete data sets with missing and noisy data with no previous assumption of data distribution, which allows it to recognise complex patterns between attributes (Vellido *et al.*, 1999). However, its long training process and lack of theoretical grounding are considered as drawbacks of ANN modelling technique.

In this study, we used feed forward backpropagation as an ANN modelling method for classifying failed firms.

3.4.5. Decision Trees

DT modelling techniques are commonly used machine-learning approaches that have been widely deployed in business failure applications. DT is a non-parametric classification method that analyses target data using a function of independent attributes (Tsai *et al.*, 2014). The conceptual idea behind using DT techniques in business failure is to classify businesses into a binary classification system. It starts with a root node that includes both classes of business' status, and then it splits into another two nodes containing the possible event based on the chosen attributes by applying a decision algorithm. In the DT structure, the leaves are marked by class labels while branches are marked with conjunctions, which lead to classifications (Sharma and Kumar, 2016). The process continues in a loop of all possible splits until optimal DT is reached, optimally partitioning the mostly active and failed firms with the lowest error

and misclassification rate. This classification method has been used widely by researchers of business failure (Tsai and Cheng, 2012; Heo and Yang, 2014; Tsai *et al.*, 2014; Wang *et al.*, 2014; Smiti and Soui, 2020). Larivière and Van den Poel (2005) adumbrated the main advantages and disadvantages of DT. The advantages are that it:

- Does not make assumptions based on data distribution.
- Allows for numerical and nominal attributes.
- Handles missing data or values.
- Is easy and interpretable.

Its main disadvantages are that it:

- Has less model robustness and optimality of performance.
- Is sensitive to data outliers and irrelevant attributes.
- Requires extreme efforts to handle missing data

Figure 3.2 displays the structure of a DT.

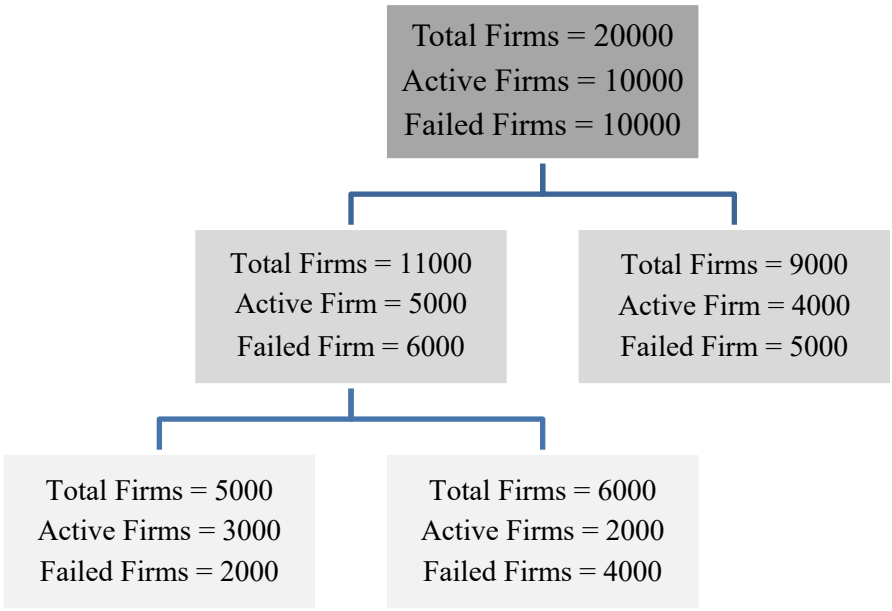

Figure 3.2: Example of a Decision tree structure

3.4.6. Naïve Bayes

NB is a statistical classifier used to predict binary class problems. The classifier relies on adopting the Bayesian theory to classify the final output of the model when the input variables space is high (Choi *et al.*, 2018). The simplicity of the method in regard to binary classification decisions means that is has gained little attention in the field of business failure prediction (Wang *et al.*, 2014).

NB modelling starts by calculating the conditional probability of a class or instance being classified, represented by independent variables *x*. For solving business failure classification problems, assume that there is a training dataset containing a number of variables in the form of { x1, x2,….., xn}, assigned with a class label *m* indicating either 'fail' or 'active'. The NB classifier trains the data and finds the mapping function that can predict the probability of a class based on all features on the dataset using the following the computational method:

$$P\ (c_i|\ x_1, \ldots\ldots, x_n) = \ P\ (x_1, \ldots\ldots, x_n \mid c_i) * \ P\ (c_i) / \ P(x_1, \ldots\ldots, x_n) \qquad (3.10)$$

where $P(c_i \mid x_i \ldots, x_n)$ is the conditional probability of a business class based on all variables *x*. In contrast, $P(x_1, \ldots\ldots, x_n \mid c_i)$ is the probability of the variables belonging to firm class. $P(c_i)$ is the prior probability of class unconditioned to any data, and $P(x_1, \ldots\ldots, x_n)$ is the probability of all variables on class c_i. The model uses minimum error probability criterion or maximum posterior probability to assign the posterior probability to the label class.

3.4.7. K-Nearest Neighbour

KNN is defined as a type of non-parametric classification method, and it is a popular data mining technique to solve classification problems (Choi *et al.*, 2018; Smiti and Soui, 2020). The KNN algorithm analyses all available data and classifies it, and then uses the classifications of the previously established categories to determine how the new cases should be classified. The first step is to count the number of closest neighbours and see if the class assignment is correct. To determine the distance between an instance and all training instances, the model first needs to know the size of each training instance. The instances are ordered according to their distance from one another, and their nearest neighbours are discovered. The process of finding the shortest path is therefore equivalent to moving from the new instance to the starting point. Nearby neighbourhoods are then visited to see if they can find the majority (Smiti and Soui, 2020). This majority of the class is a class that was predicted to be the final result (Choi *et al.*, 2018).

3.4.8. Support Vector Machine

SVM is a supervised learning modelling technique derived from statistical learning theory, which is associated with a learning algorithm used to analyse data in classification and regression. This machine learning modelling method has been widely adopted to develop classification models in the field of business failure (Barboza *et al.*, 2017; Fan *et al.*, 2017;

Choi *et al.*, 2018; Jing and Fang, 2018; Huang and Yen, 2019; Smiti and Soui, 2020; Jabeur *et al.*, 2021).

As illustrated in Figure 3.3, SVM topology has a decision surface called the optimal hyperplane, and the data points closest to it and the margins (dashed lines) are called support vectors. The support vectors are very important elements used for training the model, whereby the SVM finds the optimal hyperplane separating the input data with the maximum margin width to fit the data. SVMs is a non-probabilistic binary linear classifier that classifies a set of training to one of two categories (Huang and Yen, 2019). Kernel function mechanism can be used for the mapping process, by which the SVM can be adapted to become a nonlinear classifier using nonlinear kernels. Several binary SVM classifiers can be combined to convert SVM into a multiclass classifier.

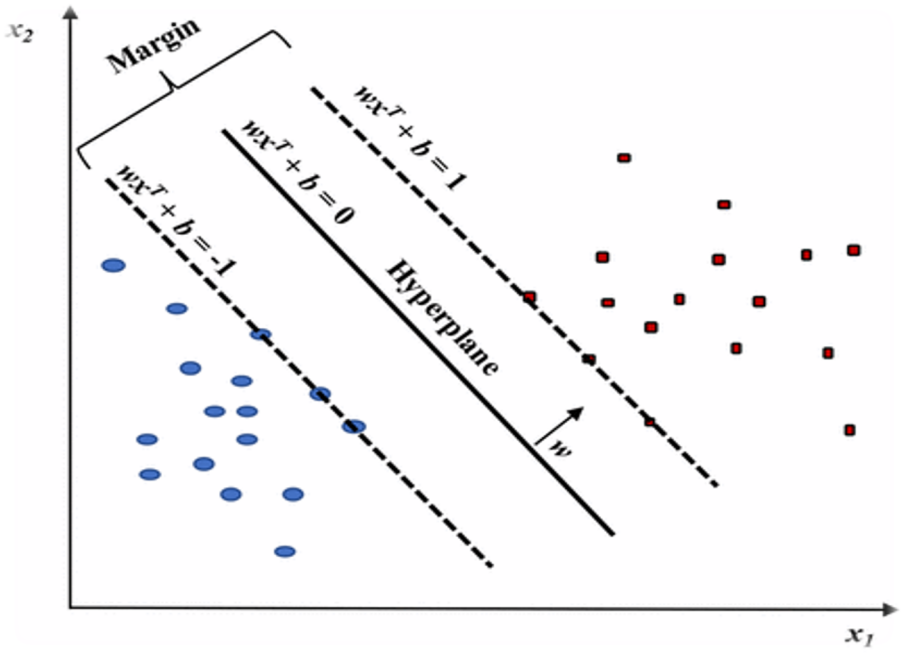

Figure 3.3: SVM model

Source: Huang *et al.* (2018)

SVM has the ability to model nonlinear data using different types of kernels that transfer data in higher dimensional space to provide higher prediction accuracy rates in comparison with other modelling techniques. However, in contrast to ANN, SVM is sensitive towards missing and noisy data (Teng *et al.*, 2010).

3.4.9. Deep Learning

Deep learning can be used as supervised classification learning model consisting of input, output, and hidden units, where most data processing work is completed. Long Short-Term Memory Recurrent Neural Network (LSTM RNN) is a very deep neural network in the time direction, developed to learn sequence and time patterns from time series or data sequences, and to learn long-term dynamics while avoiding problems of vanishing and exploded gradients (Aljawazneh *et al.*, 2021).

The network is composed of LSTM memory blocks (instead of hidden neurons), which consist of memory cells and gates that replace the hidden layers unit of the RNN. Each cell is mainly configured by three gates: the input, output, and forget gates. These cells and gates play an important role in training long-range dependency while controlling information storage. The memory block consists of one memory cell (c_t) and four gates: input (i_t), forget (f_t), input modulation (g_t), and output (o_t) gates (Jang *et al.*, 2019). Unlike in feed forward NN, deep learning LSTM introduces a directional loop that uses previous information to analyse the current output, as the previous output is related to the current output sequence, and the nodes between memory cells are connected (Jang *et al.*, 2019). Figure 3.4 shows a deep learning LSTM structure.

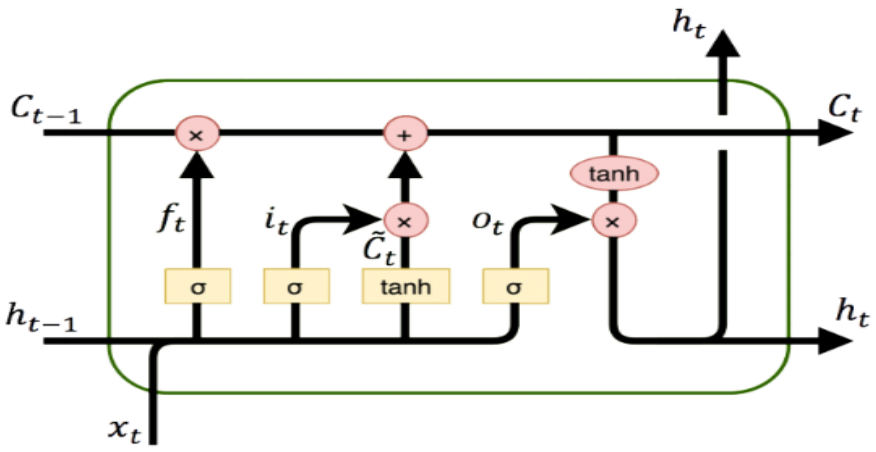

Figure 3.4: Deep learning LSTM structure

According to Figure 3.4, in a standard LSTM RNN network algorithm, the calculated h_t is the final prediction value calculated by receiving input information at time t (x_t) using previous hidden state h_{t-1}, which is represented in the following equations:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3.11)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3.12)$$

$$g_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3.13)$$

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

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o) \quad (3.15)$$

$$h_t = o_t \odot tanh(c_t) \quad (3.16)$$

where σ is the activation function (sigmoid), x_t represents input variables, t is the time unit, W and U represent the assigned weights, i_t is the input gate, o_t is the output gate, b is the bias, c_t represents the vector of memory status, σ is the sigmoid function, g_t represents the input modulation, and ⊙ is a pointwise multiplication.

The first step in the LTSM process is the forget gate procedure, where the sigmoid function is applied to identify information to be discarded (based on value). The next step applies equations 3.13, 3.14, and 3.15, where the sigmoid and the *tanh* functions are used to make the decision to update information from the input values. The final step (equation) is to calculate the final predictions.

In this thesis, the deep learning classifier was used as an individual classifier for predicting business failure one year ahead.

3.4.10. Ensemble Boosting Decision Trees

Ensemble boosting methods combine two or more classifiers to increase the prediction accuracy and improve model performance. They are more proficient and useful to handle the model instability and increased variance between multiple data subsets taken from the

population (Freund and Schapire, 1997). They let the classifier learn from the maximum variance within the dataset, using DT models as the base classifiers to do so.

Boosting is defined as the combination of learning algorithm in series, to improve learner performance from many sequentially connected weak classifiers (Freund and Schapire, 1997). For AdaBoost boosted DT modelling classifier, DTs are weak classifiers, whereby each tree attempts to improve the classifying accuracy and reduce the errors of the previous tree. Therefore, adding many trees in series, with each focusing on improving the classification errors from the previous one, results in a more efficient and accurate classification model. The classifier is designed such that in each step the data distribution is adapted to put more weight on misclassified data, and less weight to correctly classified data, in order to reduce misclassification errors from the previous classification tree. Also, more weight is assigned to stronger classifiers based on their classification performance, and the final classification result is a weighted average of all the weak classifiers. An advantage of adding trees sequentially is that the boosting mechanism learns slowly and performs better. Figure 3.5 shows the ensemble boosting process of the DT framework.

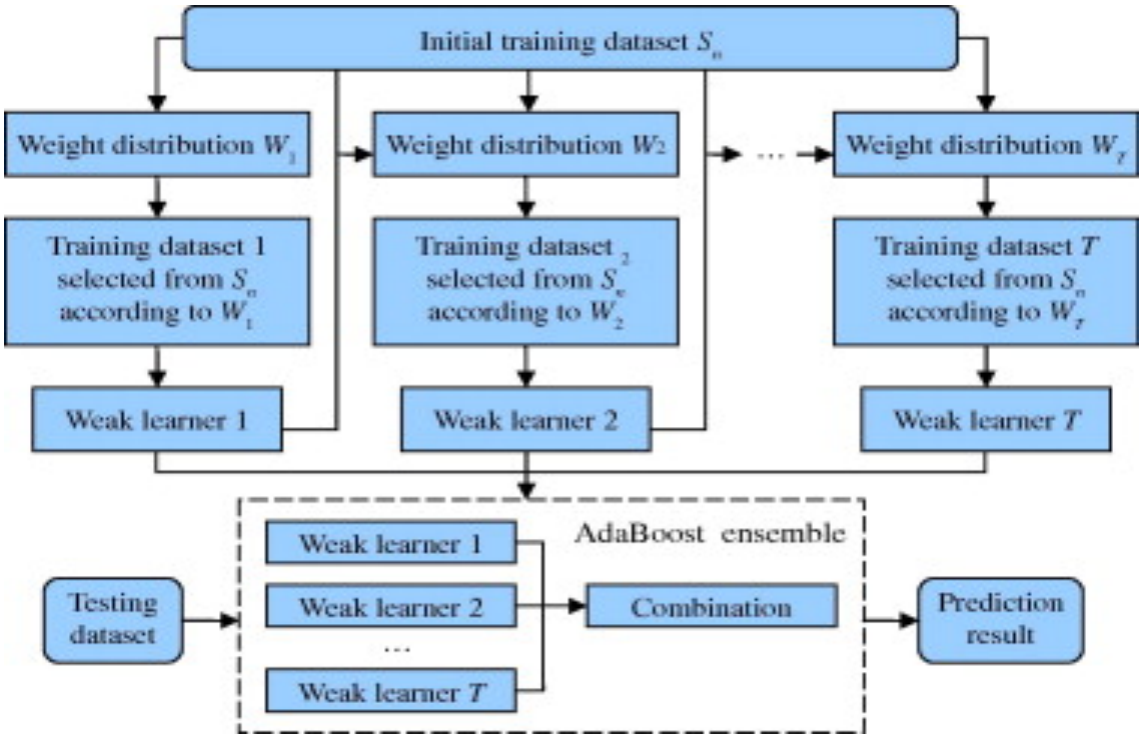

Figure 3.5: Ensemble boosting DT framework

Source: Sun *et al.* (2011)

3.5. Performance Measurements

This section explains model performance evaluation, which is ultimately the most important stage in the development process of business failure models. Each developed model in this study was tested using performance evaluation metrics derived from related studies in order to determine the extent to which these models are reliable and well-learned, to be ready to predict new real-world data. According to Lessmann *et al.* (2015), there are three important types of measure that should be taken into consideration when evaluating model performance results to reach a comprehensive conclusion on how well the developed model performed:

- Measures used to evaluate the prediction power of the model.
- Measures to assess the discrimination power of the developed model.
- Measures used to assess the extent of model accuracy in predicting instances.

In order to evaluate and validate the prediction performance of all models used in this study and to make a robust conclusion on predictive accuracy, eight performance measures were selected that can be integrated from the confusion matrix:

- Accuracy
- Sensitivity
- Specificity
- Type I Error
- Type II Error
- Area under the curve (AUC)
- Reliability diagram
- Brier Score

These measures (as shown in Table 3.4) have been widely used in business failure studies, therefore they were chosen to cover all aspects of model performance in this study.

3.5.1. Confusion Matrix

The confusion matrix is a table used to provide information about model performance results, reporting the number or percentage of correctly and incorrectly classified data. The table consists of information about the number of firms classified as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Several performance measures can be derived from this information, which has been widely used to evaluate model performance. Table 3.4 displays a confusion matrix table.

Table 3.4: Confusion matrix table

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

The following section describes the important performance metrics that can be derived from the confusion matrix.

3.5.2. Average Accuracy Rate

Average accuracy rate measures the percentage of correctly classified data instances out of the total number of the cases on the dataset. As can be seen in Table 2.2, it is the most commonly used measurement to assess model performance in business failure studies due to its simplicity of calculation. The accuracy rate is calculated using the following formula:

$$\text{Average Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{3.17}$$

However, it does not take into consideration how accurate the model performs in classifying different classes individually. As explained previously, if a dataset is imbalanced, consisting of 90% active and 10% failed firms, the developed classifier can still achieve an average accuracy rate of 90% due to correctly classifying all active firms, despite misclassifying all failed firms in the dataset. In this case, a substitute measurement that can give insight on each predicted class is preferable, in order to see if a model is biased toward a specific class such as Type I and Type II Error, and sensitivity and specificity ratios.

3.5.3. Type I and Type II Error

Based on information from the confusion matrix, Type I Error represents false negative (FP), and Type II Error represents false negative (FN). If a failed firm is misclassified as active, this is considered as Type I Error, and if an active firm is misclassified as failed, this is considered as Type II Error. These measures are frequently used in business failure studies (Barboza *et al.*, 2017; Fan *et al.*, 2017; Huang, Yen, 2019; du Jardin, 2021).

For business failure prediction, Type I Errors are related to financial and economic loss resulting from classifying failed firms as active, with high risk of quitting the market, and Type II Errors are related to potential economic loss that could occur from classifying an active firm as failed. Type I and Type II Errors can be calculated using the following:

$$\text{Type I Error} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{3.18}$$

$$\text{Type II Error} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{3.19}$$

Practically, a business failure model that is better able to correctly predict failed firms is considered more beneficial for users than a model that focusing on correctly classifying active firms. Therefore, it is of more concern for users to have a prediction model that can prevent losses that could occur from investing in failed firms, as well as giving assurance to make profit in active firms. It is essential to have an unbiased balanced model to predict business status.

3.5.4. Sensitivity and Specificity

Sensitivity and specificity represent measures that deal with each class of business status in the dataset. Sensitivity is defined as how effectively a classifier is able to identify positive class (Sokolova and Lapalme, 2009). The measure calculates the percentage of correctly classified failed firms, referred to as true positive (TP) predictions (according to the confusion matrix). Specificity is defined as how effectively a classifier identifies firms with a negative class (Sokolova and Lapalme, 2009). It measures the proportion of correctly classified active firm, which is known as true negative (TN) (Bešlić Obradović *et al.*, 2018). Sensitivity and specificity can be calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.20}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.21}$$

3.5.5. Area Under Curve Receiver Operating Characteristic Curve

AUC is a measurement used to evaluate model classification performance based on various threshold measures. Basically, it represents the area under the Receiver Operating Characteristic (ROC) curve, which is a two-dimensional graphical representation of the possible distribution of the predicted business class, as shown in Figure 3.6. The proportion of correctly predicted active firms is plotted in the (x-axis) as the true positive rate (sensitivity), and the percentage of misclassified failed firms is plotted in the (y-axis) as the false positive rate (1-specificity) (Brown and Mues, 2012). Moreover, the ROC illustrates the behaviour of a classifier without being affected by any misclassification errors, or any change in class distribution (Veganzones and Séverin, 2018).

$$\text{AUC} = \frac{\text{Sensitivity} + (1 - \text{Specificity})}{2} \tag{3.22}$$

The diagonal line is the balance of sensitivity and (1-specificity) for a random classification model, with an AUC value of 0.5. The ROC curve should be as far to the top left corner as possible for a successful classification. The best classification classifier is the ROC_1 curve in the example shown in Figure 3.6.

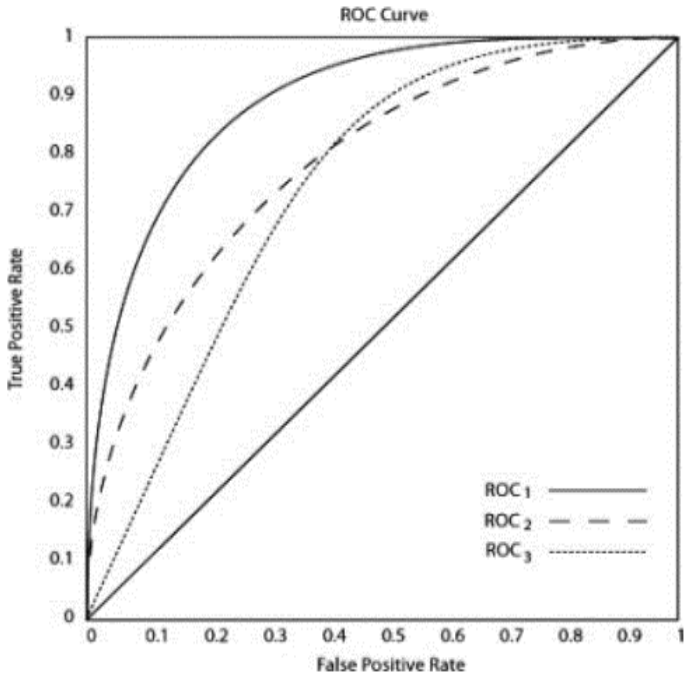

Figure 3.6: ROC illustrative example

Source: Brown and Mues (2012)

In business failure classification problems, AUC provides indications about how well the model performs in distinguishing between different classes (Son *et al.*, 2019). For instance, a model with a higher AUC value performs better in correctly classifying both classes of the dataset, whereas an excellent model has AUC values close to 1, reflecting good separability. In contrast, a model with an AUC value close to 0 has the worst separability, with an AUC value of 0 meaning that the model is reciprocating the result and predicting all 0s as 1s, and all 1s as 0s. However, an AUC value of 0.5 means that the model has no class separation capacity.

3.5.6. Brier Score

The Brier score is a cost or loss function used as a model performance measurement to measure the accuracy of probabilistic predictions (Brier, 1950). Since it is a cost function, a classifier with lower Brier score measure has more accurate prediction performance than a classifier with a higher score. Therefore, the best possible classifier can have a Brier score of 0, and the worst can have a score of 1. Brier score is an important performance measurement in evaluating classifier prediction (Lessmann *et al.*, 2015). Unlike average accuracy rate, which transforms classifier prediction into two separate classes (0 and 1) based on a threshold pre-determined value, Brier score measures the mean squared error of the classifier prediction as follows:

$$\mathrm{BS} = 1/\mathrm{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{p_i} - \mathrm{y_i})^2 \tag{3.23}$$

where N denotes the total number of firms, P_i stands for the probability of the firm i, and y_i is the actual class for the firm.

3.5.7. Reliability Curve

A reliability diagram offers a visual examination to test if the binary classifier is calibrated. It is used as a key diagnostic tool to check model calibration by plotting predictive probability values against the actual observed event. For binary classification, both statistical and machine learning classifiers generate continuous predictive probabilities within the range of 0 to 1. In these settings, to generate a reliability curve, binning and counting approach is used to arbitrary select a certain number of bins, whereby each bin includes a certain number of predicted values, and the average of these values is calculated and assigned to the bin. The same process is assigned to the target values representing the actual status of companies. Each bin value is graphically matched against the optimal diagonal line that represents the actual respective observations.

According to Prati *et al.* (2011), well-calibrated classification outputs should produce a curve close to the optimal diagonal line, resulting in a small reliability measurement, which is the weighted average of the squared vertical distances between the curve and the main diagonal. The curve represents the bias of the predictions, with curve lines lying entirely below or above the optimal diagonal line that represent negative or positive biased predictions (respectively). Regions of low resolution can be presented on the flat lines or line segments on the curve (Prati *et al.*, 2011).

3.6. Statistical Significance Testing

Testing the statistical significance of classifier prediction performance is the final important stage in business failure model development. The fact that a model achieves better prediction performance based on techniques used or performance results obtained is insufficient in itself to prove superiority. Therefore, to have a comprehensible evolution of model performance, hypothesis testing can be used to prove that the differences are statistically significant and are not due to random aspects. In choosing the best test for statistical significance, factors such as the size of the datasets, the number of models or classifiers adopted, and the measurement scale of the output (such as binary, interval, or nominal) should be taken into consideration. Applying inappropriate significance testing can undermine research conclusions, and result in misleading information about model performance (McCrum-Gardner, 2008)

Previous studies tested statistical significance using parametric methods such as paired t-test, and nonparametric methods such as Friedman test. According to Demšar (2006), parametric tests could be conceptually inappropriate, and nonparametric tests are considered statistically safer, since they do not require normal data distribution or homogeneity of variances. Therefore, the dataset used in this study was tested for normality using SPSS version 26 statistical software, which revealed that it is not normally distributed.

Nonparametric testing can be carried out to compare the performance results of more than two classifiers. Hence, Friedman (1940) test was adopted in this study in order to investigate and discover statistical differences in different models' prediction capabilities for business failure.

Friedman Test

In order to assess and compare the classification results obtained by various classifiers used, Friedman test (Demšar, 2006) was adopted in this study. It is a non-parametric randomised back analysis of variance that allows comparison of classifiers' results for the same subject.

The test is a Chi-square with *j-1* degree of freedom, where *j* is the number of repeated measures. The null hypothesis is rejected when the p-value is small (usually < 0.05). The aim of this test is to determine whether statistically significant differences exist between the algorithms examined over certain datasets. The test determines the algorithm ranks of each set of data.

Friedman test detects whether there are statistically significant differences between algorithms examined and classify algorithms from the best to the worst. If statistical significance is detected, the researcher can carry out post-hoc procedures to determine which algorithms differ significantly. In our case, the test is used to detect the significance of each classifier's predictions represented in columns, and different outputs on each sample in rows, as in Table 3.5 and Table 3.6.

Table 3.5: Floating predictions

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

Source: Author

Table 3.6: Predictions ranking

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

According to the tables, the initial step is to convert the row of floating predictions to ranking, where $rij \in \{1,2, ...,N\}$, $i \in \{1...N\}$, $j \in \{1...C\}$, $rij \neq rik \forall i \in \{1...N\}$, and $j,k \in \{1..C\}$. For example, if a row consists of the predictions (1, 0.35, 0.15, 0.56, 0.67), it will be converted to ranking row as (5, 2, 1, 3, 4) (respectively), so that higher prediction value will receive bigger ranking. After converting classifier outputs to rankings for each row, equation (3.24) is applied:

$$S = \frac{12}{NC(C+1)} \sum_{i=1}^{c} R_i^2 - 3N(c+1), where\ R_i = \sum_{j=1}^{N} r_{ij} \quad (3.24)$$

The probability of S can be approximated by chi-squared distribution if $n > 15$ or $c > 4$. The null hypothesis is rejected if the value of S is greater than the critical value of chi-squared distribution $X_p^2\ (c-1)$ for probability p.

If the null hypothesis is rejected, a post-hoc Bonferroni–Dunn pairwise comparisons test is recommended. The test measures the critical difference (CD), which is the minimum required difference in rank sums for a pair of classifiers to differ at the pre-specified alpha level of significance (Demšar, 2006). The CD statistic is calculated using the difference in rank sum averages (i.e., Rj/n), rather than rank sums.

3.7. Study Framework and Research Design

The above discussion explained the essential steps of developing a business failure model. Accordingly, the experimental design of the proposed business failure model in this thesis is shown in Figure 3.7, and summarised as follows:

1. Data collection
2. Data pre-processing
3. Data splitting
4. Feature selection
5. Model development
6. Performance evaluation
7. Statistical significance testing

3.8. Summary

This chapter provided an overview of all the methodological steps used to build and develop a business failure model. The development process of business failure encompasses several main stages, starting from the dataset collection and pre-processing to be implemented in the models. Hence, the experimental design framework adopted in this thesis demonstrates the essential steps used to develop comprehensive and reliable models of business failure in the UK. The next chapters fully demonstrate the execution of each stage of the proposed model.

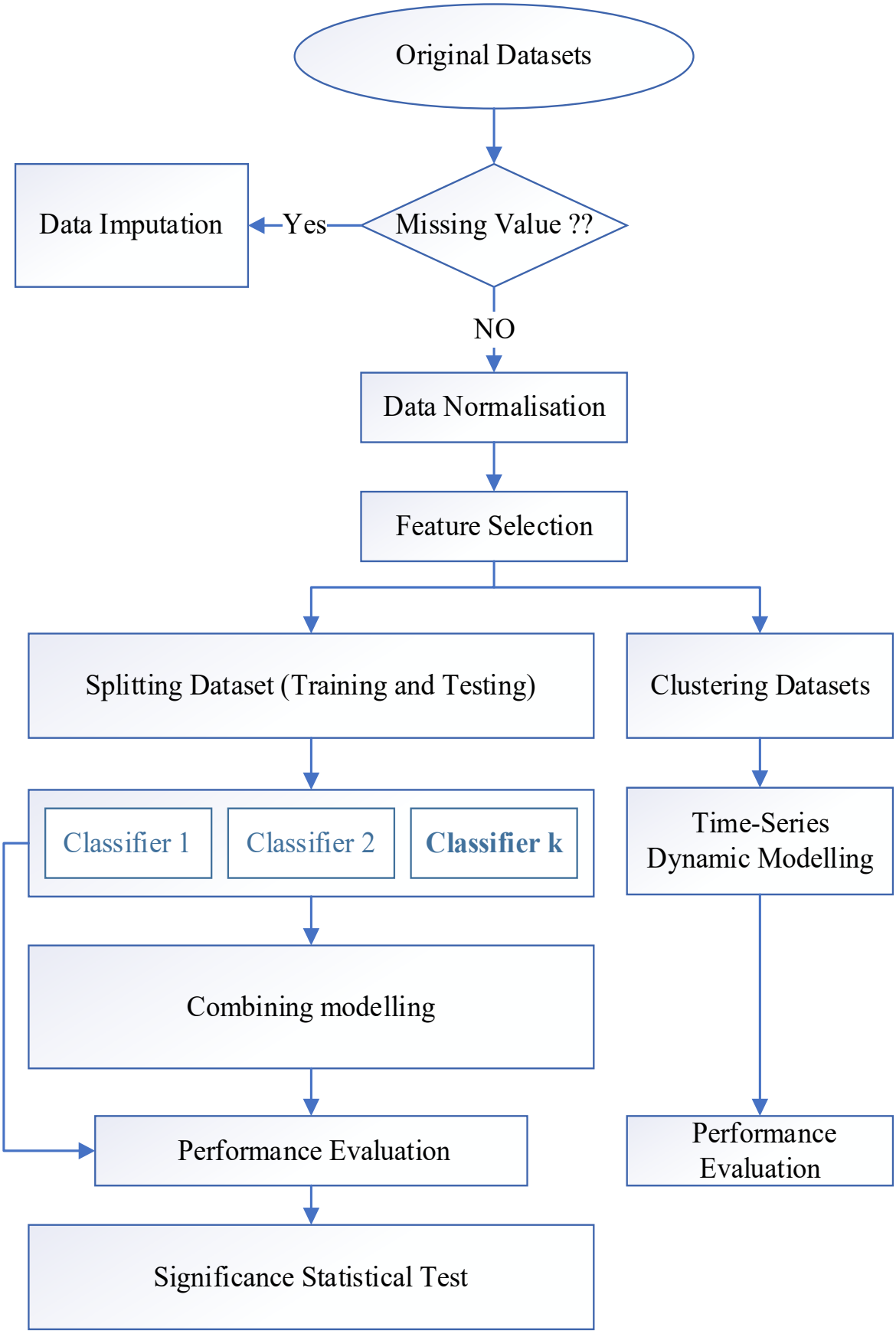

Figure 3.7: Flowchart of the experimental design

Chapter 4
Data Mining Tools for Insolvency Modelling

4.1. Introduction

This chapter presents multiple classification methods used to develop a business failure classification model and evaluates their performance over three datasets (comprising financial data related to UK companies). It illustrates the development phase of each classifier and compares its prediction capability based on its performance measurements, in order to assess how each classifier performed in predicting business failure. Table 2.2 displays several modelling techniques used as individual classifiers for business failure predictions, which are to be deployed in this chapter and different fields.

These famous individual classifiers have been used to solve binary classification problems in deferent fields, because of their ease of implementation. Nine base classifiers are used for model development and analysis in this thesis, namely ANN, SVM, DT, LR, LDA, KNN, NB, DPL, and ENS-DT. Each of these classifiers has its own running methodology, with particular advantages. However, LR is considered as the default modelling method in the field of business failure, and it was selected as a benchmark classifier to compare other performance results with (as explained previously). 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.

For individual classifiers, it is crucial to implement the best classifier parameters in order to allow the model to perform well. For business failure classification datasets, the same features are used for model development, so implementing the best model parameters plays a major role to achieve best classification performance for all datasets. In this section two major aspects are presented: (1) dataset preparation and pre-processing in order to be fed into the classifiers; and (2) the model development process and training for each classifier based on the best parameter selection (to achieve the best classification results).

4.2. Data Pre-Processing and Preparation for Training and Evaluation

As explained in Chapter 3, pre-processing and preparation before feeding financial datasets into the classifiers enhances data quality, improving classification capability to produce best results. To achieve this objective, the financial data is pre-processed by the following steps:

- Data cleaning by removing firm data with missing values.
- Normalisation of the values of some attributes of each dataset.

The datasets are then partitioned using two substantial methods for training and testing purposes depending on the classifiers used:

- Data are divided into two data segments (training, and testing) using the percentages 80% and 20%.
- Data are divided using K-fold cross-validation partitioning technique with 10 × 5 cross-validation.

The training stage builds and develops classifier models by using and selecting classifier parameters, which are then tested in order to evaluate prediction performance. The classifier building and development process is explained in the following section.

4.3. Model Development and Experimental Results

This section presents the performance of each individual classifier in tables and figures extracted from the results of testing using holdout technique and 10 × 5 cross-validation across the three datasets of UK companies. Each table includes eight performance measurements selected to evaluate and compare classifier performance, which allows us to discover the best classifier among all individual classifiers for classifying business failure. Moreover, all results are compared to the LR industrial statistical modelling technique, which is the traditional benchmark classifier in business failure literature. Tables 4.1 to 4.9 demonstrate each individual classifier's results based on all classification measurements. Moreover, the figures in this chapter show the graphical presentation of the ROC and reliability curve performance parameters of each classifier.

After describing all results, a thorough discussion and analysis process is conducted based on performance measurement results to explain each classifier's advantages and disadvantages, in order to discover the best classifier model for BF classification. The model development of each machine learning single classifier is based on parameters selected and tuned in order to achieve the highest performance.

4.3.1. Linear Regression

LR, the benchmark classifier, illustrates the outperformance of machine learning over statistical classification methods. According to Table 4.1, the average accuracy rate increased from

75.4% for year 2017 to 80.1% for year 2019. This increase was reflected in the substantial decrease of Type II Error, which means the misclassification of failed firms decreased. As can be seen, Type II Error was 29.3% in 2017, and decreased to 22% in 2019.

However, the specificity rate, which represents the ability of the model to correctly classify failed firms, is relatively higher than the sensitivity one, indicating the model's performance in classifying active firms for all years' datasets, whereas the scenario is the opposite for all other classifiers, which had higher sensitivity than specificity. Meanwhile, LR classifier only outperformed the other statistical LDA classifier based on all performance measurements; all other classifiers showed better results than LR.

All-Data results show that the classification performance of the LR classifier was enhanced in comparison with each year datasets results; the average accuracy rate of 81.8% is 1.7% higher than 2019 dataset and 6.4% higher than 2017 results. The improvement of classification is related to correctly classifying failed firms, manifest in a reduction of Type II Error to 18.8%.

Table 4.1: LR results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	80.1%	77%	75.4%	81.8%
Type II Error	22%	26.3%	29.3%	18.8%
Type I Error	17.8%	19.7%	20%	17.6%
Sensitivity	78%	73.7%	70.7%	81.2%
Specificity	82.2%	80.3%	80%	82.4%
AUC	89%	86%	85%	91%
Brier Score	0.1328	0.1531	0.1621	0.1211
Area Under Reliability Curve	0.0574	0.0685	0.0710	0.0385

The ROC graph presented in Figure 4.1 shows the curve for each year's dataset generated from the LR classifier results. The shape of the curve and the AUC shown in Table 4.1 give indications about the model performance. Based on the graph, the 2019 dataset had a better curve than 2018 and 2017.

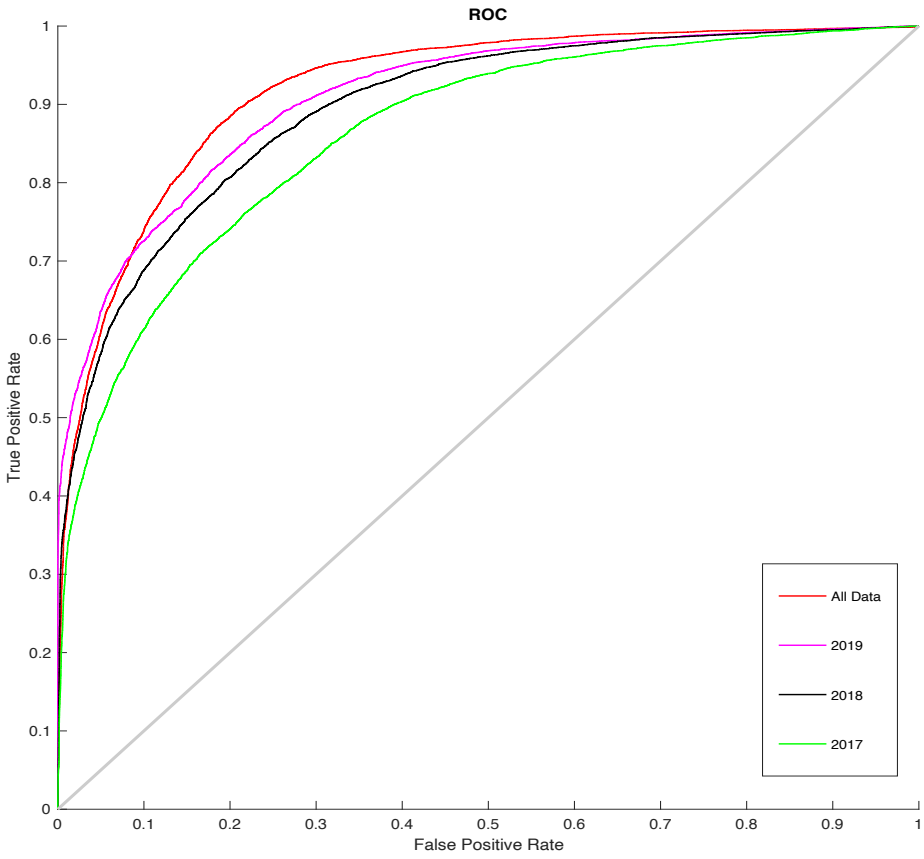

Figure 4.1: ROC curve for LR classifier

Figure 4.2 shows the reliability diagram of the LR classifier. The shape of the line of the predictions presented in 20 bins indicates very low classification error for active firms (the part of the curve above the 0.5 threshold of mean predictive value); the line should optimally be at the diagonal line or above. The line also indicates higher error rate of failed firm classification (the part of the curve below the 0.5 threshold of mean predictive value); the line should optimally be at the diagonal line or below. This mean LR is more reliable in correctly classifying active firms than failed ones, which is also reflected in the area under the reliability line shown in Table 4.1: 3.85% for All-Data, and 6.85%, 7.1%, and 5.7% for the years 2017, 2018, and 2019, respectively.

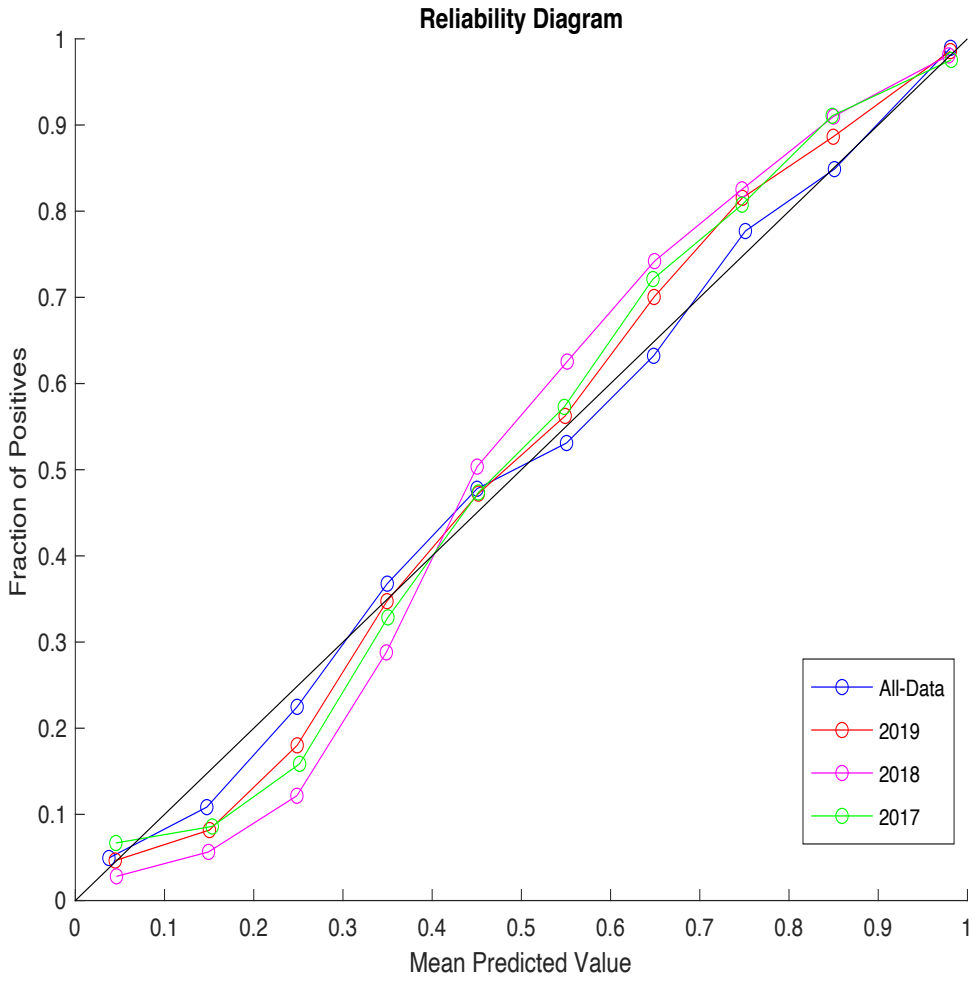

Figure 4.2: Reliability diagram for LR classifier

4.3.2. Linear Discriminant Analysis

LDA is a classification method that uses Gaussian distribution function to classify input data. The model was trained using a dataset with k-fold cross-validation with five folds. Table 4.2 shows the results of the LDA classifier, which was found to be the worst based on all performance measurements. This indicates that LDA modelling is not preferable for business failure classification using financial ratios datasets. The best accuracy rate of the model was for the 2019 dataset, with only 71.5% (the lowest rate among all other classifiers). The model has the highest Type I and Type II Error, and its Brier score has an inverse relationship with accuracy rate, reflected in the high values shown in Table 4.2. However, the classifier showed better results for the All-Data dataset, with 75% average accuracy rate. This confirms that Big Data is relatively germane for this classifier's classification performance.

Table 4.2: LDA results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	71.5%	70.5%	69.5%	75%
Type II Error	28.7%	30.4%	30.8%	24.6%
Type I Error	28.3%	28.5%	30.1%	25.4%
Sensitivity	71.3%	69.6%	69.2%	75.4%
Specificity	71.7%	71.5%	69.9%	74.6%
AUC	77%	76%	75%	80%
Brier Score	0.1978	0.215	0.2059	0.1825
Area Under Reliability Curve	0.0825	0.0799	0.0772	0.1028

According to Figure 4.3, the ROC curve for LDA model shows bad performance in comparison with all classifiers. This can be seen as all curves lay closer to the diagonal line, indicating lower AUC values. An AUC value of 80% for All-Data is considered the best performance of the model, but it is still the lowest value among all other classifiers.

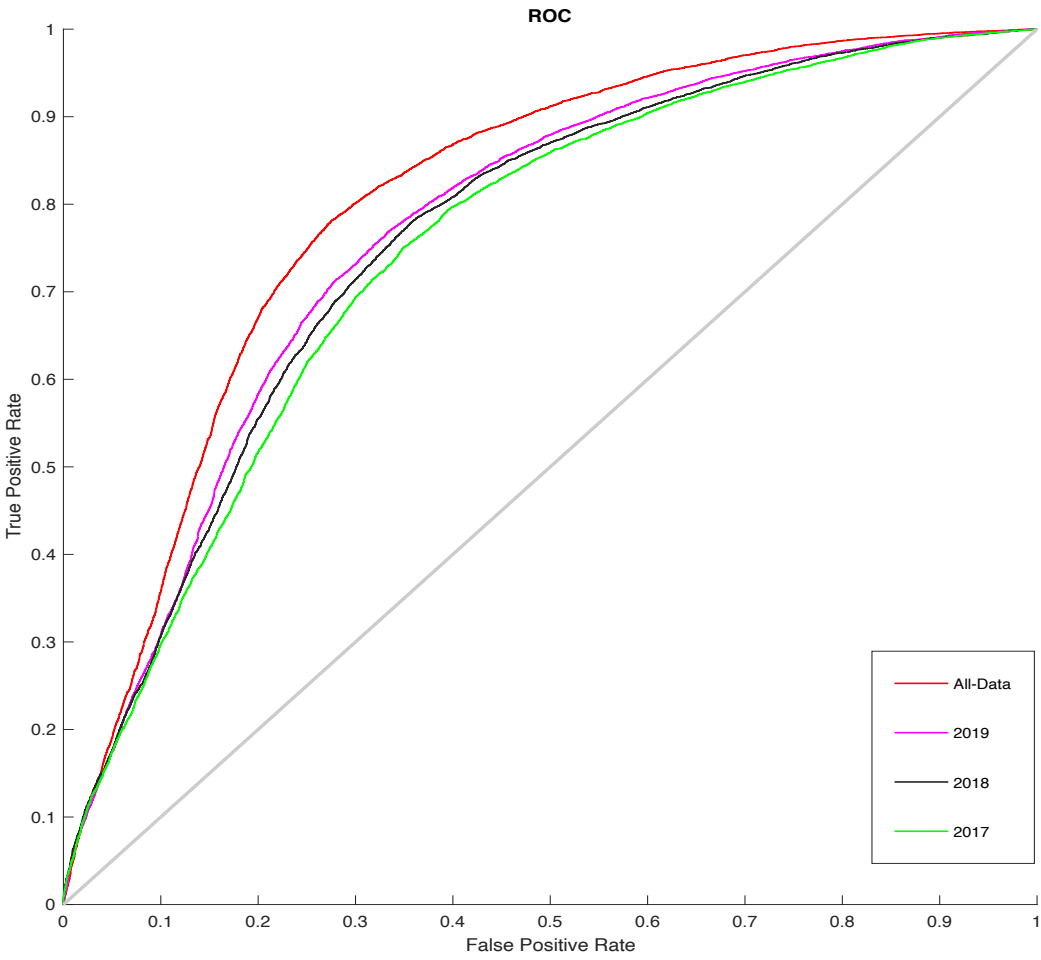

Figure 4.3: ROC curve for LDA classifier

According to the reliability diagram in Figure 4.4, the shape of the prediction line indicates a very bad performance. As can be seen from the figure, the right side of the line is below the diagonal line, whereas it should be above, which means the model is not reliable in correctly predicting firm status. The high area value between the prediction line and the optimal diagonal line shows high prediction error. These values are the highest among all other classifiers for all years' datasets. It can be noticed that the lines for all datasets lay a similar distance from the diagonal line, approved by their area values from Table 4.3.

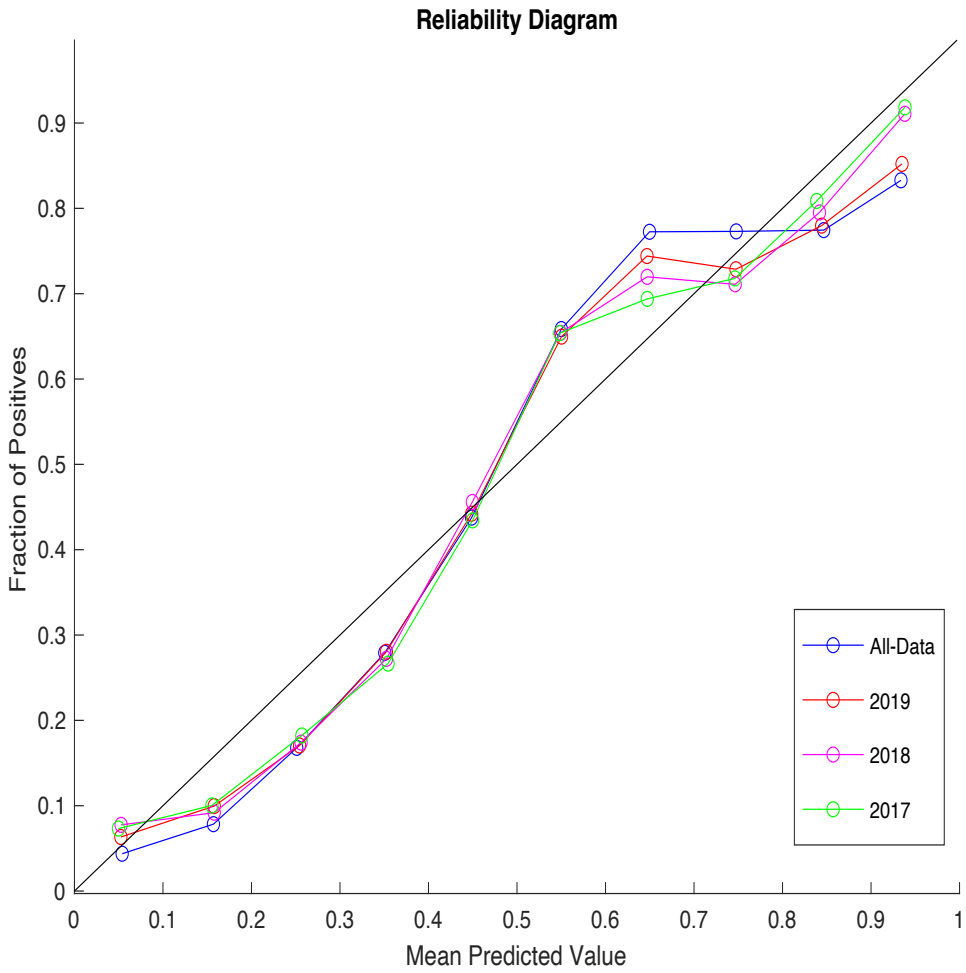

Figure 4.4: Reliability diagram for LDA classifier

4.3.3. K-Nearest Neighbour

KNN is a classification algorithm used for the approximation of local function and the execution of all computations until function evaluation. For classification purposes, the algorithm relies on computing the distance between features and targets, by normalizing the features to achieve the highest classifying accuracy rate. Therefore, to solve classification problems, the algorithm assigns more weights to nearer neighbours, which contributes to the average more than more distant neighbours. It is crucial to define the number of neighbours, which is assigned as 10 for the numerical financial dataset. The algorithm method is the nearest neighbour search method, which has two selections (either 'Kdtree' or 'Exhaustive'). The Exhaustive algorithm outperformed Kdtree for binary classification, as it is more efficient and flexible with respect to distance matrix choice. Therefore, Exhaustive method was selected to find the nearest neighbour by computing the distance values from all features to the target using Euclidean distance as a distance metric with the following formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4.1)$$

Where x_i is the value of the variables of the input data, y_i is the target class of the firms in the dataset, and n represent the number of firms in the dataset. The distance weight function is set as 'squared inverse' (weight is equal 1/distance^2). The tie-breaking parameter set as 'smallest', which is used when multiple classes have the smallest cost.

According to Table 4.3, KNN classifier shows relatively weak performance in terms of classification accuracy in comparison with the other classifiers except for LR and LD classifiers. Its average accuracy rates of 81.2%, 80.4%, and 79.4% for the years 2019, 2018, and 2017 (respectively) are considered to indicate relatively low classification performance. However, the decline in average accuracy through year 2017 to 2019 is relatively low, which indicates relatively stable performance among all datasets. Although the model's high Type I Error of 25.5% for the year 2017 dataset decreased slightly to 22.9% for the year 2019 dataset, the error values still reflect the weakness of the model to correctly predict failed firms in relation to the other datasets. It is worth mentioning that when using All-Data dataset, the classifier showed a small enhancement of classification accuracy, with a 1.8% increase in comparison to 2019.

Table 4.3: KNN results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	81.2%	80.4%	79.4%	82.9%
Type II Error	14.7%	15.6%	15.7%	12.7%
Type I Error	22.9%	23.5%	25.5%	21.6%
Sensitivity	85.3%	84.4%	84.3%	87.3%
Specificity	77.1%	76.5%	74.5%	78.4%
AUC	90%	89%	88%	91%
Brier Score	0.1311	0.1362	0.1428	0.1194
Area Under Reliability Curve	0.0534	0.047	0.0446	0.0592

The ROC curve of KNN classifier shown in Figure 4.5 displays a smooth decrease from the year 2017 to 2019. This is considered compatible with the classifier performance according to

other performance parameters shown in Table 4.3. An AUC of 91% for All-Data is considered low in comparison with other classifiers, but it is still higher than LR and LD; KNN outperformed these classifiers in terms of all performance parameters.

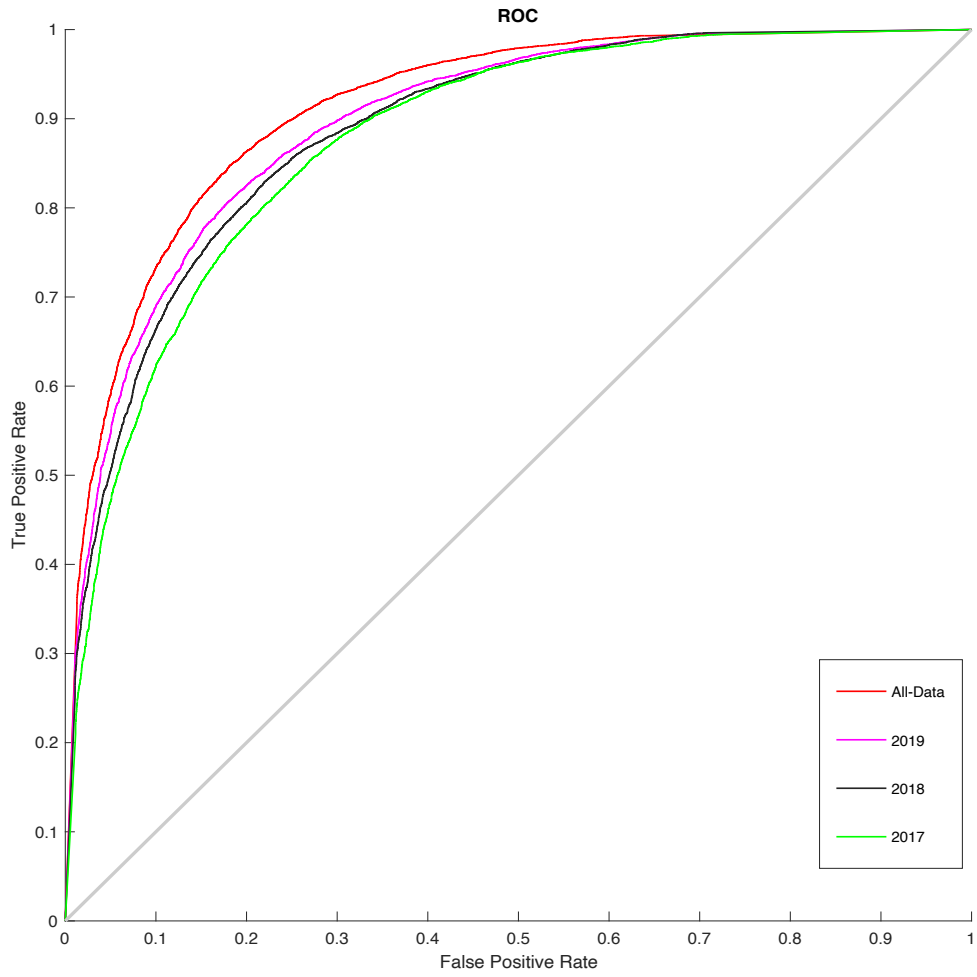

Figure 4.5: ROC curve for KNN classifier

Figure 4.6 shows the reliability diagram for KNN classifier, plotting the predictions line below the diagonal. This indicates low reliability of the classifier prediction performance, even though the lines for all datasets were close to the optimal diagonal line. This can be explained by its weak prediction of failed firms.

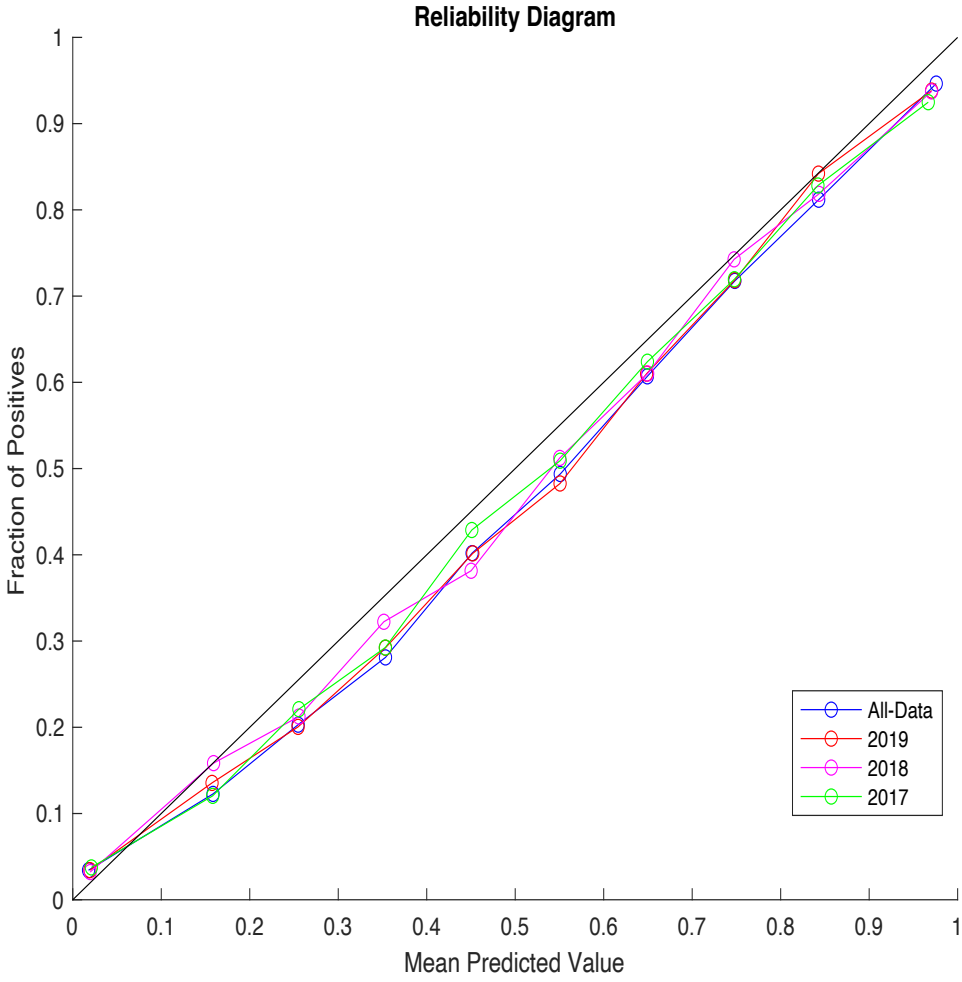

Figure 4.6: Reliability diagram for KNN classifier

4.3.4. Artificial Neural Networks

In order to build a classification network using ANN, it is crucial to find a suitable network topology structure that includes the best hidden neurons and hidden layers of the network. Moreover, an important step of building the network is to assign the optimal learning function, training function, transfer function, and the learning speed of the network. To build the network, there are three training function available to train the datasets *traingda, traingdx, and trainlm*. traingda can train any network as long as its weight, net input, and transfer functions have derivative functions. Backpropagation is used to calculate derivatives of performance *dperf* with respect to the weight and bias variables X. Each variable is adjusted according to gradient descent using equation 4.2:

$$\mathrm{dX} = lr * dpref/dx \tag{4.2}$$

where lr represent the learning rate. At each epoch, if performance decreases toward the goal, then the learning rate is increased by the factor lr_inc. If performance increases by more than

the factor max_perf_inc, the learning rate is adjusted by the factor lr_dec and the change that increased the performance is not made.

The function traingdx combines adaptive learning rate with momentum training. It is invoked in the same way as traingda, except that it has the momentum coefficient mc as an additional training parameter. traingdx can train any network as long as its weight, net input, and transfer functions have derivative functions. Backpropagation is used to calculate derivatives of performance *dperf* with respect to the weight and bias variables X. Each variable is adjusted according to gradient descent with momentum using the equation 4.3,

$$dX = mc * dXprev + lr * mc * dpref/dX \quad (4.3)$$

where *dXprev* is the previous change to the weight or bias.

For each epoch, if performance decreases toward the goal, then the learning rate is increased by the factor lr_inc. If performance increases by more than the factor max_perf_inc, the learning rate is adjusted by the factor lr_dec and the change that increased the performance is not made.

Trainlm supports training with validation and test vectors if the network's NET.divideFcn property is set to a data division function. Validation vectors are used to stop training early if the network performance on the validation vectors fails to improve or remains the same for max_fail epochs in a row. Test vectors are used as a further check that the network is generalizing well, but do not have any effect on training. trainlm can train any network as long as its weight, net input, and transfer functions have derivative functions.

Backpropagation is used to calculate the Jacobian jX of performance perf with respect to the weight and bias variables X. Each variable is adjusted according to Levenberg-Marquardt through the equation 4.4, 4.5, and 4.6,

$$jj = jX * jX \quad (4.4)$$

$$je = jX * E \quad (4.5)$$

$$dX = -(jj + I * mu) \backslash je \quad (4.6)$$

where E is all errors and I is the identity matrix.

The adaptive value mu is increased by mu_inc until the change above results in a reduced performance value. The change is then made to the network and mu is decreased by mu_dec.

To select the optimal training function, *trainlm* was selected to allow the network to adjust input weights to achieve the optimal output value and minimise the error of the classification where the *trainlm* is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. This function assumes that model's performance is a mean of squared errors. Therefore, networks trained with this function must use the Mean Squared Error performance function in order to minimise the classification error of the model. This function is often the fastest backpropagation algorithm in the toolbox and is highly recommended as a first-choice supervised algorithm, although it does require more memory than other algorithms (*traingdx, traingda*) where *traingdx* is a network training function that updates weight and bias values according to gradient descent momentum and an adaptive learning rate and *traingda* is a network training function that updates weight and bias values according to gradient descent with adaptive learning rate.

The network structure consists of input, hidden, and output layers. One hidden layer is selected, where it is crucial to select the optimal number of neurons in order to improve network classification capability (the greater the number of neurons, the more complex the model). For data training, the number of ten neurons was selected based on trial-and-error, which produced the best performance based on all measurement parameters.

For the hidden layer the *tansig* hyperbolic tangent is selected as a transfer function, which is the most common function used for network development:

$$tanh(x) = \frac{2}{1-e^{-2x}} + 1 \quad (4.1) \qquad (4.7)$$

For the output layer, pure linear 'purlin' function was used as a transfer function. This uses hidden layer output as the final decision of the network classifier.

Table 4.4 shows the performance of NN classifier. It has better performance than KNN, LD, and LR in terms of average accuracy. The model achieved 84.9% average accuracy for All-Data dataset. Its average accuracy of 83.1% for 2019 is just 1.6% higher than for the 2017 dataset, which indicates relatively stable classification rates among three datasets. However, these rates are still lower than those other classifiers apart from those mentioned above.

On the other hand, the classifier has relatively balanced Type I Error and Type II Error, which illustrates the model's capability to perform well in predicting both active and failed firms. However, Type I Error, which reflects model failure to predict active firms, only outperformed

LD and LR classifiers for all years' datasets. The model's Brier score results were lower than those of KNN and LD, indicating lower predictions error.

Table 4.4: NN results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	83.1%	82.5%	81.5%	84.9%
Type II Error	16.4%	17.5%	18.8%	13.7%
Type I Error	17.4%	17.4%	18.2%	16.5%
Sensitivity	83.6%	82.5%	81.2%	86.3%
Specificity	82.6%	82.6%	81.8%	83.5%
AUC	92.02%	91.2%	90%	93%
Brier Score	0.1155	0.1212	0.1454	0.1073
Area Under Reliability Curve	0.0867	0.0881	0.113	0.0958

Figure 4.7 shows the ROC curve for NN classifier. It indicates an increase in the performance of the classifier from the year 2017 to 2019, as the curve shifted away from the diagonal line. However, the shape of the curves is still better than KNN, LR, and LD classifiers; moreover, the AUC values for NN for all years' datasets are greater than those classifiers.

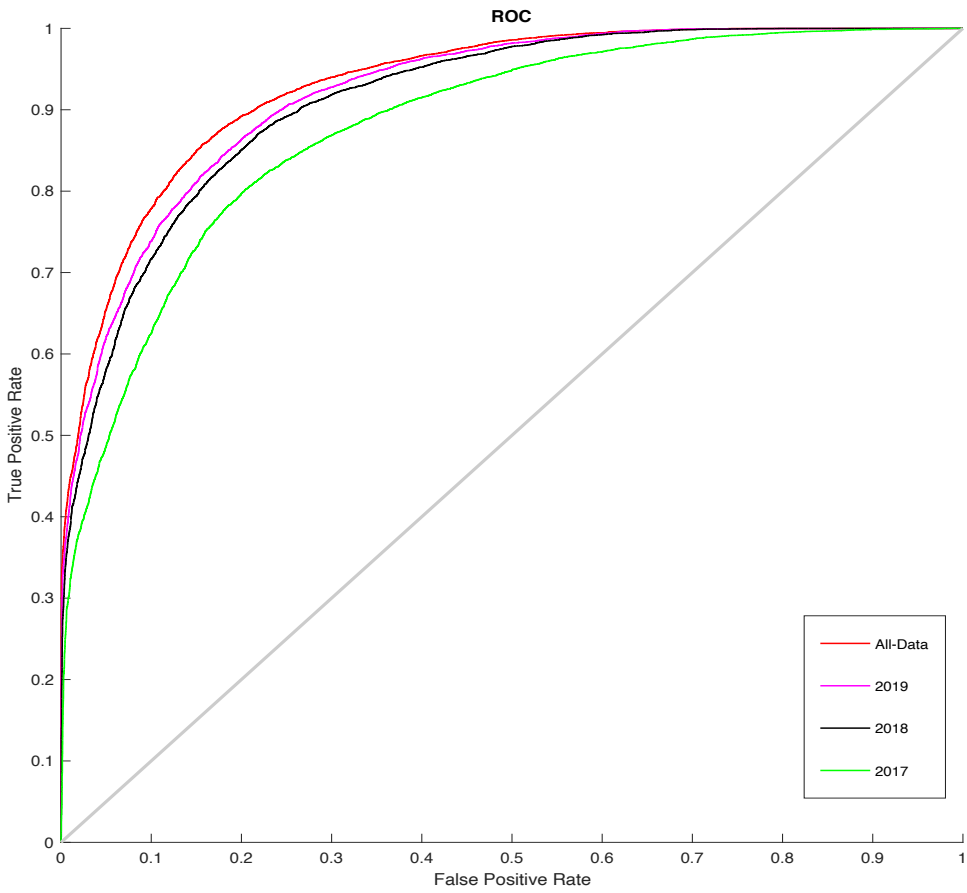

Figure 4.7: ROC curve for ANN classifier

Figure 4.8 shows the reliability diagram resulting from the predictions of the NN model. The classifier achieved the best line shapes for all datasets among all classifiers. This is reflected in the lowest area between the lines and the diagonal line, based on the values cited in Table 4.4. This is due to the balanced classifier's capability to predict both active and failed firms for all years, which gives the classifier a prediction advantage over other classifiers.

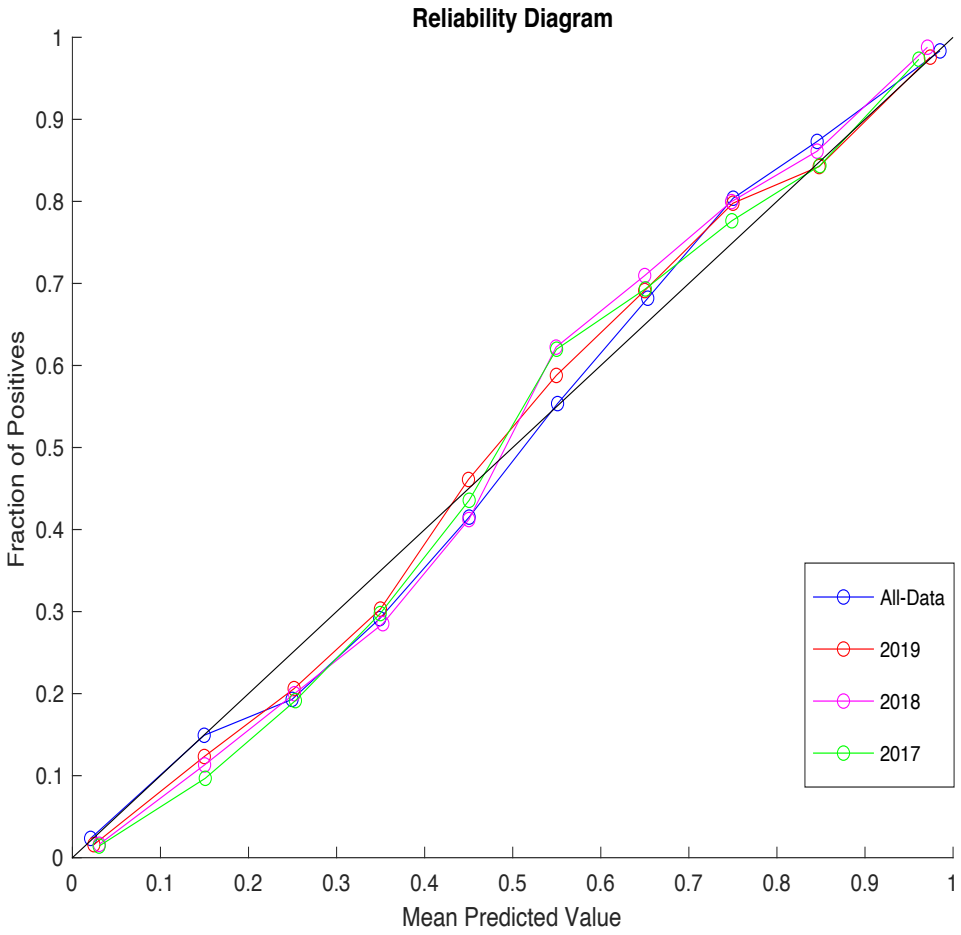

Figure 4.8: Reliability diagram for ANN classifier

4.3.5. Support Vector Machine

SVM has a decision surface called the optimal hyperplane, and the data points closest to it and the margins (dashed lines) are called support vectors. The support vectors are very important elements used for training the model, whereby the SVM finds the optimal hyperplane separating the data with maximum margin width to fit the data. SVM is a non-probabilistic binary linear classifier that classifies a set of training to one of two categories. There are four kernel functions mostly used by SVM: Linear k, Quadratic, Polynomial, and Gaussian Radial Basis function. After training the model using each of this functions, Quadratic SVM outperforms all other functions with a kernel function set as 'Quadratic', and a kernel scale set as 'Automatic'. The value of the kernel scale parameter for the trained model is set as '1.5945'.

Table 4.5 demonstrates SVM performance results for all years' datasets. The model achieved a relatively higher average accuracy rate of 87.7% for the All-Data dataset in comparison with the previous classifiers. Moreover, the accuracy rate for each single year dataset is slightly higher than LR, NN, and LD classifier results. Based on the sensitivity and Type II error results of the model over the years, it has a more stable prediction accuracy of active firm than NN,

KNN, LD, and LR, with a rate gap of only 1.4%. Also, the SVM outperformed all other classifiers except for DT, ENS-DT, and DPL in terms of sensitivity and Type II Error. This is an indicator of its superior ability to correctly classify failed firms.

Table 4.5: SVM results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	86.7%	86.8%	85.3%	87.7%
Type II Error	11.6%	11%	13%	10.9%
Type I Error	14.9%	15.4%	16.3%	13.7%
Sensitivity	88.4%	89%	87%	89.1%
Specificity	85.1%	84.6%	83.7%	86.3%
AUC	94%	94%	93%	95%
Brier Score	0.1063	0.0995	0.1845	0.0843
Area Under Reliability Curve	0.1694	0.1461	0.2825	0.0796

Figure 4.9 demonstrates the ROC for the SVM classifier, showing slight gaps between the curves for all years' datasets. The All-Data dataset clearly has the best curve, which means that the classifier's performance is enhanced with larger datasets. However, all the curves have a good shape, reflecting good classifier performance.

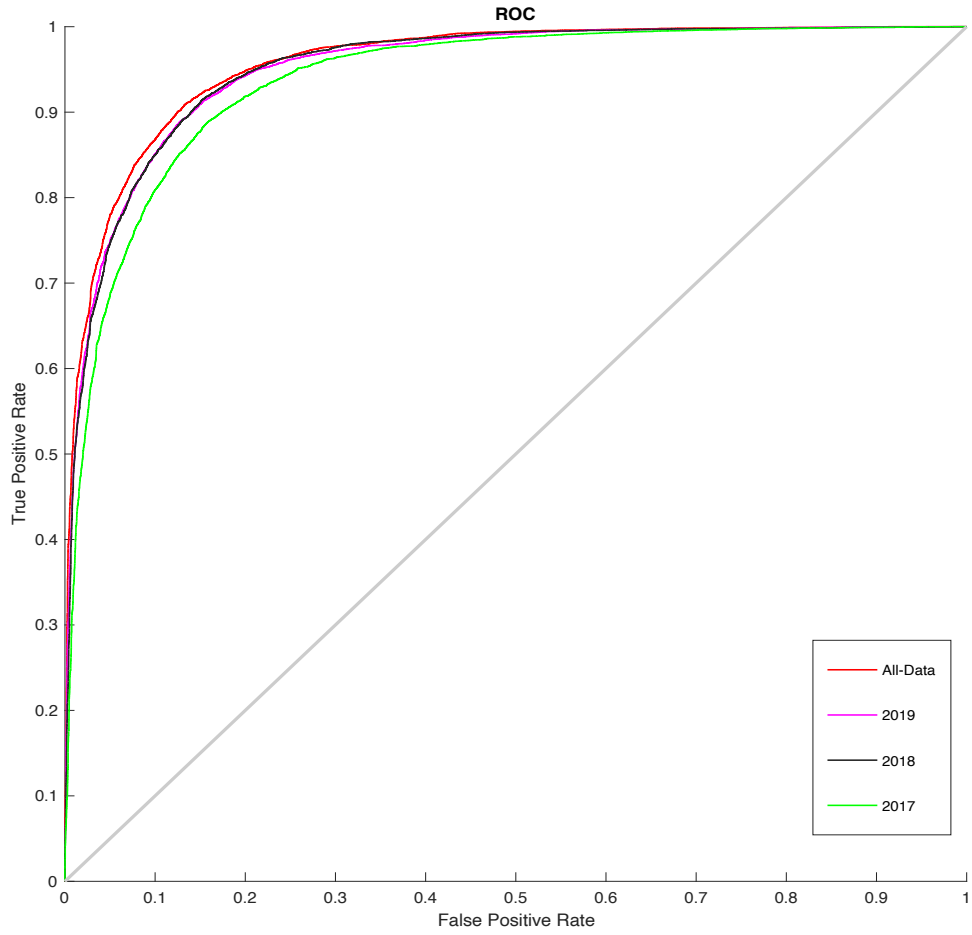

Figure 4.9: ROC curve for SVM classifier

Source: Author

According to Figure 4.10 the reliability diagram of the SVM classifier shows a relatively better shape for the All-Data and 2019 datasets. The line shifted away from the diagonal line with the worst distance for the 2017 dataset. The area between the predictions lines and the diagonal is the highest for year 2017, with 0.2825 (Table 4.5), and it is considered the worst among all other classifiers.

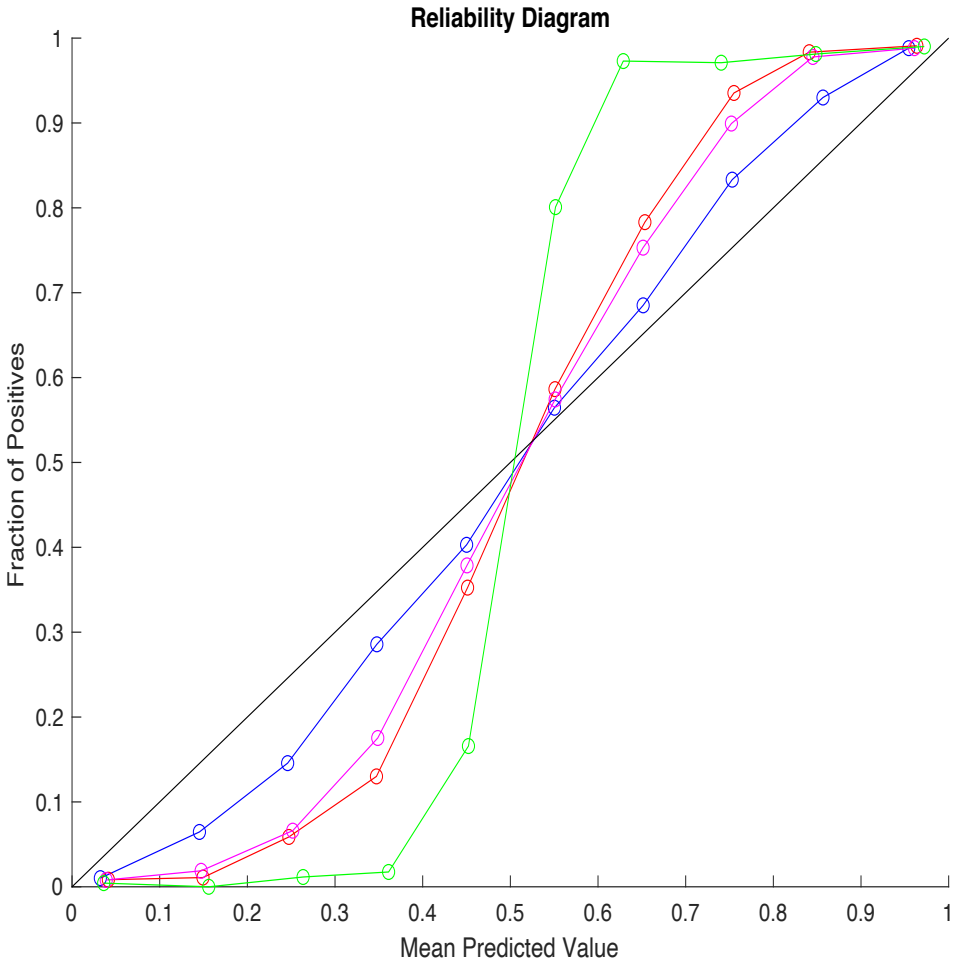

Figure 4.10: Reliability diagram for SVM classifier

4.3.6. Naïve Bayes

NB is a highly scalable classification technique that uses linear time, rather than expensive iterative approximation algorithms used by other classifiers. The algorithm parameter depends on the distribution name for numeric predictors, number of kernels, kernel type and support. The predictor's distribution is set as 'kernel smoothing density estimates', and the number of kernels is 24 in regard to the number of the features of the dataset. The type of kernel is set as 'normal'. The Support parameter is set as 'unbounded', which allows support density to be applied to all real values of the attributes. The activation function of the kernel is selected as 'Gaussian'. Thereafter, the training data is fed to the model to classify firms using all features based on the model parameters selected to obtain the best results.

Based on Table 4.6, NB classifier outperformed the abovementioned classifiers in terms of average accuracy, achieving 89% for All-Data and 86.3%, 87%, and 86.8% for 2019, 2018, and 2017, respectively. It can be seen that the classifier has approximately equal accuracy for all years' datasets; surprisingly, 2018 has the highest classification accuracy, in contrast to the

previous classifiers which achieved their best classification for the 2019 dataset. Type I and II Error were the most stable over all years' datasets in comparison to other classifiers, except for DT, ENS-DT, and DPL. However, the high value of Type II Error indicates weak performance in correctly classifying failed firms. Therefore, it is obvious that the higher average accuracy rate can be explained by the capability of the classifier to predict active firms more correctly than failed ones.

Table 4.6: NB results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	86.3%	87%	86.8%	89%
Type II Error	8.5%	7.8%	7.1%	6.6%
Type I Error	17.95	18.2%	19.3%	15.3%
Sensitivity	91.5%	92.2%	92.9%	93.4%
Specificity	82.1%	81.8%	80.7%	84.7%
AUC	95%	95%	95%	97%
Brier Score	0.1098	0.1090	0.1090	0.0926
Area Under Reliability Curve	0.2908	0.2540	0.2538	0.3374

Figure 4.11 shows the ROC curve for the NB classifier. Based on the shape of the curves for all years' datasets, NB classifier has better performance than previous classifiers, which can be attributed to its higher AUC values (Table 4.6).

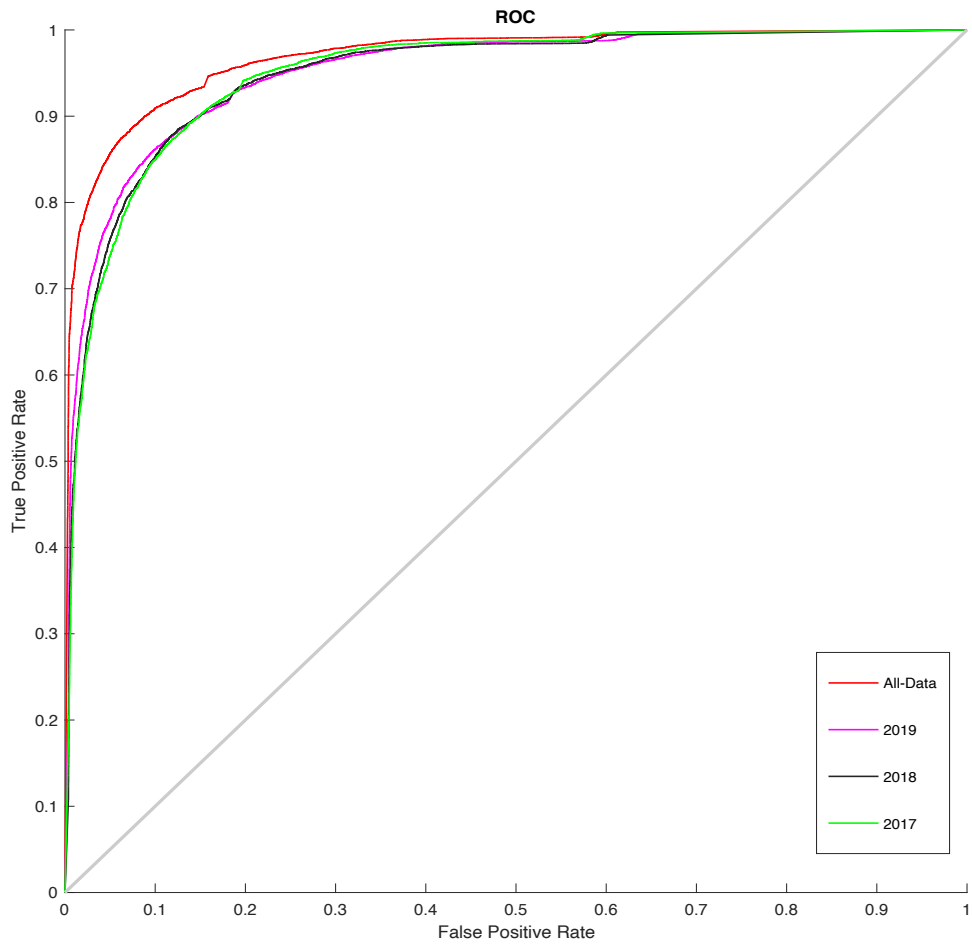

Figure 4.11: ROC curve for NB classifier

Figure 4.12 shows the reliability diagram for the NB classifier. It has the worst shape compared to all other classifiers, as illustrated by the high area between the predictions line and the diagonal line. As shown in Table 4.6, this is based on the value of 0.2908 for the year 2019 dataset, and 0.3374 for the All-Data dataset. Although these values are surprisingly lower in 2018 and 2017, NB still has the worst value among all other classifiers, and the classifier has poor performance to correctly predict failed firms.

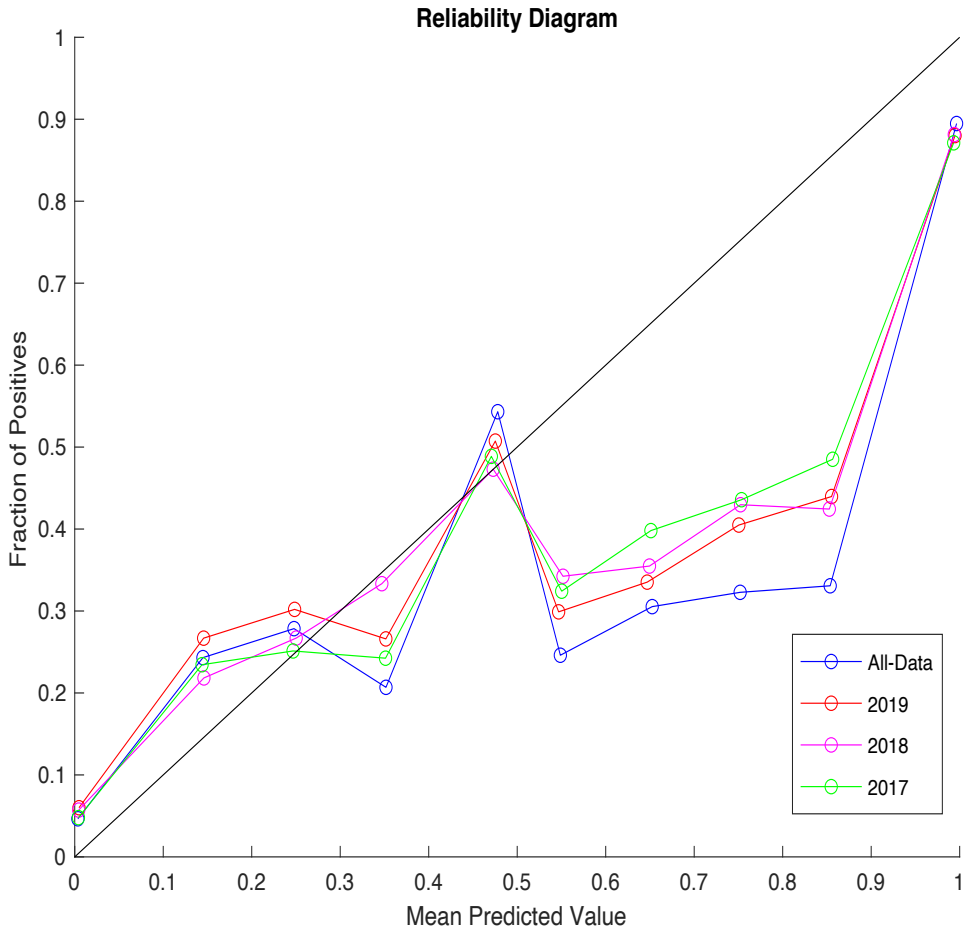

Figure 4.12: Reliability diagram for NB classifier

4.3.7. Decision Tree

DT is a common approach used by researchers for predictive modelling in statistics, data mining, and machine learning, because of its predictive power capability to solve classification problems. Its modelling starts from observation about the target, which is financial ratios in this study, represented in the nodes linked by corresponding branches to terminal (leaf) nodes, to make a conclusion about the target classification represented on the leaf nodes that take the values of '1' and '0'.

As financial ratios were assigned as input data in different classes, with each class labelled to a single node. Each node was labelled to a possible target value. The classification tree is classified when all nodes are labelled with a probability distribution over the target classes. The training process of the model is executed to classify firms' status using the optimal algorithm. The algorithm type is set as 'fine tree', with a maximum number of splits of 20, and the split criterion Gini's Diversity Index is applied. After setting the optimal parameters, training data is fed to the model with k-fold cross-validation, to train the model to classify firm's status.

Based on Table 4.7, DT classifier outperformed all previous classifiers in terms of all measurement parameters. The average accuracy rate increased from 93% for the year 2017 to 94.2% for 2019. This is considered a relatively small difference in comparison to the previous classifiers. The classifier showed more stability in classifying firms over the years. Moreover, the model achieved low values of Type I and Type II Error, which explain the superiority of the model to correctly classify both active and failed companies. However, these values are still higher than those of ENS-DT and DPL classifier. In regard to the specificity values, DT classifier showed more capability to classify failed companies, achieving the highest accuracy rates among the classifiers whose results have been presented thus far.

Table 4.7: DT results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	94.2%	93.6%	93%	95.3%
Type II Error	5.5%	5.1%	5.2%	4.9%
Type I Error	6.5%	7.6%	8.7%	4.5%
Sensitivity	94.5%	94.9%	94.8%	95.1%
Specificity	93.5%	92.4%	91.3%	95.5%
AUC	98%	98%	97%	99%
Brier Score	0.0460	0.04880	0.0576	0.0373
Area Under Reliability Curve	0.085	0.0635	0.1203	0.1035

Figure 4.13 demonstrates the ROC curve for DT. The curve has the optimal shape in comparison with previous classifiers, which reflects the superior performance of the model. The curve shifted upward slightly as the accuracy of classification of companies' status increased from 2017 to 2019. This can be explained by the AUC values for DT shown in Table 4.7.

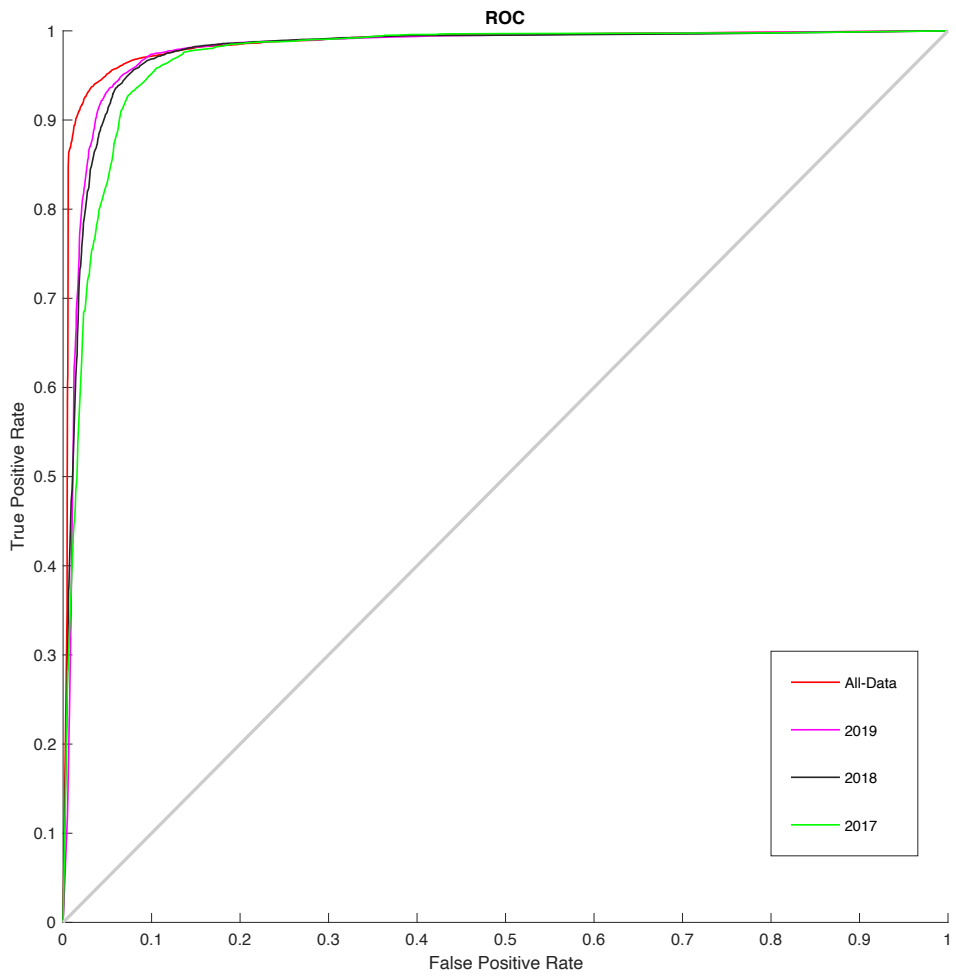

Figure 4.13: ROC curve for DT classifier

Figure 4.14 shows the reliability diagram for DT classifier. The diagram illustrates the good performance of the model for classifying companies' status. Surprisingly, the All-Data dataset has a greater area between the predictions and diagonal lines than the years 2019, 2018, and 2017.

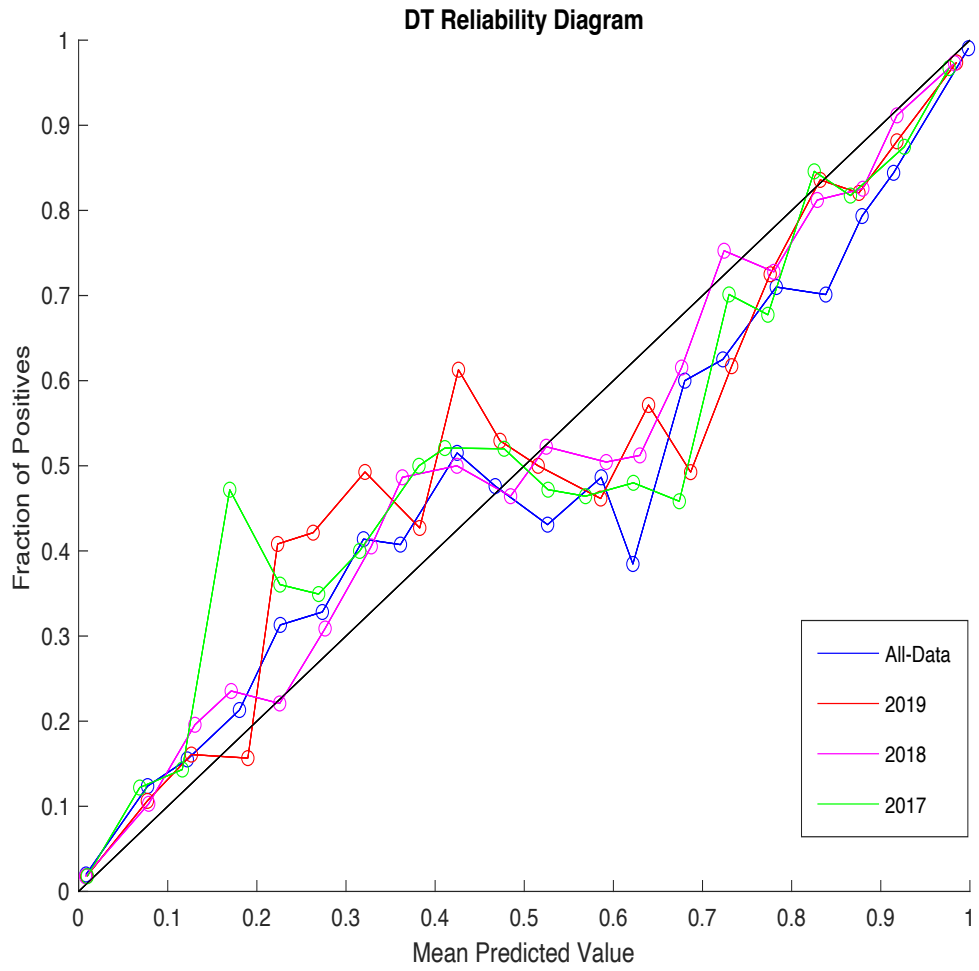

Figure 4.14: Reliability diagram for DT classifier

4.3.8. Ensemble Boost-Decision Tree

Ens-DT is a classification tool used to improve the performance of other types of classification algorithm. It is known by its capability of building a strong classifier by combining the outputs of weaker classifiers in weighted sums to represent the final output (i.e., the boosted classifier). Therefore, for best classification, the type of ensemble boost is selected as 'tree', whereby 'Decision Tree' classifier is selected as the weak learner. The ensemble method is set as 'Adaboost' algorithm, and the model is set with maximum number of splits of 20, and the number of learners as 30, with a learning rate of 0.1. The main advantage of using AdaBoost is that it feeds the weak outputs at each stage of the algorithm into the growing tree algorithm, to give more focus on misclassified outputs in order to improve classification accuracy and reduce errors. K-fold cross-validation was selected to validate model results, with five folds.

Table 4.8 demonstrates the performance results for EnsBoost-DT. As this classifier was used to improve the performance of DT classifier by its strong algorithm, which deals with weak classification outputs, it outperformed all previous classifiers in terms of all measurement

parameters. The results show the model achieved 96.2% average accuracy rate for the All-Data dataset. Also, the Ens-DT classifier was able to correctly classify 94.8%, 94.4%, and 94% of firms for the years 2019, 2018, and 2017, respectively. Therefore, the model preforms well and shows more stability for classifying firms' status. Moreover, the model achieved the lowest values of Type I Error among all other previous classifiers, with values ranging from a high of 8% for the year 2017 to a low of 6.2% for the year 2019. This enhancement in classifying failed companies makes the model more balanced and accurate for all years' datasets.

Table 4.8: ENS-DT results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	94.8%	94.4%	94%	96.2%
Type II Error	4.2%	3.8%	4%	4.8%
Type I Error	6.2%	7.5%	8%	3.5%
Sensitivity	95.8%	96.2%	96%	95.2%
Specificity	93.8%	92.5%	92%	96.5%
AUC	99%	99%	99%	99%
Brier Score	0.0414	0.0440	0.0470	0.0346
Area Under Reliability Curve	0.1194	0.1213	0.126	0.1681

Figure 4.15 demonstrates the ROC curves for Ens-DT. It shows the optimal performance of the classifier over the years, with the curves shifting up in small gaps from 2017 to 2019. Moreover, it has the highest AUC values among all classifiers.

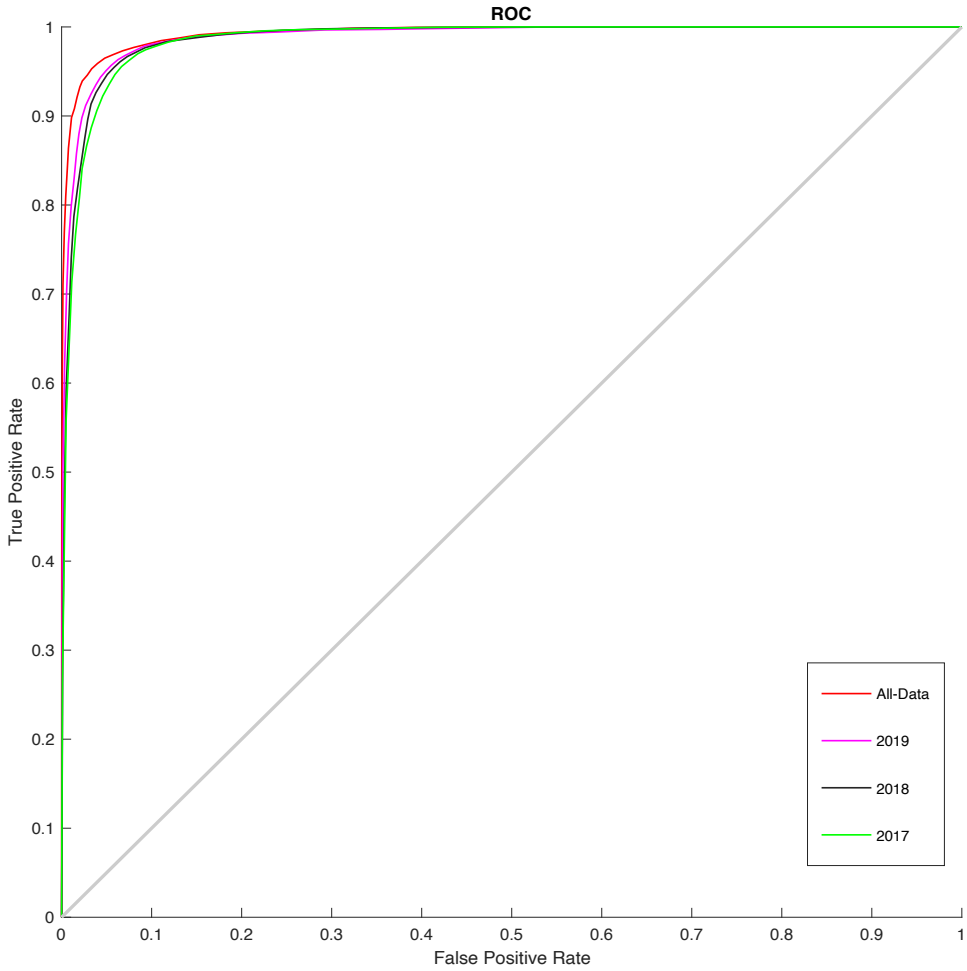

Figure 4.15: ROC curve for ENS-DT classifier

Figure 4.16 shows the reliability diagram of the Ens-DT classifier. The shape of the diagram is considered as the best so far, especially in comparison with DT classifier. The shape of the curve indicates a high calibration of the classifier, and high reliability in its classification.

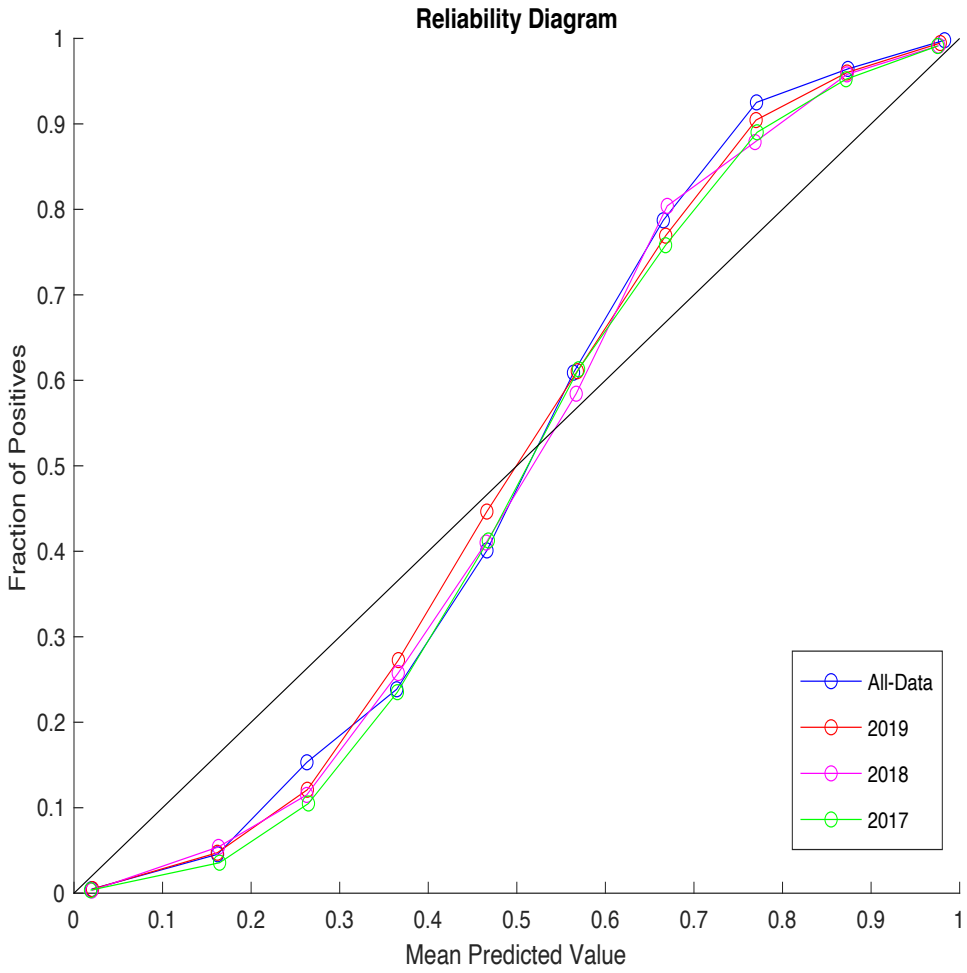

Figure 4.16: Reliability diagram for ENS-DT classifier

4.3.9. Deep Learning

DPL is a robust classification algorithm can be built using deep learning to solve binary classification by using multiple layers, which progressively extract information from the raw input data. Deep learning methods are commonly used to train data using supervised learning, providing input data to predict binary classification. The architecture of the model consists of building a layer-by-layer model.

The model is built using Long Short-Term Memory (LSTM), in which the core components of the network are a sequence input layer. For network creation, a layer containing a sequence input layer is implemented, followed by the LSTM layer, fully connected to a Softmax layer, linked to a classification output layer. SofMax activation function is selected because of its ability to handle multiple classes and its usefulness for output neurons. The input size is set as 24, representing the number of features of the input data used to feed the sequence input layer in the network. The number of hidden units is set as 24, and the number of classes is set as 2 (representing the two classes of the target output). The maximum epochs are set as 2000, and

the minimum batch size as 1000. After creating the optimal model structure, the model is trained and tested using a training and testing dataset extracted form original dataset (with percentages of 80% and 20%, respectively).

As shown in Table 4.9, the DPL classifier outperformed all of the other classifiers based on all measurements. The model achieved the highest average accuracy rates for all years' datasets, with the smallest gap of only 1% between year 2019 and year 2017. The average accuracy rate of 97.2% for the All-Data dataset is considered the highest in comparison with all other classifiers. Moreover, the model has the best performance of classifying both active and failed companies, illustrated by its lowest values of Type I and Type II Error. Based on the specificity and sensitivity measurements, the model has the most balanced classification capability to classify companies' status, as these values are similar. This allows the model to be ranked first, based on its superior performance as an individual classifier.

Table 4.9: DPL results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	96.3%	95.1%	95.3%	97.2%
Type II Error	2.8%	3.9%	3%	3%
Type I Error	4.5%	5.9%	6.3%	2.6%
Sensitivity	97.2%	96.1%	97%	97%
Specificity	95.5%	94%	93.7%	97.4%
AUC	99.35%	98.9%	98.82%	99.04%
Brier Score	0.0282	0.0370	0.0374	0.0237
Area Under Reliability Curve	0.0405	0.0414	0.045	0.0412

Figure 4.17 shows the ROC curves for the DPL classifier. Obviously, the model has the best ROC shape among all classifiers, with small shifts in curves from year 2017 to year 2019. For All-Data and the year 2019 the classifier has the highest AUC values of 99.04% and 99.35% (respectively).

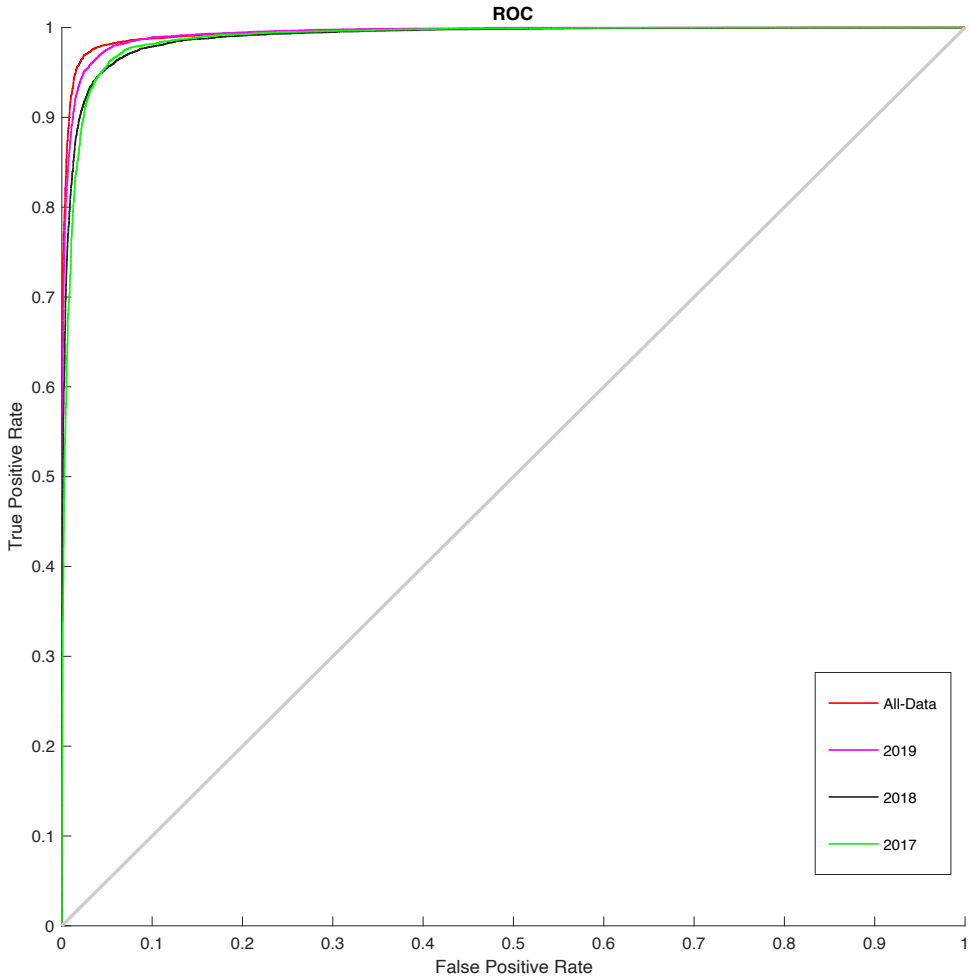

Figure 4.17: ROC curve for DPL classifier

Figure 4.9 shows the reliability diagrams of all datasets using DPL classifier. The classifier has the optimal shape, with prediction lines close to the diagonal line. This indicates the high classification performance of the classifier.

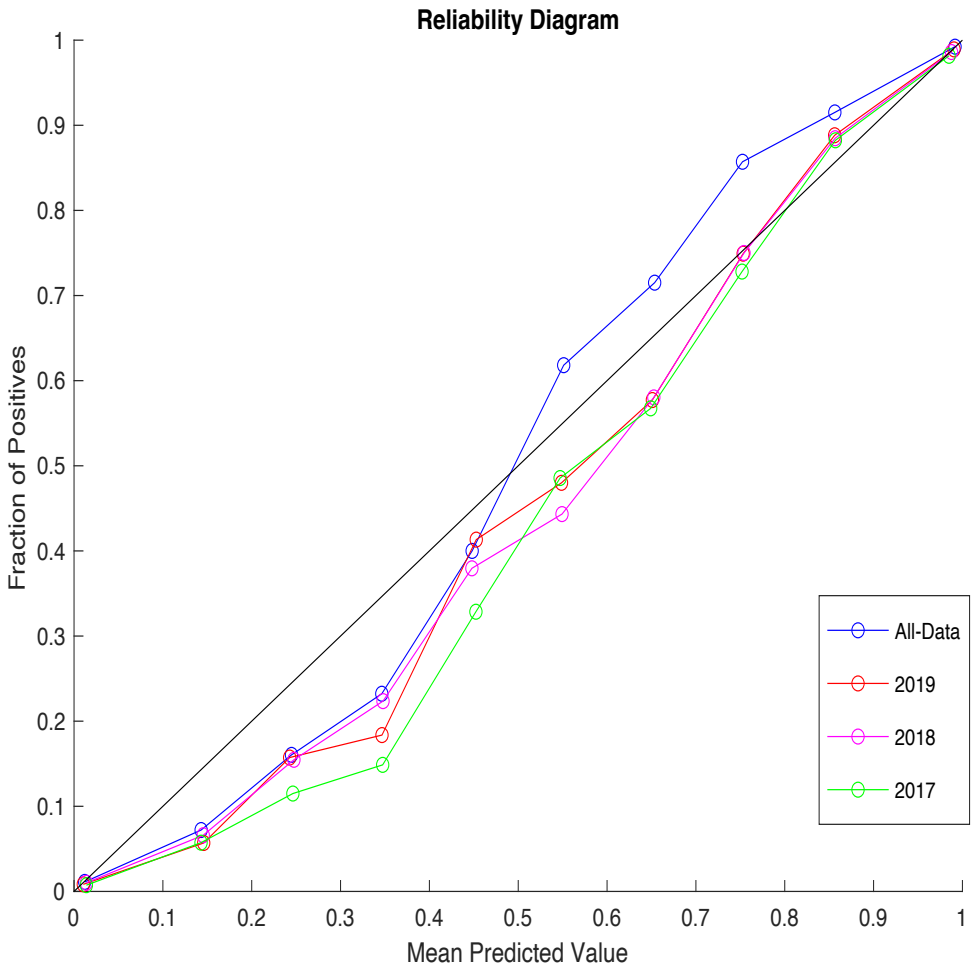

Figure 4.18: Reliability diagram for DPL classifier

4.4. Analysis and Discussion

This section analyses and discusses all classifiers' results to reveal the optimal classifier for classifying firm status. It is important to pin each classifier's strengths and weaknesses, to understand how they performed based on all measurement parameters. Tables 4.10 to 4.13 illustrate the classification performance of all classifiers related to each dataset used in this study.

The LDA classifier clearly had the worst results, with only 75% accuracy rate for the All-Data dataset, which is far below the other classifiers' results. The classifier was able to classify 2019 firms with an average accuracy of 71.5%, an increase of 2% in comparison to the outcomes for 2017 dataset. Moreover, its higher Brier score and AURC combined with lower AUC results show that the model has higher error for classifying companies for all years. Although the model has a balanced classification rate for both failed and active company status (based on its specificity and sensitivity rates), it only beat KNN classifier in terms of classifying failed firms

for the year 2019 dataset. However, the model achieved bad performance that resulted in it having the lowest rank as a classifier.

The LR classifier, the traditional classification method in the field of business failure, had similar results to KNN, in comparison to which it had superior ability to correctly classify failed companies for all years, but KNN sharply outperformed LR in classifying active companies. Both classifiers have relatively similar classification error values in total, based on their Brier scores and AURC values. Moreover, in comparison with other classifiers, both LR and KNN have low performance in classifying failed companies, as shown by their higher Type I Error rates and lower specificity rates (except for LDA).

NN classifier exhibited generally improved classification accuracy in comparison with LD, LR, and KNN. However, the majority of performance improvements were due to its powerful classification of active companies, when it is more important in this context to improve the classification of failed firms. For example, LR classifier has approximately similar specificity results to NN, and surprisingly it outperforms NN in predicting failed companies in year 2018 dataset.

SVM classifier performed substantially well in classifying failed companies. Despite having a lower average accuracy rate than NB, DT, ENS-DT, and DPL classifiers, it had a higher specificity rate with lower Type I Error, which reflects the true classification of failed companies. Therefore, SVM is considered as a more powerful classifier for detecting failed companies than LR, LD, KNN, NN, and NB classifiers. Moreover, it improved the classification of active companies, for which it ranked higher than NN. Its higher sensitivity and lower Type II Error rates among all year datasets represent a substantial improvement on the overall performance of previous classifiers. However, it is still more important to improve the classification of failed companies, which is useful for flagging concern about companies' health.

As the main purpose of the classifier to correctly identify failed companies and distinguish them from healthy ones, DT classifier, as a successor of SVM, had substantially improved classification of failed companies, achieving the highest specificity rates except for ENS-DT and DPL. It achieved a rate of 95.5% for the All-Data dataset, accompanied with a relatively higher sensitivity rates, which were higher than the other classifiers except for ENS-DT and DPL. Hence, DT classifier performance was more powerful and balanced than LR, LD, KNN, NN, SVM, and NB in classifying business status. Moreover, its Brier score had the lowest

values range from 0.0373 to 0.061, which is inversely related with the accuracy rates achieved. Therefore, DT classifier is more reliable in terms of classifying companies' health.

On the other hand, to improve weak classification of DT classifier, ENS_DT classifier achieved these improvements by increasing the accuracy of predicting companies' health, especially for failed ones. The model outperformed DT for all years' datasets and was the most balanced classifier discussed so far in this section. Also, the classifier showed an improvement in the Brier score. This overall outperformance of the classifier made it the most reliable classifier to be used for classifying companies' health overall for the abovementioned classifiers except for DPL.

DPL classifier outperformed all other individual classifiers based on all measurement parameters. The classifier had the highest average accuracy rate of 97.2% for All-Data, and 96.3%, 95.1%, and 95.3% for the years 2019, 2018, and 2017, respectively. Also, the model achieved the lowest Type I and Type II Error over all years' datasets. Tables 4.10 to 4.12 show the yearly classifier results for 2017-2019 (respectively), and Table 4.13 shows the All-Data results for all classifiers. It can be seen that DPL had the best classification accuracy rate of failed companies, and specificity rates, as well as the lowest Brier score, reflecting the model's low classification error. Therefore, it was ranked first as an individual classifier to be reliably used to classify the status of firms in the UK datasets.

Table 4.10: All classifiers 2017 results

	Year 2017 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier Score	AURC
LR	77%	26.3%	19.7%	73.7%	80.3%	86%	0.1531	6.85%
LD	69.5%	30.8%	30.1%	69.2%	69.9%	75%	0.2059	7.72%
KNN	79.4%	15.7%	25.5%	84.3%	74.5%	88%	0.1428	4.46%
NN	81.5%	18.8%	18.2%	81.2%	81.8%	90%	0.1454	11.3%
SVM	85.3%	13%	16.3%	87%	83.7%	93%	0.1845	28.25%
NB	86.8%	7.1%	19.3%	92.9%	80.7%	95%	0.1090	25.38%
DT	93%	5.2%	8.7%	94.8%	91.3%	97%	0.0576	12.03%
ESM	94%	4%	8%	96%	92%	99%	0.0470	12.6%
DPL	95.3%	3%	6.3%	97%	93.7%	98.82%	0.0374	4.5%

Table 4.11: All classifiers 2018 results

	Year 2018 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier Score	AURC
LR	83.9%	16.5%	15.7%	83.5%	84.3%	92%	0.1170	9.67%
LD	70.5%	30.4%	28.5%	69.6%	71.5%	76%	0.215	7.99%
KNN	80.4%	15.6%	23.5%	84.4%	76.5%	89%	0.1362	4.7%
NN	82.5%	17.5%	17.4%	82.5%	82.6%	91.2%	0.1212	8.81%
SVM	86.8%	11%	15.4%	89%	84.6%	94%	0.995	14.61%
NB	87%	7.8%	18.2%	92.2%	81.8%	95%	0.1090	25.4%
DT	93.6%	5.1%	7.6%	94.9%	92.4%	98%	0.4880	6.36%
ESM	94.4%	3.8%	7.5%	96.2%	92.5%	99%	0.440	12.13%
DPL	95.1%	3.9%	5.9%	96.1%	94%	98.9%	0.0370	4.14%

Table 4.12: All classifiers 2019 results

	Year 2019 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier Score	AURC
LR	80.15	22%	17.8%	78%	82.2%	89%	0.1328	5.74%
LD	71.5%	28.75	28.3%	71.3%	71.7%	77%	0.1978	8.25%
KNN	81.2%	14.7%	22.9%	85.3%	77.1%	90%	0.1311	5.34%
NN	83.1%	16.4%	17.4%	83.6%	82.6%	92.02%	0.1155	8.67%
SVM	86.7%	11.6%	14.9%	88.4%	85.1%	94%	0.1063	16.94%
NB	86.3%	8.5%	17.95%	91.5%	82.1%	95%	0.1098	29.08%
DT	94.2%	5.5%	6.5%	94.5%	93.5%	98%	0.0460	10.35%
ESM	94.8%	4.2%	6.2%	95.8%	93.8%	99%	0.0414	11.94%
DPL	96.3%	2.8%	4.5%	97.2%	95.5%	99.35%	0.0282	4.05%

Source: Author

Table 4.13: All classifiers' All-Data results

	All-Data Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier Score	AURC
LR	81.8%	18.8%	17.6%	81.2%	82.4%	91%	0.1211	3.85%
LD	75%	24.6%	25.4%	75.4%	74.6%	80%	0.1825	10.28%
KNN	82.9%	12.7%	21.6%	87.3%	78.4%	91%	0.1194	5.92%
NN	84.9%	13.7%	16.5%	86.3%	83.5%	93%	0.1073	9.58%
SVM	87.7%	10.9%	13.7%	89.1%	86.3%	95%	0.0843	7.96%
NB	89%	6.6%	15.3%	93.4%	84.7%	97%	0.0926	33.74%
DT	95.3	4.9%	4.5%	95.1%	95.5%	99%	0.0373	6.13%
ESM	96.2%	4.8%	3.5%	95.2%	96.5%	99%	0.0346	16.81%
DPL	97.2	3%	2.6%	97%	97.4%	99.04%	0.0237	4.12%

4.5. Summary

In this chapter, nine individual classifiers were used and demonstrated in an attempt to discover best classification model. Financial data related to 20,000 companies, divided equally into active and inactive, were used to develop and evaluate the classifiers based on eight performance measurements. At the beginning, data were pre-processed and cleaned from missing values. They were then fed to the classifier based on partition techniques, consisting of training and testing, using 10 × 5 cross-validation and holdout method. Finally, classifier parameters were set in order to achieve the optimal classification results out of 50 runs.

Based on each classifier results, LR as a benchmark was outperformed by all machine learning classifiers, and LDA classifier had the worst results. Classifiers' results varied from classifier to classifier, with some showing more capability in classifying one of the classes and less for others. However, DPL showed the best performance overall, and ranked first as the best classifier among the group, representing a major improvement in classification accuracy in comparison to LR. In the next chapter, the committee machine approaches are used in order to enhance classification performance using classifier outputs.

Chapter 5
Business Failure Classification using Committee Machine Classifiers

5.1. Introduction

As single classifier was implemented to classify business failure in the previous chapter. The current chapter addresses the question of how to combine and use classifier outputs to enhance classification performance. This can be done through combining functions, to convert single classifiers' predictions into new output classifications. There are different types of functions to convert single classifier classification results as input data x_i, to be fed to a new combining classification model in the form of $f\ (x_1.\ x_2.\ x_3.\ x_4.\ x_5) = x_*$. In an attempt to explore how single classifiers could work together, committee machine learners are implemented using the following rules algorithms:

- Min Rule (MIN)
- Max Rule (MAX)
- Median
- Consensus (Cons)
- Majority Voting
- Weighted Average
- Fuzzy Combiner

5.2. Committee Machine Classifiers

This section demonstrates the used combiners and their mathematical combining functions, noting their strengths and weaknesses in terms of classification performance for the studied data type. Each combiner is illustrated in a mathematical diagram representing each step of the model. Thereafter, some recommendations are discussed to address the fitting of the model deployed.

5.2.1. Min Rule

The operation rule of Min combiner allocates the minimal value among all single classifier predictions for a single company. According to Figure 5.1, the model selects predictions that are considered to have the lowest value among all classifiers, and then compares them with the

optimal threshold of Max Rule. Therefore, the final output for each company is the lowest prediction value extracted from all single classifiers' results for that company. Although it is a simple rule for combining and selecting each company classification, the model is highly affected by the predictions of classifiers with low sensitivity rate performance. For example, if a company's actual status is active, with a class of 1, and it has been correctly classified by all single classifiers except one classifier, which misclassified it as failed, with a class of 0, the Min Rule will select the latest value as the lowest among all predictions to be the final prediction output, therefore it will incorrectly classify the company as failed. As a result, in most cases Min Rule predicts failed companies much better than active companies.

In order to overcome this weakness in the model, two crucial steps should be considered and implemented: omitting the predictions of all classifiers with low sensitivity rates, to avoid allocating false negative predictions (0 class); and adjusting the predictions threshold using threshold lowering, by which all classifiers' predictions is scaled based on a lower threshold in an attempt to avoid misclassifying of active companies. Only the latter was implemented in this work to enhance combiner predictions, using the optimal threshold.

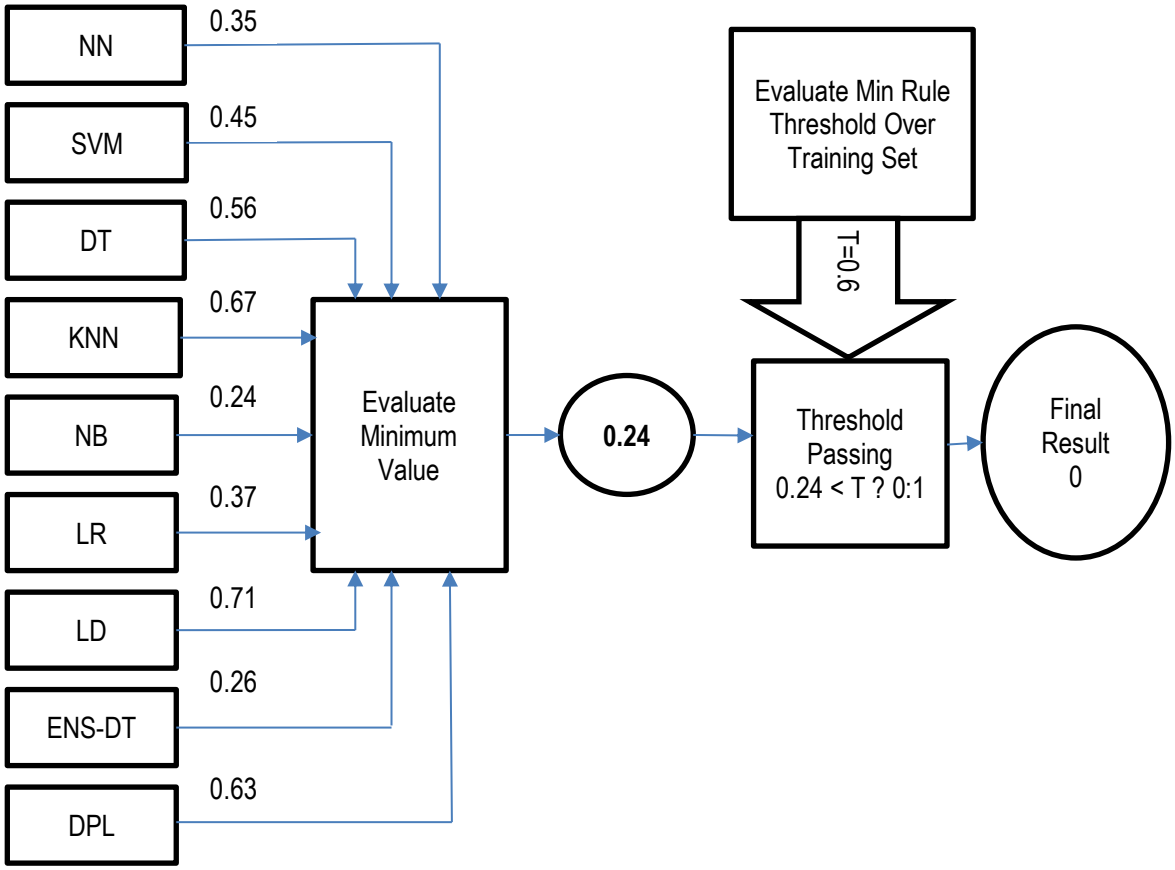

Figure 5.1: MIN combiner example

5.2.2. Max Rule

The combiner operates by selecting the maximal prediction value of all classifier predictions for each company as a final prediction output. Based on Figure 5.2, the Max Rule combiner takes the highest classification value among all classifier prediction values and then compares it to the optimal threshold obtained for training set using Min Rule. Its function is similar to Min Rule, but in the converse, whereby the combiner is more affected by the results of the classifiers with the lowest specificity rates. Using the default threshold, the combiner classifies active companies better than failed ones, which it results in higher sensitivity performance but very low specificity. For instance, if a failed company been classified correctly with a class of '0' (True Negative) based on all classifiers except for one, in which it was classified as active (False Positive), the Max combiner will select the false positive vale as the final prediction, and consequently misclassify the company as active. Therefore, the predictions of the classifiers with low specificity rate should be removed, and a higher threshold needs to be implemented to overcome the biased of the combiner. In this case, MAX combiner is favourable when the number of active companies is higher than that of failed ones.

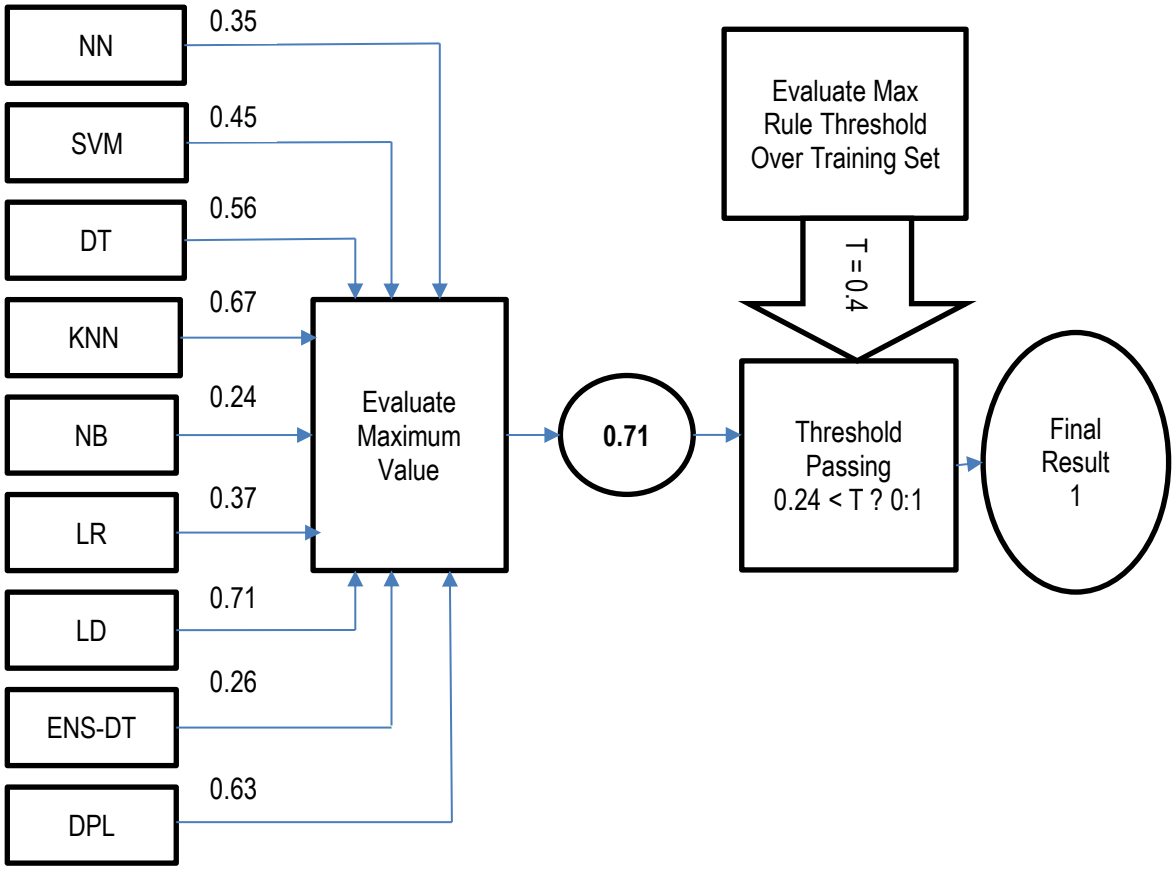

Figure 5.2: MAX combiner example

5.2.3. Average Rule

The combiner is developed by taking the mean value of all classifier predictions to be the final output classification. According to Figure 5.3, AVG combiner sums all single classifier predictions of a single company and then divides this by the number of classifiers used, to compare the output result with the default threshold. Unlike Min and Max combiner, adjusting the threshold for the predictions is unnecessary when it stays at its default value (0.5). AVG is a more reliable combiner than MAX and MIN, since it includes all classifier predictions in the mathematical calculation of the final output. Moreover, another advantage of the combiner is its balanced performance for sensitivity and specificity rates, when the data has a balanced input of both classes. On the other hand, a disadvantage of using AVG combiner appears when the predictions of single classifiers significantly vary in values, in which case the calculation of the mean could be biased because of outliers, resulting in wrong classification.

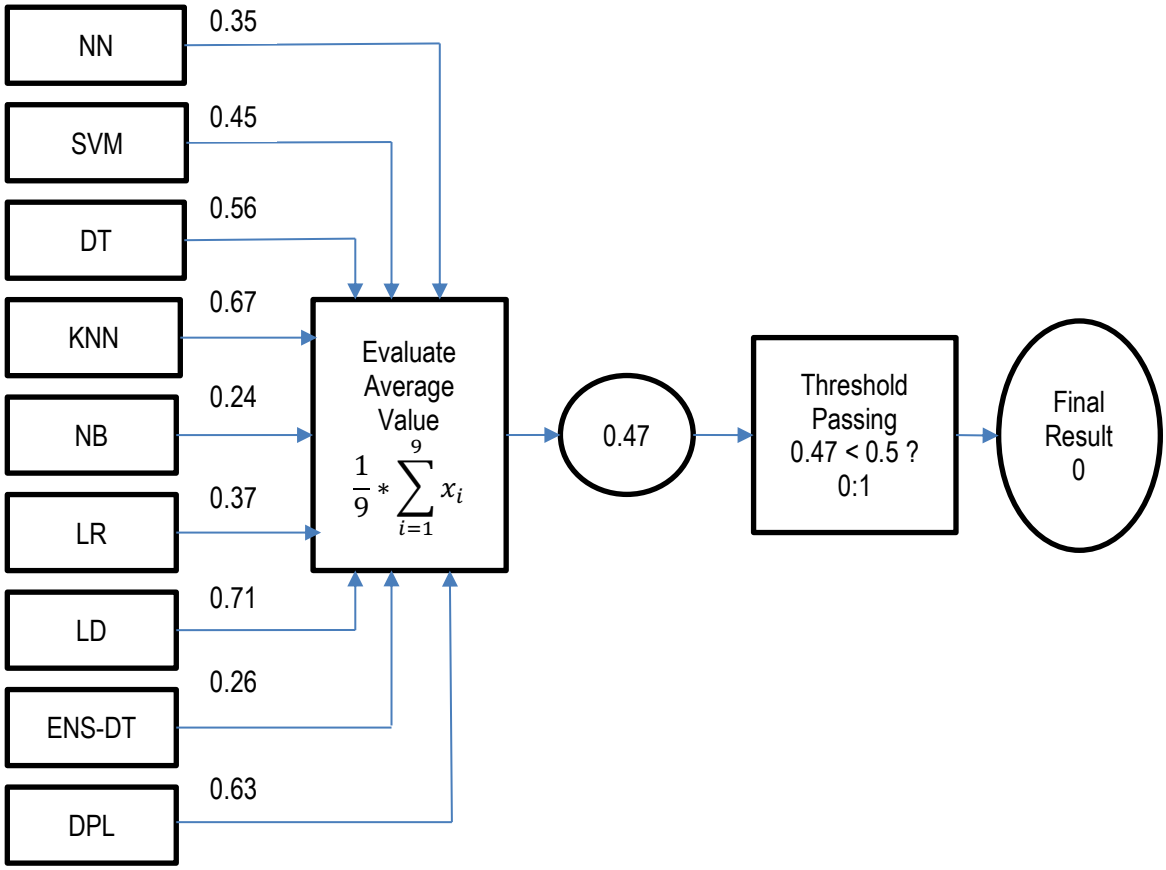

Figure 5.3: AVG combiner example

5.2.4. Median Rule

Median Rule combiner ranks all classifier predictions in ascending order and takes the value of the median prediction. Based on Figure 5.4, after sorting each company's predictions for all classifiers in an ascending order, the combiner takes the midpoint to be the final answer. Similar to AVG combiner, changing the threshold is an unnecessary step, and it also has a default value of 0.5. An advantage of using Median rule is that the final output is unaffected by the prediction's outliers (extreme low or high scores), since these scores is out of the calculation of the final output, as the combiner focuses only on the middle scores. However, a disadvantage appears when the number of the input values (number of classifiers) is even, whereby more calculation is needed to get the final result. This disadvantage is considered insignificant, since the number of classifiers is odd (9), and the median score needs no extra calculation. Median combiner is considered to be less reliable than AVG, as it ignores most classifier prediction values and focuses only on the middle-ranked prediction.

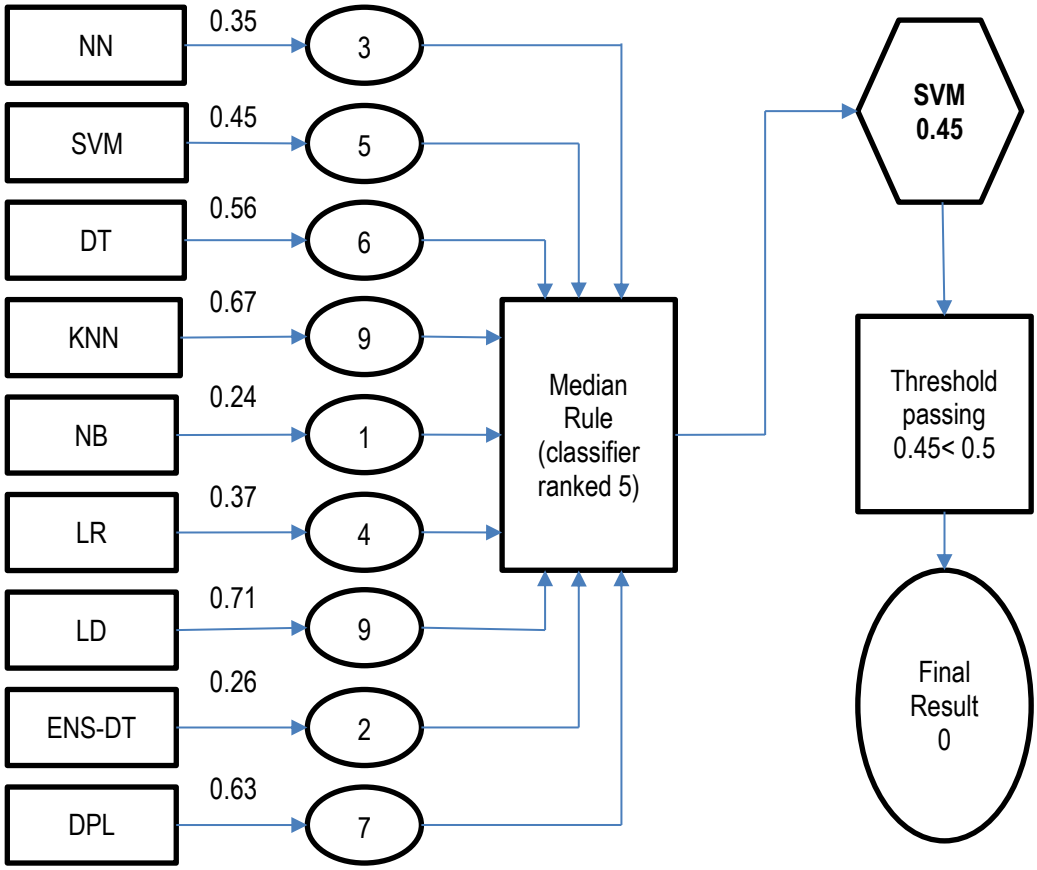

Figure 5.4: Median combiner example

5.2.5. Weighted Average Rule

The combiner rule here takes the average score of all classifier predictions in terms of each classifier's associated weight, based on the overall performance of the classifier. According to Figure 5.5, the WAVG combiner has similar calculations to AVG; while both combiners calculate and take the average of the classifiers scores, there is a difference in relying on the enhancement of WAVG final results, resulting from the allocation of higher weight to classifiers with higher accuracy, and lower weight to those with lower accuracy. Hence, weights' coefficients are evaluated based on each single classifier's performance, which allows more accurate classifiers to contribute more to the final output; conversely, less accurate classifiers contribute less. Due to these features, WAVG combiner is considered more reliable than the normal AVG combiner, and it performs with higher accuracy.

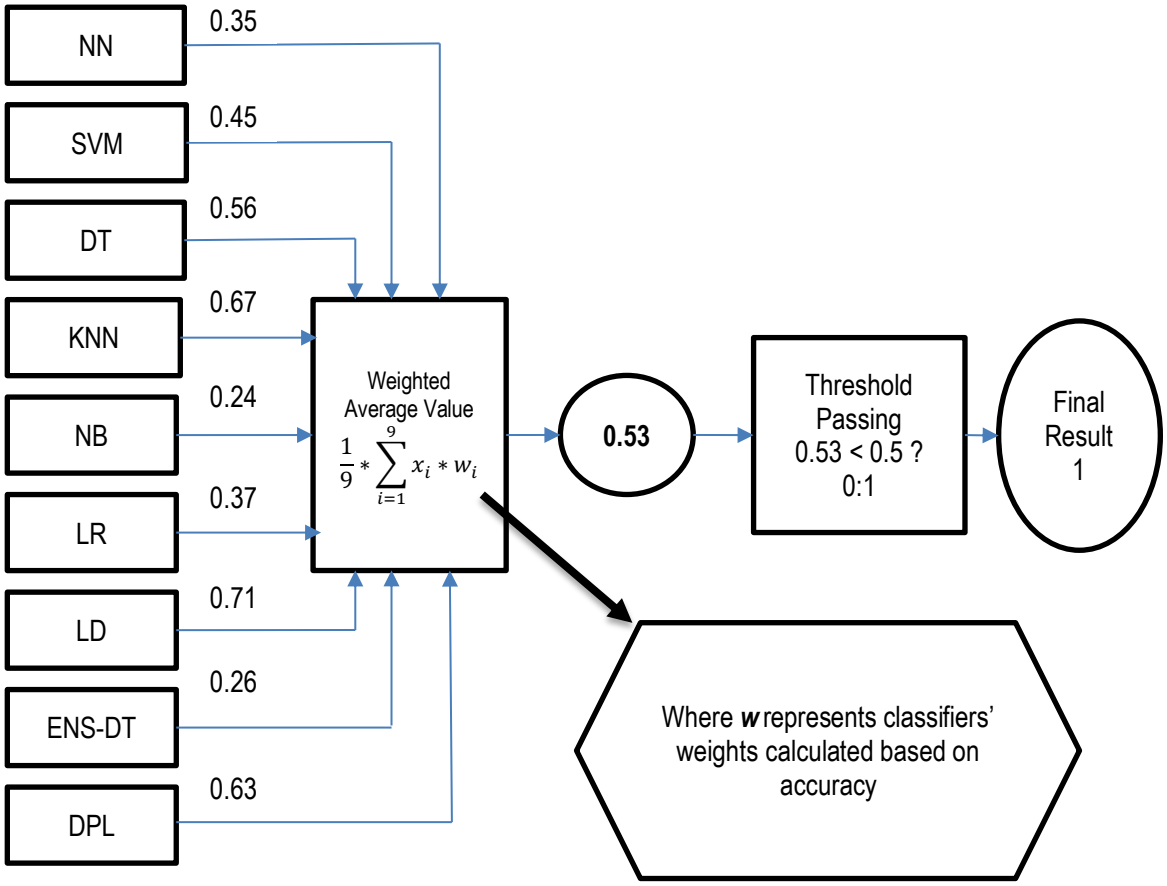

Figure 5.5: Weighted AVG combiner example

5.2.6. Majority Voting Rule

Majority voting combiner calculates the final output result based on voting method, in which the final prediction for a company is made based on the most frequent score or class on which the majority of classifiers agree. As displayed in Figure 5.6, the combiner firstly rounds all prediction values to the nearest integer (0 or 1), and then selects the most agreed class as the final answer for each company. An advantage of using this combiner is that it incorporates classes unlikely be misclassified by most single classifiers. Therefore, its results comply with those with higher accuracy rates, and are less affected by bad efficiency classifier predictions when the majority of classifiers classify the same company correctly. Moreover, the combiner does not require changing the threshold, since final results are based on comparing the numbers of votes given for each score of the classifiers.

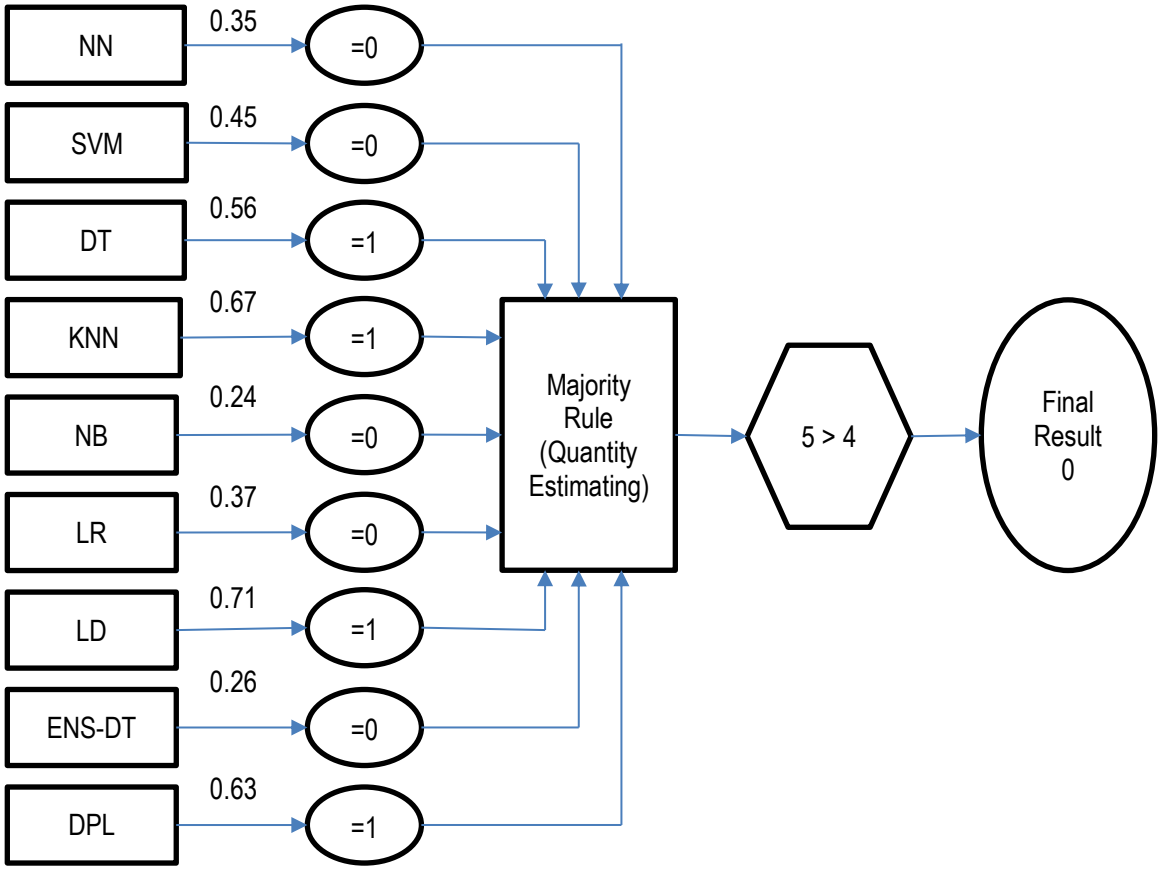

Figure 5.6: Majority voting combiner example

5.2.7. Consensus Combiner

The aim of the combiner is to minimise the uncertainty of the decisions made by all classifiers through a discourse process between each classifier and other classifiers involved in the ensemble. This combiner mechanism enables each classifier to check its results in regard to other classifiers' results, in order to calculate the certainty of its results, based on which weights are assigned to be part of the calculation of the final output. Consensus algorithm is considered a powerful decision-making technique to enhance heterogeneous classifier performance. Figure 5.7 shows the design of the Consensus combiner and explains the recursive process of a linear function for adjusting classifier weights to reach final decisions.

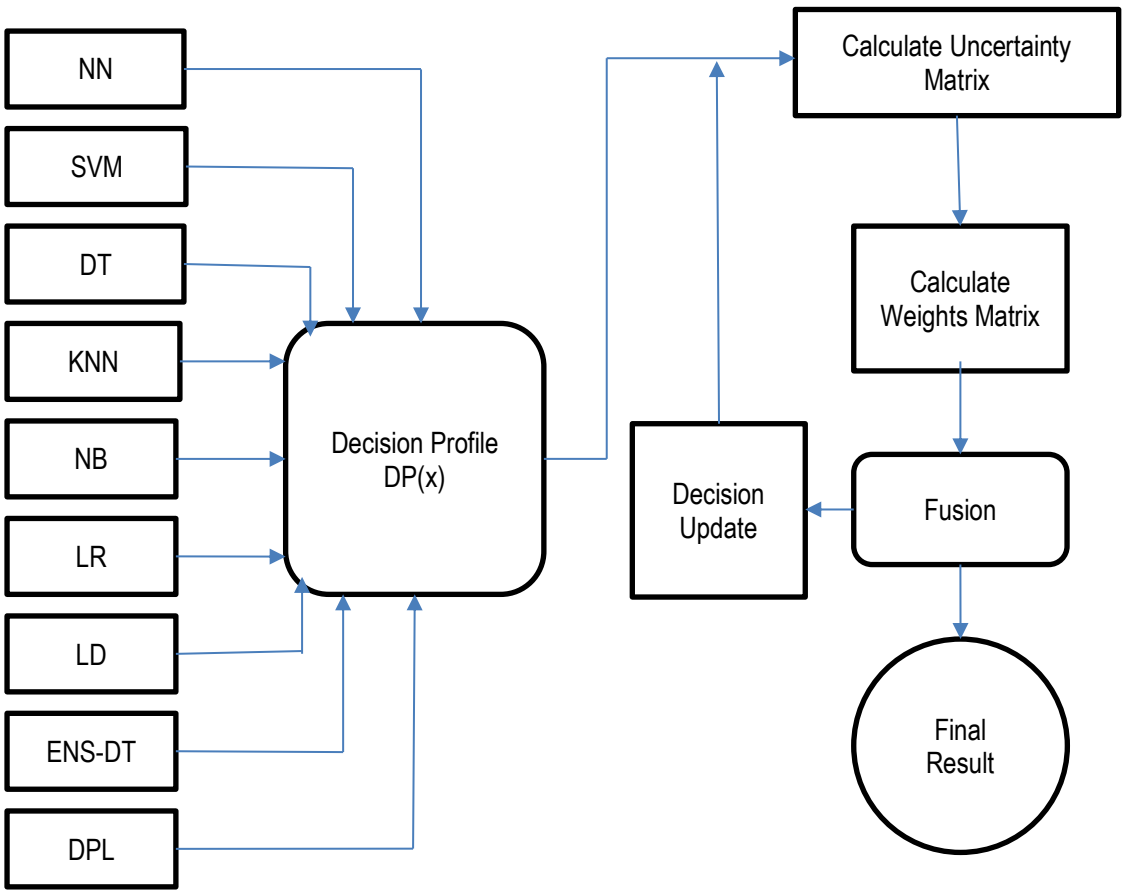

Figure 5.7: Cons combiner example

The combiner algorithm operates the following steps, explained in detail in the following subsections (based on MATLAB coding):

- Build a Profile Decision for each classifier.
- Calculate the uncertainty of each classifier's output, then create an uncertainty matrix with the results.
- Calculate new classifier weights based on uncertainty levels.
- Update weights.

5.2.7.1. Step 1: Decision Profiles

The Consensus combiner combines the nine single classifiers' outputs to generate the final output classifications. The first step is to build a decision profile for all classifiers. To explain the process, the nine individual classifiers are denoted by $C_i = C_1, C_2, C_3, ..., C_9$, which generate classification outputs of possible answers $R = (y_1, ..., y_m)$. For each possible output of the single classifier, an estimate A_i function is assigned to all answers. The estimates of A_i take values

between ‘0’ and ‘1’, to represent the desirability of the corresponding output (weights). The calculation of the initial weighted values of each classifier is based on the following equations:

$$\sum_{k=1}^{m} A_i(y_k) = 1 \; \forall i \in \{1..N\} \tag{5.1}$$

$$\sum_{k=1}^{m} A_i(\gamma_k|\Gamma_j) = 1 \; \forall i \in \{1..N\} \tag{5.2}$$

Once all classifiers been given initial weights, the decision profile is represented by:

$$DP = \begin{bmatrix} A_1(r_1) & A_1(r_2) & A_1(r_3) & \dots & A_1(r_n) \\ A_2(r_1) & A_2(r_2) & A_2(r_3) & \dots & A_2(r_n) \\ A_3(r_1) & A_3(r_2) & A_3(r_3) & \dots & A_3(r_n) \\ A_4(r_1) & A_4(r_2) & A_4(r_3) & \dots & A_4(r_n) \\ A_5(r_1) & A_5(r_2) & A_5(r_3) & \dots & A_5(r_n) \\ A_6(r_1) & A_6(r_2) & A_6(r_3) & \dots & A_6(r_n) \\ A_7(r_1) & A_7(r_2) & A_7(r_3) & \dots & A_7(r_n) \\ A_8(r_1) & A_8(r_2) & A_8(r_3) & \dots & A_8(r_n) \\ A_9(r_1) & A_9(r_2) & A_9(r_3) & \dots & A_9(r_n) \end{bmatrix} \tag{5.3}$$

where n is the number of companies on the dataset, r_i *is* the *i-th* classifiers predictions, and $A_j(r_i);j \in 1\ldots9$ is the j-th weighted value of each prediction of each classifier included in the combiner. The resulting DP matrix is considered the initial stage of the combiner, and the next step is to calculate the uncertainty between classifiers using this DP matrix.

5.2.7.2. Step 2: Calculating Uncertainty

After generating the initial Decision Profile matrix for all companies point by point, the next step is to calculate the uncertainty of these observation using an appropriate calculation function. Thereafter, weights are calculated using uncertainty values, and are assigned to each classifier decision, with higher weights assigned to more certain classifier decisions. Matrix creation at this point involves local and global uncertainty. Local uncertainty evaluates the level of certainty of each classifier about their decisions, whereas global uncertainty evaluates the level of certainty of the classifier decisions after knowing other classifiers’ decision in the combiner. Hence, 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 result about a company's status of all classifiers together. These calculations allow each classifier to improve its decision in regard to other classifiers' performance. The outcome of this process is an uncertainty matrix that includes all classifiers' uncertainty values, which are later used to calculate the new weights in succeeding steps. The form of the uncertainty matrix is displayed below:

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} & \dots & U_{1N} \\ U_{21} & U_{22} & U_{23} & \dots & U_{2N} \\ U_{31} & U_{32} & U_{33} & \dots & U_{3N} \\ U_{41} & U_{42} & U_{43} & \dots & U_{4N} \\ U_{51} & U_{52} & U_{53} & \dots & U_{5N} \\ U_{61} & U_{62} & U_{63} & \dots & U_{6N} \\ U_{71} & U_{72} & U_{73} & \dots & U_{7N} \\ U_{81} & U_{82} & U_{83} & \dots & U_{8N} \\ U_{91} & U_{92} & U_{93} & \dots & U_{9N} \end{bmatrix} \quad (5.4)$$

where local uncertainty is donated by U_{ij}, i, j $\in$1…9 for the i-th classifier when i = j; and global uncertainty is donated by U_{ij}, i, j $\in$1…9 for the i^{th} classifier when i $\neq$ j. To illustrate the evaluation of the uncertainty levels on the matrix, assume $A(\gamma k)$ is the weighted output of a single classifier decision, and $A_i(\gamma k|\Gamma j)$ is the weighted output of a single classifier decision when it knows the decision of another classifier *j* on the combiner. Based on that, the following equations are used to calculate both local and global uncertainty:

$$U_{ii} = \sum_{k=1}^{n} A_i(\gamma_k) \log_M(A_i(\gamma_k)) \quad (5.5)$$

$$U_{ij} = \sum_{k=1}^{n} A_i(\gamma_k|\Gamma_j) \log_M(A_i(\gamma_k|\Gamma_j)) \quad (5.6)$$

Eq. 5.5 represents the calculation of local uncertainty of i-th classifiers in the combiner, and Eq. 5.6 represents global uncertainty based on another classifier's output. These two equations are applied after equations 5.5 and 5.6 are fulfilled. For binary classification output '0' and '1', where M = 2, the above equations are converted into the following:

$$A_i(0) + A_i(1) = 1, A_i(0|\Gamma_j) + A_i(1|\Gamma_j) = 1 \quad (5.7)$$

where A(1) is the weight of class '1' decision of i-th classifier, and A(0) is the weight of class '0' decision. If $A_i = A(1)$ and $A(\Gamma_j) = A_i(1|\Gamma j)$, then $A_i(0) = 1 - A_i$, and $A_i(0|\Gamma j) = 1 - A_i(\Gamma j)$.

The resulting equations are as follows:

$$U_{ii} = -A_i \log_2(A_i) - (1 - A_i)\log_2(1 - A_i) \quad (5.8)$$

$$U_{ij} = -A_i(\Gamma_j)\log_2(A_i(\Gamma_j)) - (1 - A_i(\Gamma_j))\log_2(1 - A_i(\Gamma_j)) \quad (5.9)$$

Uncertainties are calculated in the above equations using a logarithm to the base '2' (number of classes). Based on equation 5.1 and 5.2, the more the ranking value of the classifier decision is close to the edges of [0,1] interval, the less the uncertainty value with values almost near zero. In contrast, the closer the classifier ranking is to the threshold (0.5 by default), the higher the uncertainty. Figure 5.8 illustrates the relationship between classifier ranking and uncertainty level.

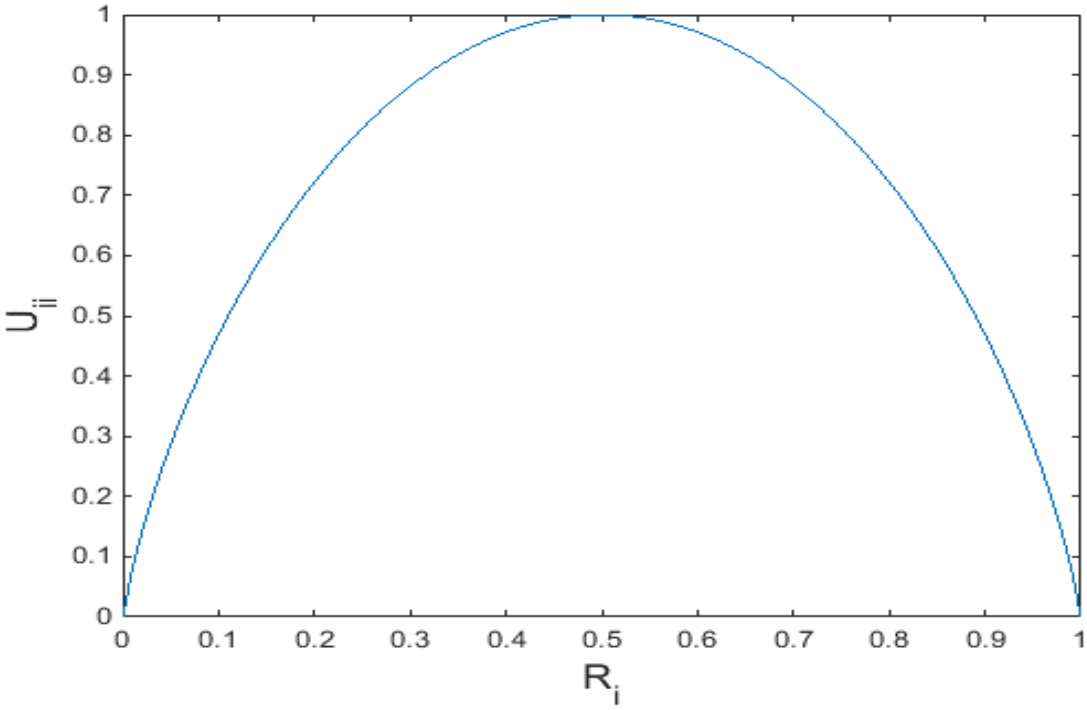

*Figure 5.8: Uncertainty value **U**ii as a function of the parameter **R**i*

5.2.7.3. Step 3: Calculating Weights

Once all uncertainty values of the classifiers have been calculated and presented in a matrix, the next step is to calculate the new weights of the classifiers. The calculation of the new weights depends on the uncertainty level of the classifier, as included in the calculation equation. The new weights are presented in a matrix the same as the uncertainty matrix. The following equation shows the weights calculation:

$$W_{ij} = \frac{1}{U_{ij}^{2}\Sigma_{k\in A}U_{ki}^{-2}} \tag{5.10}$$

After applying the equation to all classifier decisions, these new weights are assigned to each related classifier.

5.2.7.4. Step 4: Aggregating the Combiner Scores to Calculate the Decision

To generate the final decision of the combiner when all classifiers obtain their decisions based on uncertainty-adjusted weights, all these decisions are summed together, given adjusted weights, and the aggregation of these weights equals '1'. The aggregation is calculated using the following equation:

$$A_g(\gamma_k) = \sum_{k=0}^{n} A_i(\gamma_k) * w_{ij} \tag{5.11}$$

Moreover, these weights are updated and adjusted based on an error modification factor, in order to improve the final decision of the combiner on the following final step.

5.2.7.5. Step 4: Updating Classifiers' Weights

After all weights been calculated and assigned to all classifiers and a decision is made using the aggregation equation, the weights go through another updating step based on the weights modification factor. These updates are executed based on the iteration process of the algorithm. The error is calculated by taking the difference between each decision of the combiner inside the iteration loop. The weights modification factor 'sigma' is calculated for each classifier, taking the percentage of error values to the total aggregated errors of all classifiers. The following equations illustrate the updating weights process:

$$E_{ij} = 1 - \sum_{i=1}^{n} |A_i(\gamma_k) - consA_i(\gamma_k)| \tag{5.12}$$

$$Sigma_{ij} = \frac{0.1 * E_{ij}}{|\sum_{k=i}^{n} E_{ij}|} \tag{5.13}$$

$$W_{new} = W_{ij} + Sigma_k \tag{5.14}$$

where E_{ij} is the predication error between previous and adjusted weight predictions of the classifier, and Sigma is the adjusted modification error.

5.2.8. Fuzzy Logic Combiner

Fuzzy Logic is a convenient method to map input to output data. It is conceptually germane to an easy decision-making process, based on its simplicity of mathematical computations. Flexibility and less complexity allow more functionality to be added to it without having to start again from scratch. Moreover, it builds more understanding of the data space, such as modelling nonlinear functions of arbitrary complexity, in which it allows for more tolerance of

imprecise data that can match any set of input-output data. This is justifies its usefulness in application as a classifier to classify companies' status. Therefore, a fuzzy logic algorithm is implemented using single classifiers' outputs (predictions) as input data space, to be mapped with actual companies' status, in an attempt to improve companies' classification. The following subsections illustrate the development steps of the proposed classification model.

5.2.8.1. Step 1: Calculating Single Classifiers' Means and Standard Deviation

The initial step to build the fuzzy logic algorithm is the calculation of means and standard deviation values for both predictions and actual targets of single classifiers. These calculations are calculated using reliability functions, whereby predictions are assigned to 20 bins based on their values, then the means and standard deviation values are calculated based on predictions and targets values in each bin. After these values have been calculated and assigned to each classifier, confidence level and pooled standard deviation can be calculated.

5.2.8.2. Step 2: Calculating Confidence Levels and Pooled Standard Deviation

In this step the confidence level and pooled standard deviation is calculated in order to measure how much each classifier is confident about its predictions. As the purpose of the fuzzy logic is to map input data space to an output data space, these values are considered as the parameters to measure how close input classifiers' predictions are to the targets. The following equations show the calculation of these values:

$$A = 1 - \sqrt{\frac{(P_i - X)\text{^}2 + (T_i - x)\text{^}2}{N}} \tag{5.15}$$

$$\sigma_{PT} = \sqrt{\frac{\sigma_p^2 + \sigma_T^2}{N}} \tag{5.16}$$

where A is the confidence level, P_i and T_i are the *i-th* classifier prediction and target value (respectively), x is the optimal mean for the same prediction bin, and σ_{PT} is the pooled standard deviation of the prediction and the target laying in the same bin. The next step is calculating the predictions fuzzy sets using these parameters and the classifiers predictions.

5.2.8.3. Step 3: Applying Fuzzy Function

After assigning all prediction and target means and standard deviation values related to each classifier, and all confidence level and pooled standard deviation calculations are set accordingly, a membership function is applied to define how each point in the input space is mapped to the output by giving prediction a membership value. The input data comprise the classifiers' predictions values, referred to as the universe of discourse. The output of the membership function is a number known as the membership value, between 0 and 1, which is designated μ.

To satisfy the purpose of the membership function, assume a dataset set F is expressed as $F = \{x \mid x > 0.05\}$. A fuzzy set is an extension of this set and is expressed as $F = \{x, \mu F(x) \mid x \in X\}$, where $\mu F(x)$ is the membership function of x in F, when X is the universe discourse whose elements are donated by x. The results are mapped for each value of the X universe to a membership value between 0 and 1.

There are 11 built-in membership function types available on MATLAB based on the following four basic functions: Piecewise Linear Functions, Gaussian Distribution Function, Sigmoid Curve, and Quadratic and Cubic Polynomial Curves. The piecewise linear distribution creates a nonparametric representation of the cumulative distribution function (cdf) by linearly connecting the known cdf values from the sample data. This function defined by different functions for each part of the range of the entire function that has a discontinuity at one or more values mainly because of the denominator of a function is being zero at those points.

The sigmoidal membership function, which is either open left, right, and Asymmetric and closed (i.e. not open to the left or right) membership functions can be synthesized using sigmoidal functions. Using this function is not beneficial in the case since it is biased to one of the edges of the function either the left side or the right side of the function.

The Polynomial based curves account for several of the membership functions in the toolbox. Three related membership functions are the Z, S, and Pi curves, all named because of their shape. The function zmf is the asymmetrical polynomial curve open to the left, *smf* is the mirror-image function that opens to the right, and *pimf* is zero on both extremes with a rise in the middle. However, the Polynomial models have poor extrapolatory properties. Polynomials may provide good fits within the range of data, but they will frequently deteriorate rapidly outside the range of the data which increase the uncertainty.

However, in our case, a simple Gaussian distribution function is selected to calculate all classifiers' membership function values. This function has the advantage over the piecewise function in our case due to the smoothness of the shape of the function and the continuity of the function. Another advantage of the Gaussian function is that the data is in the centre instead of the left and right edges as in the sigmoidal function. Moreover, it has less uncertainty than polynomial function which make it more beneficial for the datasets used to classify business failure. This function computes fuzzy membership values using a Gaussian membership function. The following equation shows the Gaussian function deployed:

$$G(C_j) = EXP\left(\frac{-(X - C_j).^2}{2 * \sigma PT_j^2}\right) \tag{5.17}$$

where C is classifiers' predictions, and X, is the discourse universe. It is crucial to take into consideration the rule's weights. In this algorithm, the weights are defined by the confidence level values of each single classifier element, added to the equation as in the following:

$$G(C_j) = EXP\left(\frac{-(X - C_j).^2}{2 * \sigma PT_j^2}\right) * A(c_j) \tag{5.18}$$

This is to take into consideration the effect of confidence in each classifier's output. After all elements of all classifiers been given a membership function value, an average function is performed to assign the final decision set of the fuzzy function for all classifiers in the combiner.

5.2.8.4. Step 4: Aggregating All Outputs

The step represents the unify process of all classifiers' fuzzy sets by joining their parallel threads. This is done through taking all fuzzy sets of each single classifier on the algorithm and combining them into a single fuzzy set. The following equation shows the aggregation process:

$$\text{Combined Fuzzy Set} = \frac{\sum_{k=1}^{j} G(C_j)}{J} \tag{5.19}$$

The process only occurs once for each element of all the classifiers. The inputs for the aggregation function are the truncated output functions returned by the implication process for each classifier. The result of the aggregation is one fuzzy set, in preparation for the fifth and final step, defuzzification.

5.2.8.5. Step 5: Defuzzifying Output

In this step, as the final step of the algorithm, the aggregated output fuzzy set becomes the input for the defuzzification process, and the results are single numbers (whereby crispness is recovered from fuzziness at last). During the intermediate steps, fuzziness enables the model to evaluate the outcomes and encompasses them in a range of output values of each element, whereby the defuzzification helps to assign the final output of it as a single value, to be the final prediction answer for each company. The following shows the defuzzify command:

$$\text{Final\_Predictions} = \text{defuzz}(x, \text{mf}, \text{type}) \tag{5.20}$$

The results are defuzzified value out, of a membership function mf positioned at associated variable value x based on defuzzification strategies, according to the argument and type. The variable type can be set according to the following methods:

- centroid: centroid of area method.
- bisector: bisector of area method.
- MOM: mean of maximum method.
- SOM: smallest of maximum method.
- LOM: largest of maximum method.

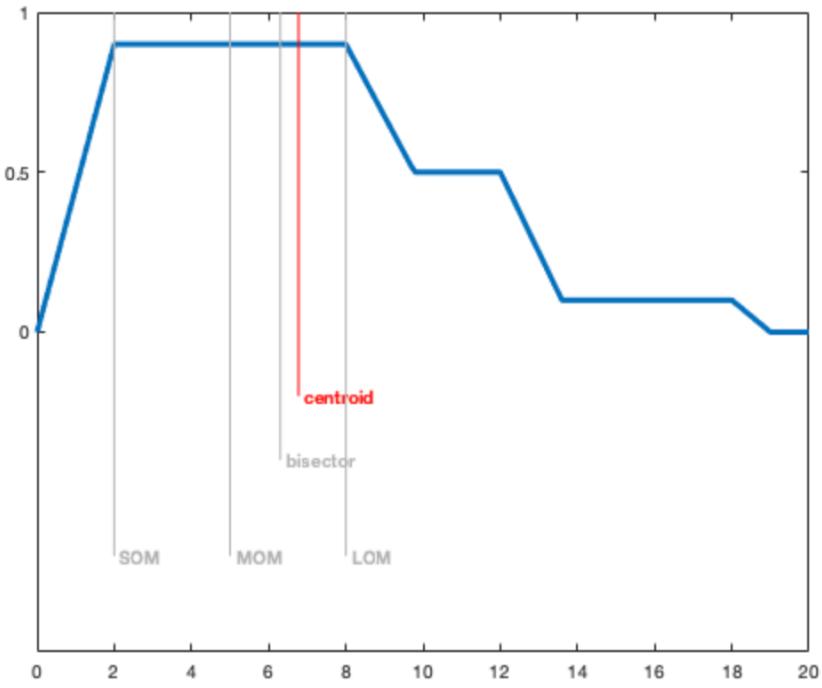

Figure 5.9: Example of defuzzification methods

According to Figure 5.9, the centroid method calculation is the most popular method for defuzzification; it returns the centre of area under the curve (centre of gravity). However, the Bisector method select the output in the area under the output fuzzy set and finds the vertical line that divides the fuzzy set into two sub-regions of equal area. MOM, SOM, and LOM stand for middle, smallest, and largest of maximum, respectively, in which MOM is the mean of the values for which the output fuzzy set is maximum, LOM is Largest value for which the output fuzzy set is maximum, and SOM is the smallest value for which the output fuzzy set is maximum as shown in the Figure above. After applying all of the above methods to classify companies' status, centroid method was found to achieve the highest balanced accuracy rate in which the output is the centre of the area under the output fuzzy set. The following example illustrates this method:

x = -10:0.1:10

mf = *trapmf*(x,[-10 -8 -4 7])

xx = *defuzz*(x,mf,'centroid')

5.3. Experimental Results

In this section, all of the results of testing data are generated using six traditional combiners, and are evaluated based on eight performance measures. These results are measured by using 10 x 5 cross-validation, in which the result is the average of 50 testing sets. The input data comprises the base single classifier predictions generated in the previous chapter, whose combination (using combination rules) enhances prediction performance. Tables 5.1 to 5.6 demonstrate the results of each combination method used across all data sets.

5.3.1. Min Rule Results

Table 5.1 illustrates the classification results achieved by the MIN combiner. The average accuracy rates achieved by the datasets were as follows: All-Data (95.5%), 2019 (94.4%), 2018 (94.1), and 2017 (93.8). Obviously, the combiner gives better results for active company classification than for failed companies, as indicated in higher sensitivity rates over specificity. The reason for the imbalance between sensitivity and specificity rates is due to adjusting the threshold. Consequently, this combiner rule is favourable when there are more active companies than failed ones in the dataset.

Table 5.1: MIN combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	94.4%	94.1%	93.8%	95.5%
Type I Error	7.3%	7.7%	8.4%	5.6%
Type II Error	3.8%	4.2%	4.1%	3.5%
Sensitivity	96.2%	95.8%	95.9%	96.5%
Specificity	92.7%	92.3%	91.6%	94.4%
AUC	98.54%	98.16%	98.17%	99.19%
Brier Score	0.0664	0.0688	0.0736	0.0430
Area Under Reliability Curve	0.1497	0.1393	0.1337	0.1362

Figure 5.9 shows the ROC curve representing the classification performance of the combiner model across yearly and All-Data datasets. It is clearly that All-Data has the best curve, shifted up with a higher gap between it and the 2019 curve. The AUC values evaluating the gap difference between these curves indicate that All-Data achieved the largest value. Moreover, the shifting upward movement of the curve across years from 2017 to 2019 reflects increasing AUC values.

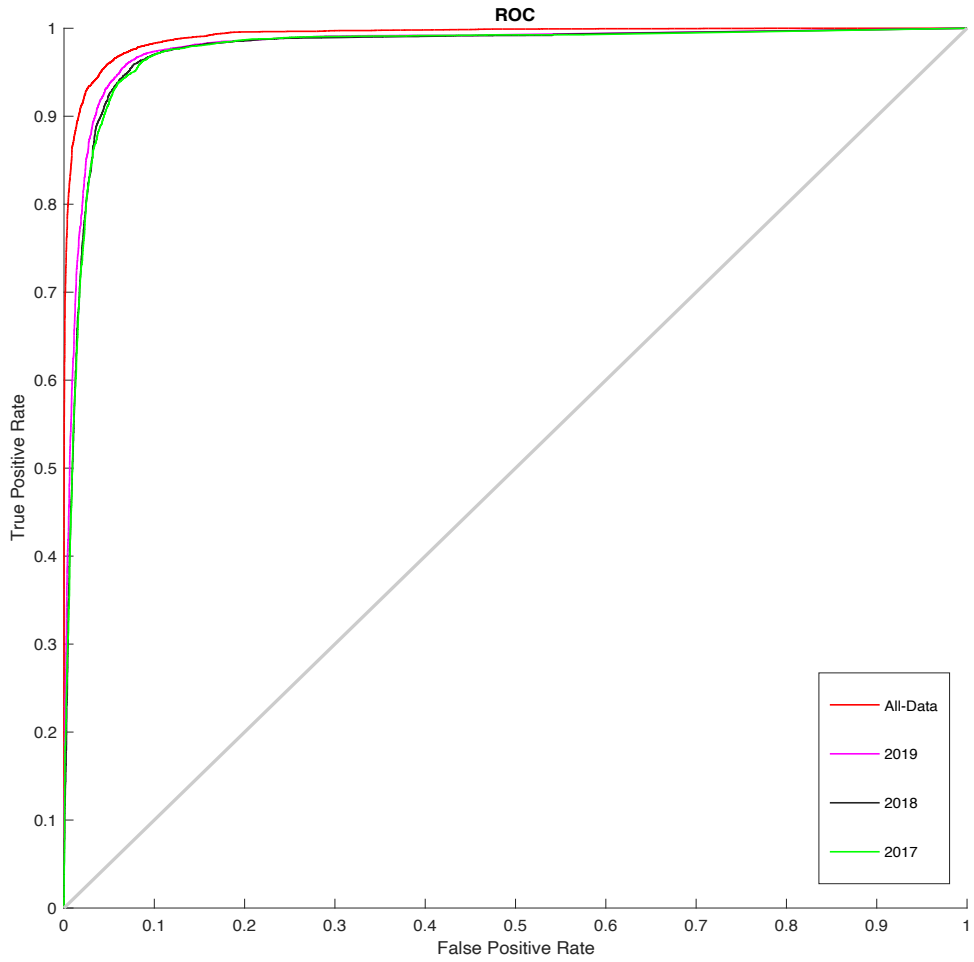

Figure 5.9: ROC curve for MIN combiner

In Figure 5.10, the reliability diagram indicates a highly calibrated curve for All-Data, and all years' datasets have relatively similar line shapes. However, the part of the curve above the 0.5 threshold has a better shape than those below it, illustrating the imbalanced accuracy rates between classifying active and failed companies. The All-Data dataset has higher Type I Error, pertaining to misclassifying failed companies, which results in the wide area between the reliability and diagonal lines. Moreover, it has higher AURC values, representing the amount of the misclassification of the combiner.

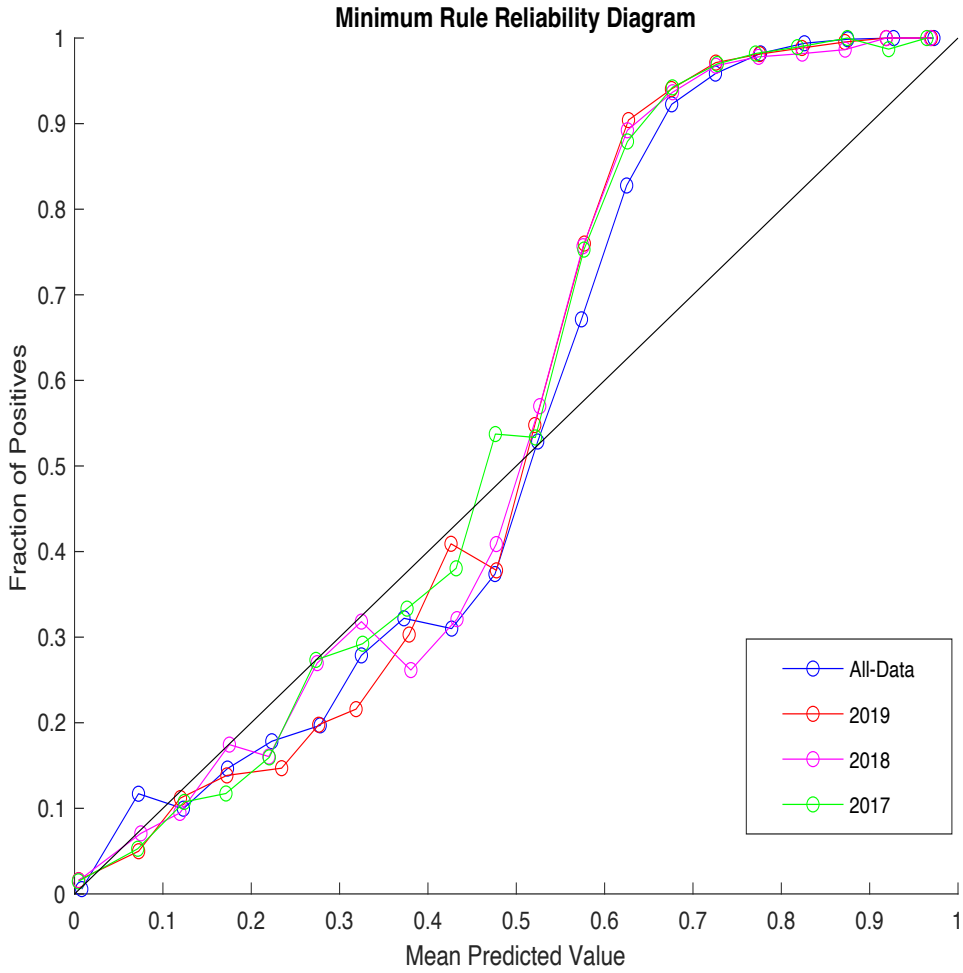

Figure 5.10: Reliability diagram for MIN combiner

5.3.2. Max Rule Results

Max Rule combiner is similar to Min Rule combiner, but it calculates final predictions in the opposite way, and it is better to describe and analyse their results together. Both combiners strongly rely on the outperformance of single classifiers and how well they have been trained. According to Table 5.2, the average accuracy rate of MAX combiner is relatively close to those achieved by the Min combiner, with 96% for All-Data, and 94.1%, 93%, and 92.7% for year 2019, 2018, and 2017, respectively. Also, in contrast to Min Rule, Max Rule combiner works better in classifying failed companies, which is shown in its higher specificity rates. However, the combiner has less gap between specificity and sensitivity rates in comparison with Min Rule results. On the other hand, its Brier scores are higher, especially for All Data, indicating higher classification errors.

Table 5.2: MAX combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	94.1%	93%	92.7%	96%
Type I Error	4.4%	5.5%	6.6%	3.1%
Type II Error	7.3%	9.5%	8%	4.9%
Sensitivity	92.7%	91.5%	92%	95.1%
Specificity	95.6%	94.5%	93.4%	96.9%
AUC	97.51%	97.29%	97.21%	99.34%
Brier Score	0.0724	0.0781	0.0837	0.0411
Area Under Reliability Curve	0.1699	0.1534	0.1642	0.1378

According to Figure 5.11, the ROC has a good shape for all datasets. The curve shifted up across the years, showing higher movement than the Max Rule ROC. This is shown in Table 5.2, where the combiner achieved higher AUC values in comparison with Max Rule.

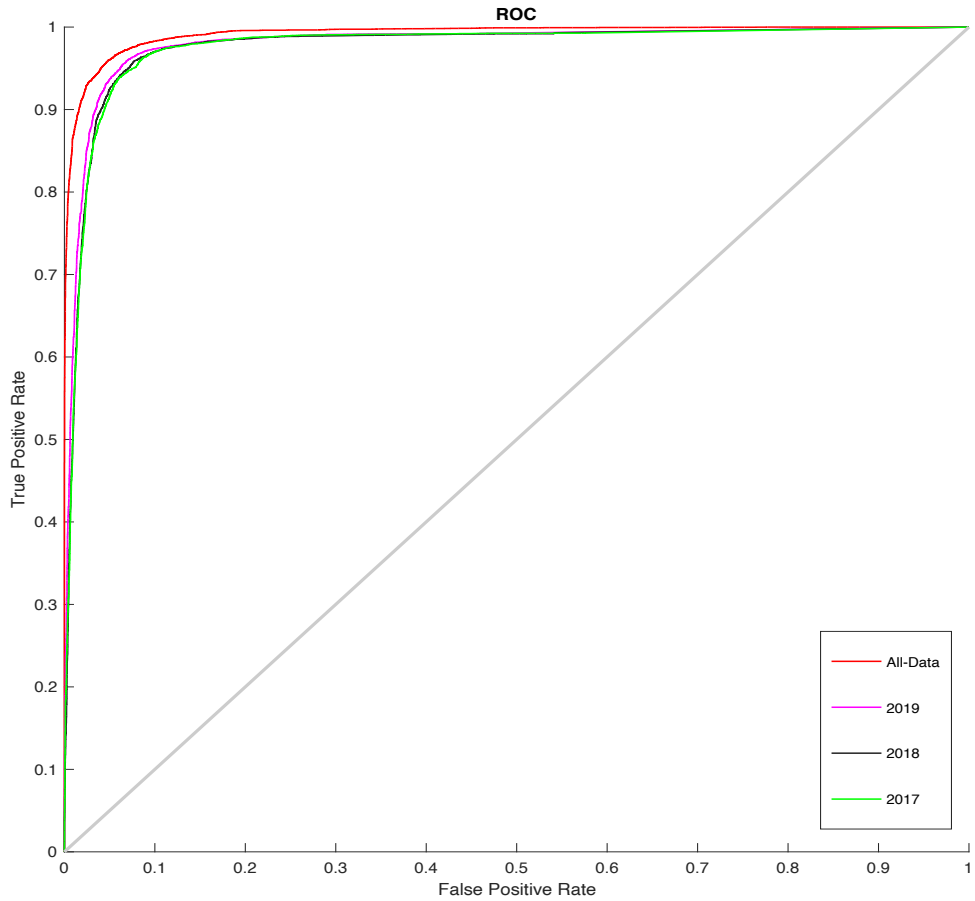

Figure 5.11: ROC curve for MAX combiner

Source: Author

Figure 5.12 demonstrates the reliability diagram for the Max Rule. The line has a perfect shape for companies with the '0' class (failed). This can be explained by the higher specificity rate achieved by the rule, which is reflecting the combiner performance of classifying failed companies correctly. The shape of the diagram illustrates the favourability of the Min Rule to classify '0' class in contrast with the Max Rule.

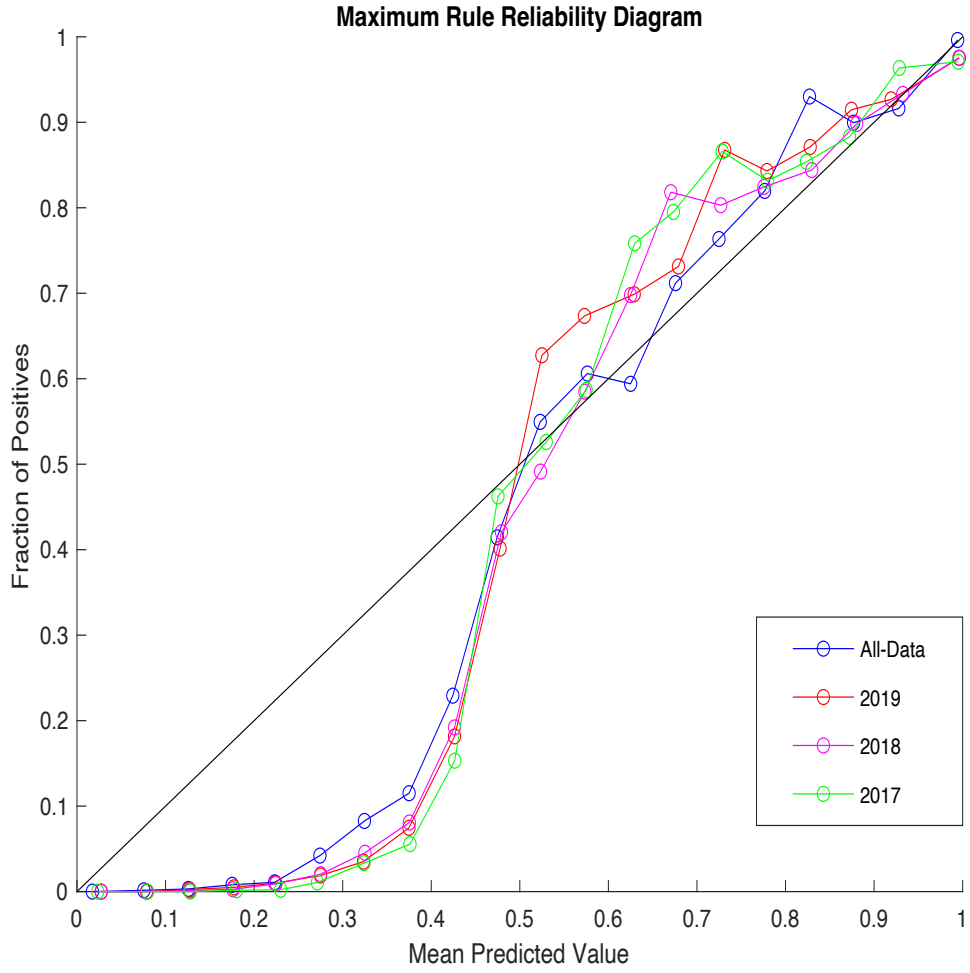

Figure 5.12: Reliability diagram for MAX combiner

5.3.3. Median Rule Result

According to Table 5.3, the Median rule combiner works well on the All-Data dataset and achieved an average accuracy of 96.5% which is higher than the equivalent accuracy of the individual classifier that ranked second (DT classifier). However, the combiner shows worse performance than DT, Max, and Min Rules combiners for each year's datasets. As the combiner assigns the median value across all predictions of all nine classifiers used in the combiner, its final answer is sensitive to the default threshold used in the rule. Hence, it is obvious from the specificity and sensitivity rates that the combiner is more capable of correctly classifying active

companies than failed ones. Moreover, based on the All-Data results, the combiner has a better Brier score value than both Max and Min combiners, reflecting lower error of classification.

Table 5.3: Median combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	92.6%	92.5%	91.5%	96.5%
Type I Error	8.1%	8.8%	10.1%	3.3%
Type II Error	6.7%	6.3%	6.8%	3.7%
Sensitivity	93.3%	93.7%	93.2%	93.3%
Specificity	91.9%	91.2%	89.9%	96.7%
AUC	98.15%	98.07%	97.72%	99.53%
Brier Score	0.619	0.0623	0.0743	0.0342
Area Under Reliability Curve	0.1458	0.1399	0.1852	0.1734

Source: Author

Figure 5.13 shows the ROC graph for the Median rule combiner. The combiner produced a better curve for All-Data than the Max and Min Rule combiners. Moreover, the curve outperformed those curves related to the year 2018 and 2017 datasets.

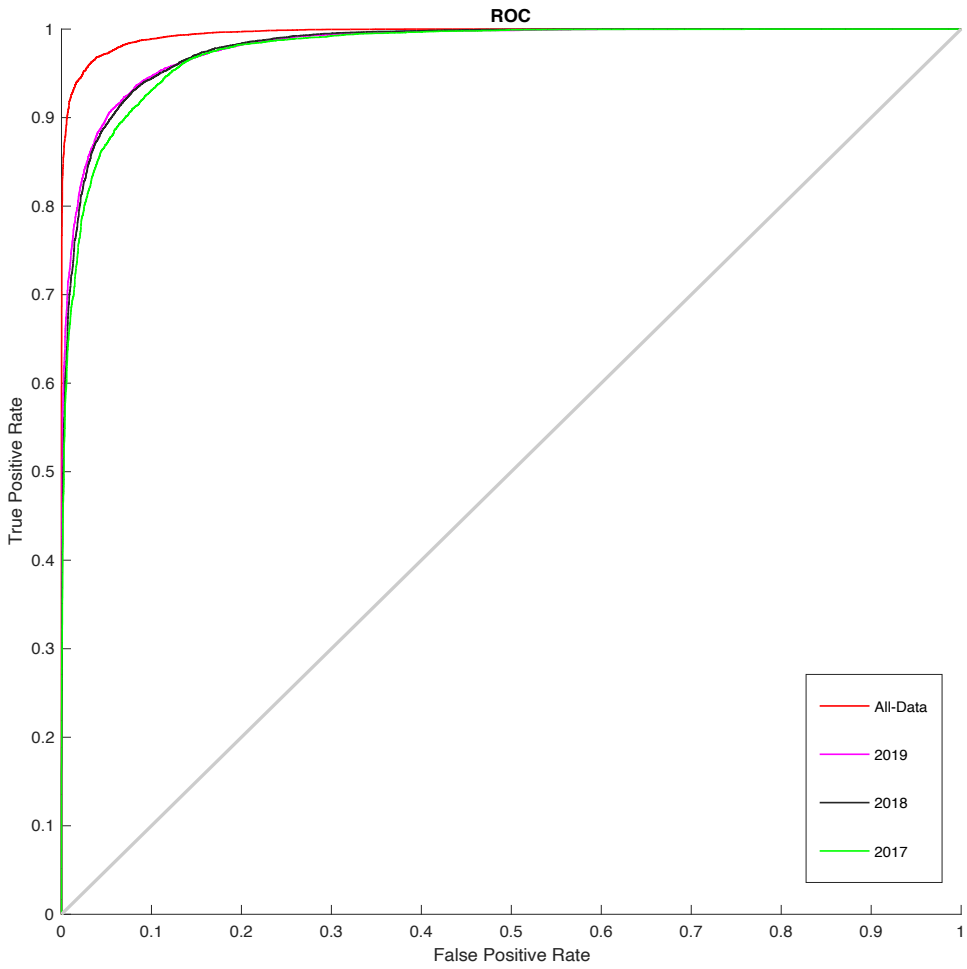

Figure 5.13: ROC curve for Median combiner

According to Figure 5.14, the Median rule combiner has an approximately an optimal line shape in the reliability diagram. Data is equally assigned to its designated bins in a balance, whereby half of the data falls in the first 10 bins, which refer to failed data; and the other half falls in the remaining 10 bins, related to active companies.

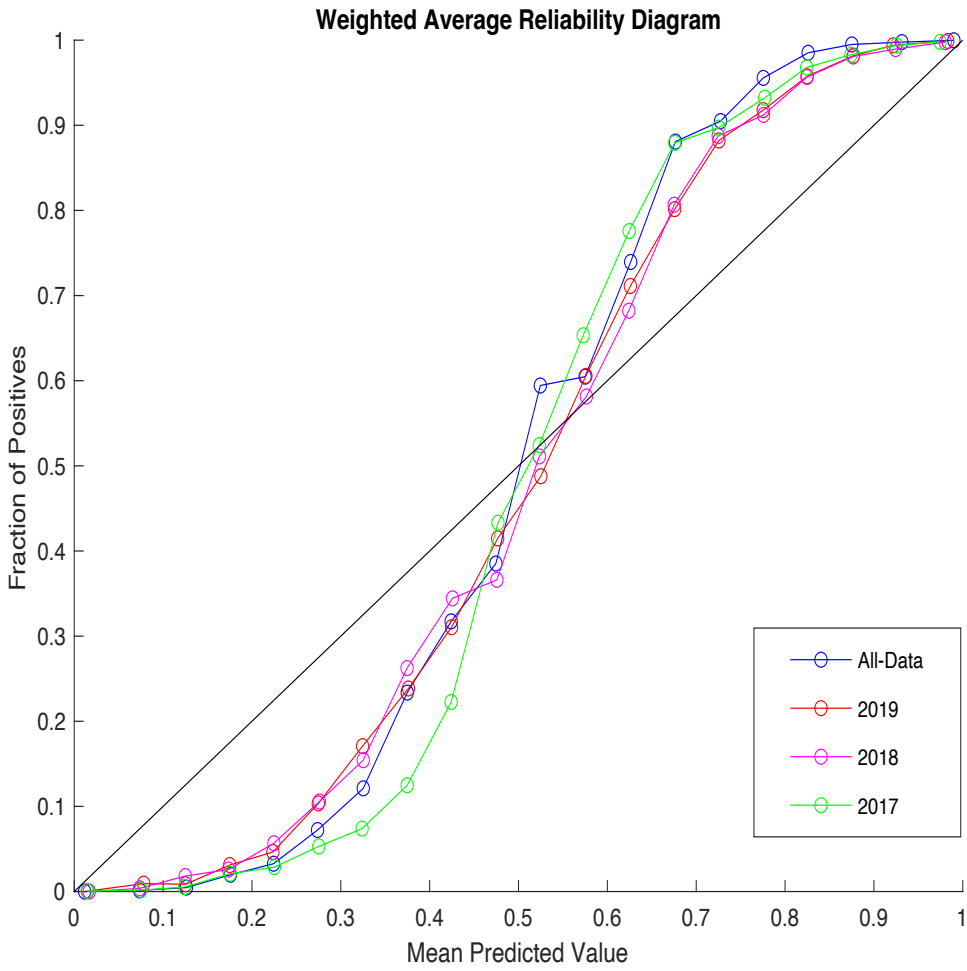

Figure 5.14: Reliability diagram for Median combiner

5.3.4. Average Rule Results

Table 5.4 shows the performance results for the Average rule combiner. The combiner achieved 96.2% in average accuracy for All-Data, and 94.5%, 94.1%, and 94% for 2019, 2018, and 2017, respectively. According to the results of the specificity and sensitivity rates, the combiner shows higher performance in classifying active companies over failed companies. However, the combiner outperformed Min, Max, and Median for the year 2018 and 2017 datasets in terms of average accuracy.

Table 5.4: AVG combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	94.5%	94.1%	94%	96.2%
Type I Error	6.5%	7.4%	8.1%	4%
Type II Error	4.6%	4.3%	3.8%	3.5%
Sensitivity	95.4%	95.7%	96.2%	96.5%
Specificity	93.5%	92.6%	91.9%	96%
AUC	98.7%	98.55%	98.51%	99.5%
Brier Score	0.0646	0.0662	0.0777	0.0451
Area Under Reliability Curve	0.1982	0.1956	0.2217	0.2010

Figure 5.15 shows the ROC graph for the Average rule combiner. The curves show the good performance of the combiner in classifying companies in the dataset in general. The combiner achieved higher AUC values in comparison with Min, Max, and Median combiners.

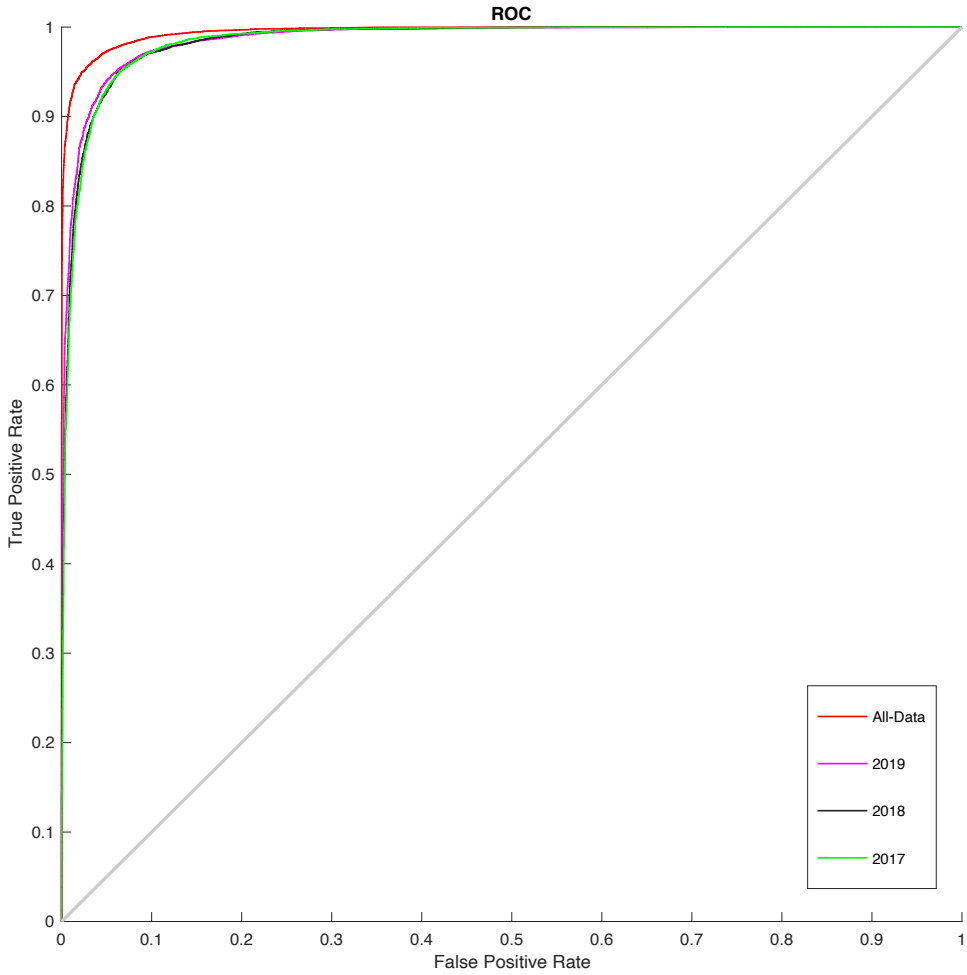

Figure 5.15: ROC curve for AVG combiner

Figure 5.16 shows the reliability diagram for the Average rule combiner. Similar to the Median, the combiner has an optimal shape with All-Data dataset, with the closest line to the diagonal.

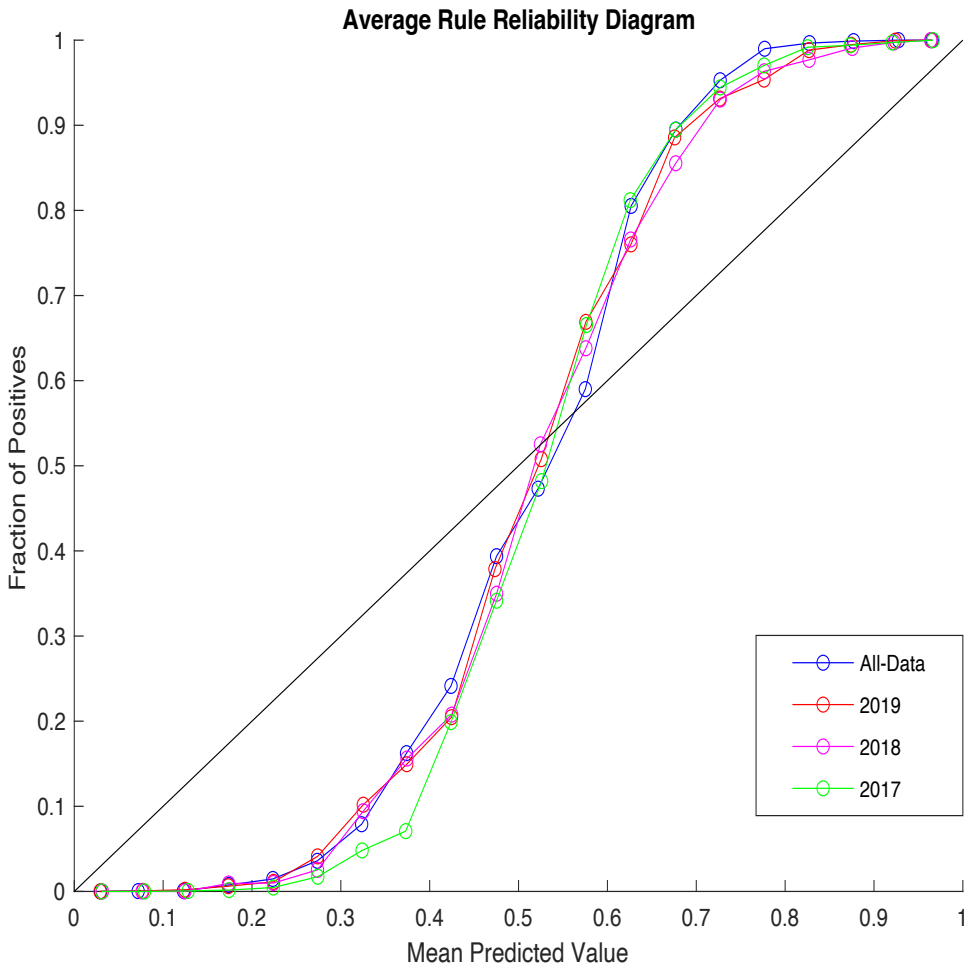

Figure 5.16: Reliability diagram for AVG combiner

5.3.5. Majority Rule Result

According to Table 5.5, the Majority rule combiner results show good performance for All-Data and year 2019, resulting from higher specificity rates than sensitivity rates. This means the combiner classify failed companies more accurately than active ones. However, for the years 2018 and 2017 the case changed, and specificity rates became less than sensitivity ones.

Table 5.5: Majority combiner results

	Year 2	**Year 3**	**Year 4**	**All-Data**
Average Accuracy	96.3%	93.8%	92.6%	96.3%
Type I Error	3.2%	7.5%	9.9%	3.2%
Type II Error	4.3%	4.9%%	5.4%	4.3%
Sensitivity	95.7%	95.1%	94.6%	95.7%
Specificity	96.8%	92.5%	90.6%	96.8%
AUC	99.48%	97.94%	97.21%	99.48%
Brier Score	0.0342	0.0487	0.0576	0.0342
Area Under Reliability Curve	0.1681	0.0682	0.1209	0.1681

Figure 5.17 shows that the Majority Rule combiner achieved good performance. The curve shifts up as it moves from the year 2017 dataset to the following years' datasets. This can be noticed in the increase of AUC values across years' datasets in Table 5.5.

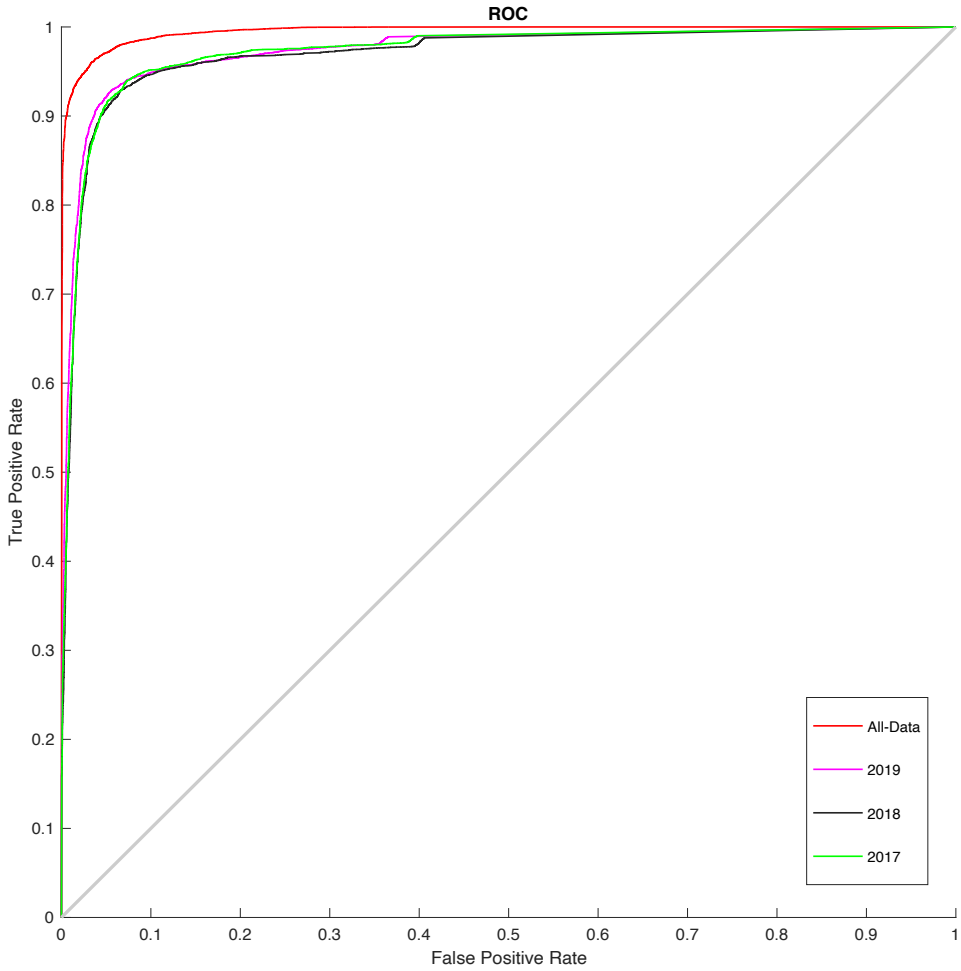

Figure 5.17: ROC curve for Majority Voting combiner

Figure 5.5 shows the reliability diagram for Majority rule combiner. The graph indicates outstanding performance for the All-Data dataset. Additionally, it can be seen that the lines lay very close to the optimal diagonal line, which reflects good reliability of the combiner classifications.

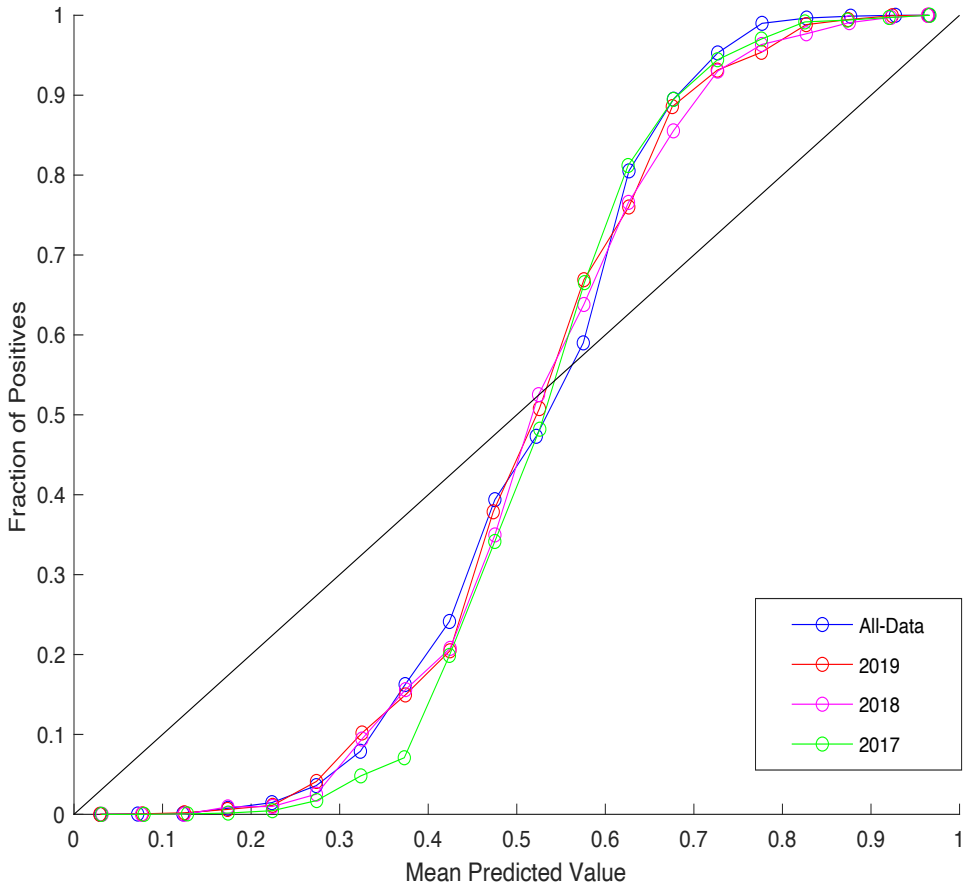

Figure 5.18: Reliability diagram for Majority Voting combiner

5.3.6. Weighted Average Rule

According to Table 5.6, Weighted Average combiner results outperformed all other traditional combiners. The Combiner rule is intended to improve on the Average rule, and it has the highest average accuracy rates for all years' datasets and the All-Data dataset. This is attributable to giving higher weights to single classifiers with higher accuracy rates. The combiner has relatively low classification errors, indicated by Type I and Type II Error. Moreover, the Combiner achieved the highest performance when using all years' data, and a high accuracy rate for All-Data. The Brier score range shows very low classification error.

Table 5.6: Weighted-AVG combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	95.7%	95.1%	94.1%	96.7%
Type I Error	4.5%	4.9%	6.9%	4.1%
Type II Error	4.5%	4.9%	2.9%	2.5%
Sensitivity	95.5%	95.1%	95.1%	97.5%
Specificity	95.9%	95.1%	93.1%	95.9%
AUC	99.51%	99.3%	99.27%	99.78%
Brier Score	0.0221	0.0316	0.0318	0.0205
Area Under Reliability Curve	0.1299	0.1154	0.1218	0.1249

Figure 5.19 shows the ROC curve for weighted average combiner, indicating the high performance of the classifier for all years, with minor gaps between each year's dataset.

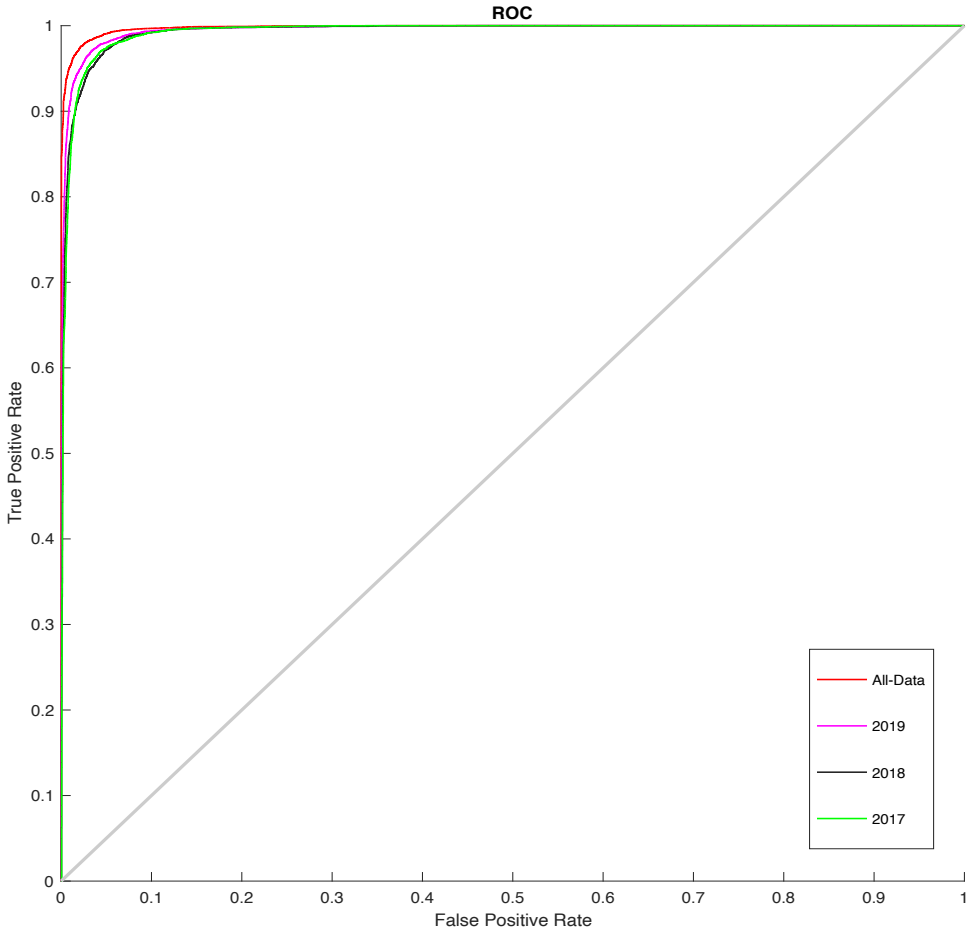

Figure 5.19: ROC curve for Weighted AVG combiner

Based on Figure 5.20, the Weighted Average combiner showed good performance, as all the lines lay very close to the optimal diagonal line.

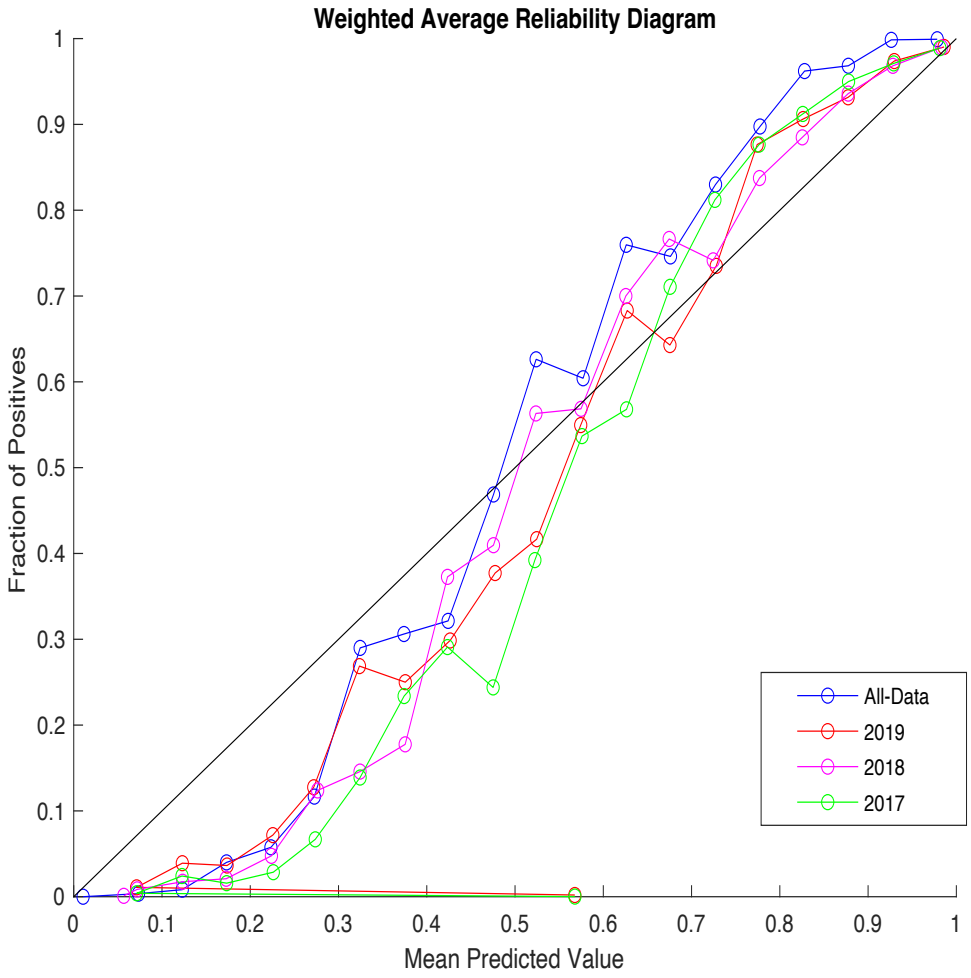

Figure 5.20: Reliability diagram for Weighted AVG combiner

5.3.7. Fuzzy Logic Combiner

Table 5.7 demonstrates the performance results for the Fuzzy Logic combiner. The average accuracy rates achieved by the combiner are lower than all traditional combiners. The combiner has higher Type I Error for all years' datasets. This indicates that the combiner is inefficient to correctly classify failed companies in comparison with other traditional combiners.

Table 5.7: Fuzzy combiner results

	Year 2	**Year 3**	**Year 4**	**All-Data**
Average Accuracy	90.8%	90.2%	89.3%	94.7%
Type I Error	10.6%	11%	12.5%	5.2%
Type II Error	7.7%	8.7%	8.9%	5.4%
Sensitivity	92.3%	91.3%	91.1%	94.6%
Specificity	89.4%	89%	87.5%	94.8%
AUC	97.7%	97.38%	97%	99.09%
Brier Score	0.0604	0.0659	0.0698	0.0369
Area Under Reliability Curve	0.0501	0.0416	0.0477	0.0513

Figure 5.21 show the ROC curve for the Fuzzy Logic combiner, and Figure 5.22 shows its reliability diagram.

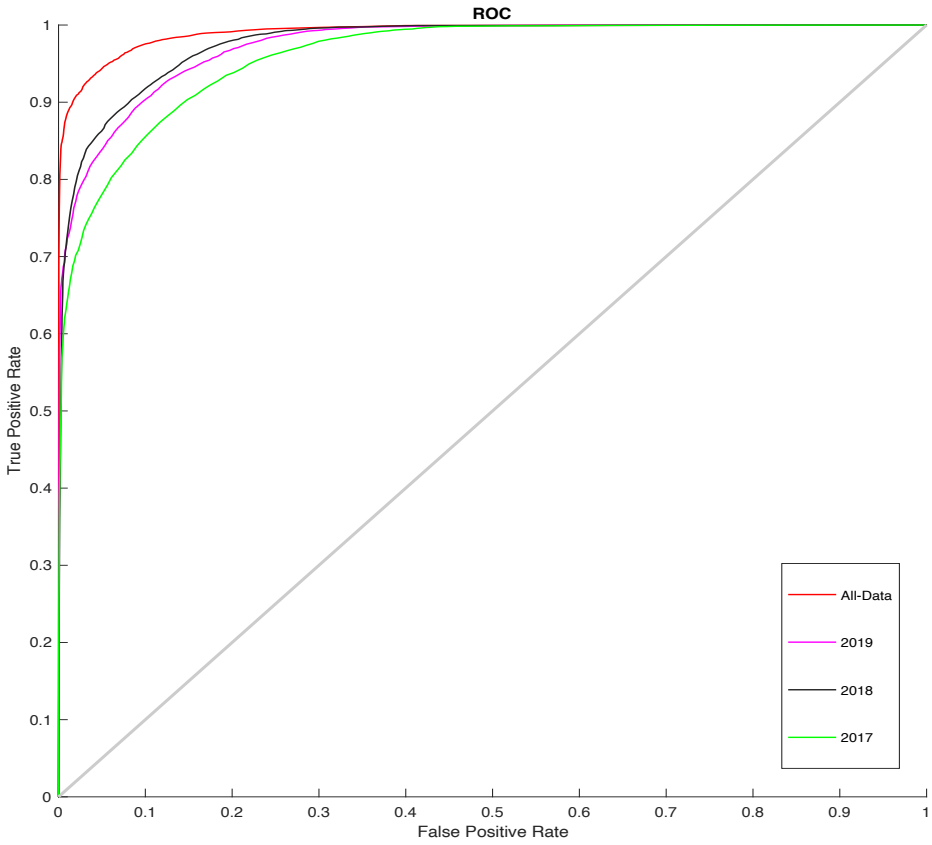

Figure 5.21: ROC curve for Fuzzy combiner

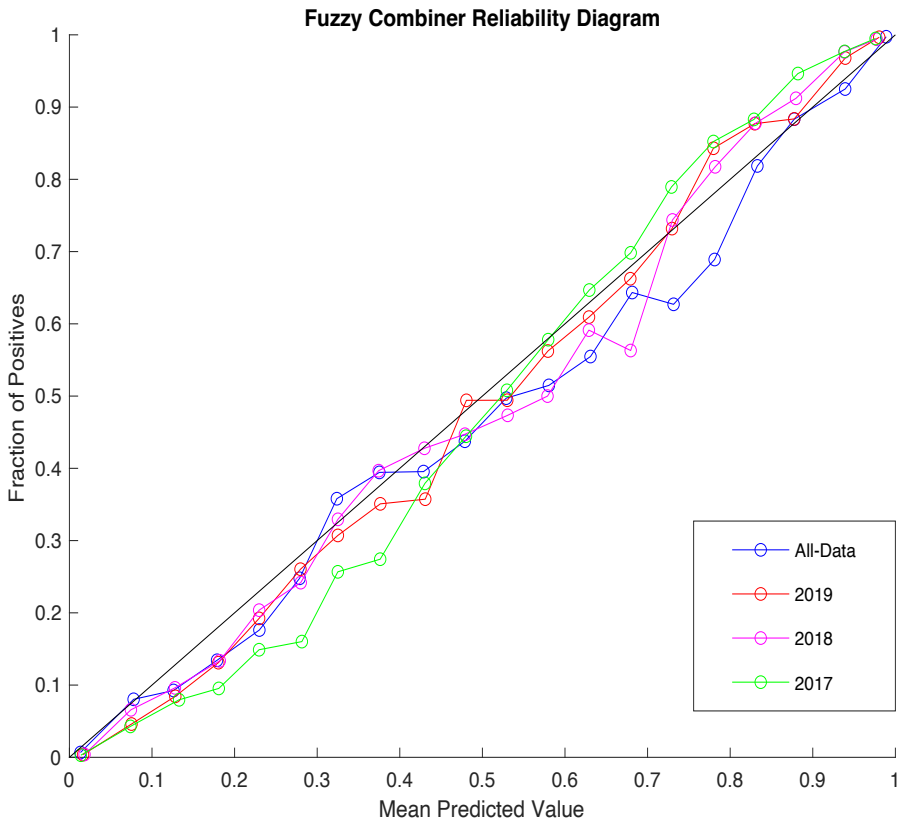

Figure 5.22: Reliability diagram for Fuzzy combiner

5.3.8. Consensus Combiner

According to Table 5.8, the Consensus (Cons) combiner outperformed all other combiners based on all performance measurements. An average accuracy rate of 97.7% was the highest rate achieved using the All-Data dataset. Albeit this rate showed a slight decrease for the yearly datasets, it remained relatively higher level than all other combiners. The combiner shows outstanding performance in classifying failed companies across all data based on the studied performance measurements.

Table 5.8: Cons combiner results

	Year 2019	**Year 2018**	**Year 2017**	**All-Data**
Average Accuracy	96.7%	96.1%	96.1%	97.7%
Type I Error	4%	4.9%	5.5%	2.1%
Type II Error	2.5%	2.9%	2.3%	2.4%
Sensitivity	97.5%	97.1%	97.1%	97.5%
Specificity	96%	95.1%	94.5%	97.9%
AUC	99.51%	99.3%	99.27%	99.78%
Brier Score	0.0221	0.0316	0.0318	0.0205
Area Under Reliability Curve	0.1299	0.1154	0.1218	0.1249

Figure 5.23 shows the ROC curve for the Consensus combiner. Approximately all curves lay close to each other, with very small gaps, reflecting the AUC ratios in Table 5.8.

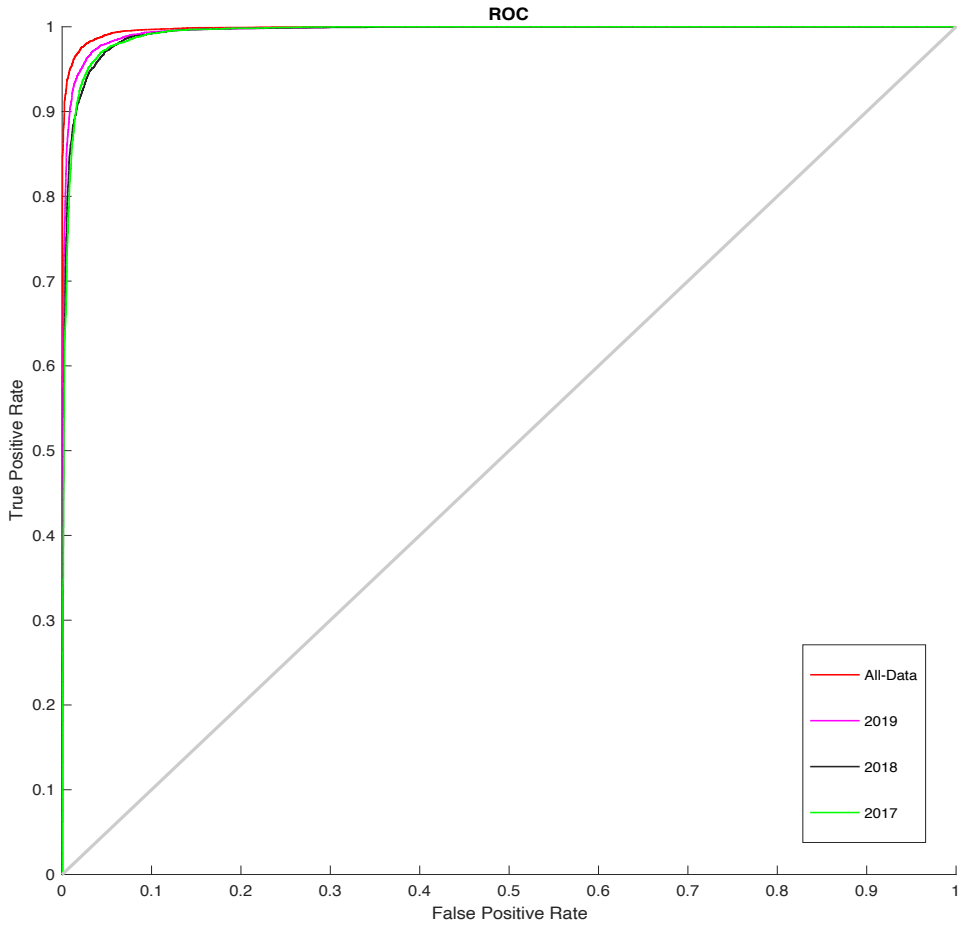

Figure 5.23: ROC curve for Cons combiner

Figure 5.15 shows the reliability diagram for the Consensus combiner. Obviously, the shape of the lines for all years' data indicates the high certainty of the combiner to classify companies correctly.

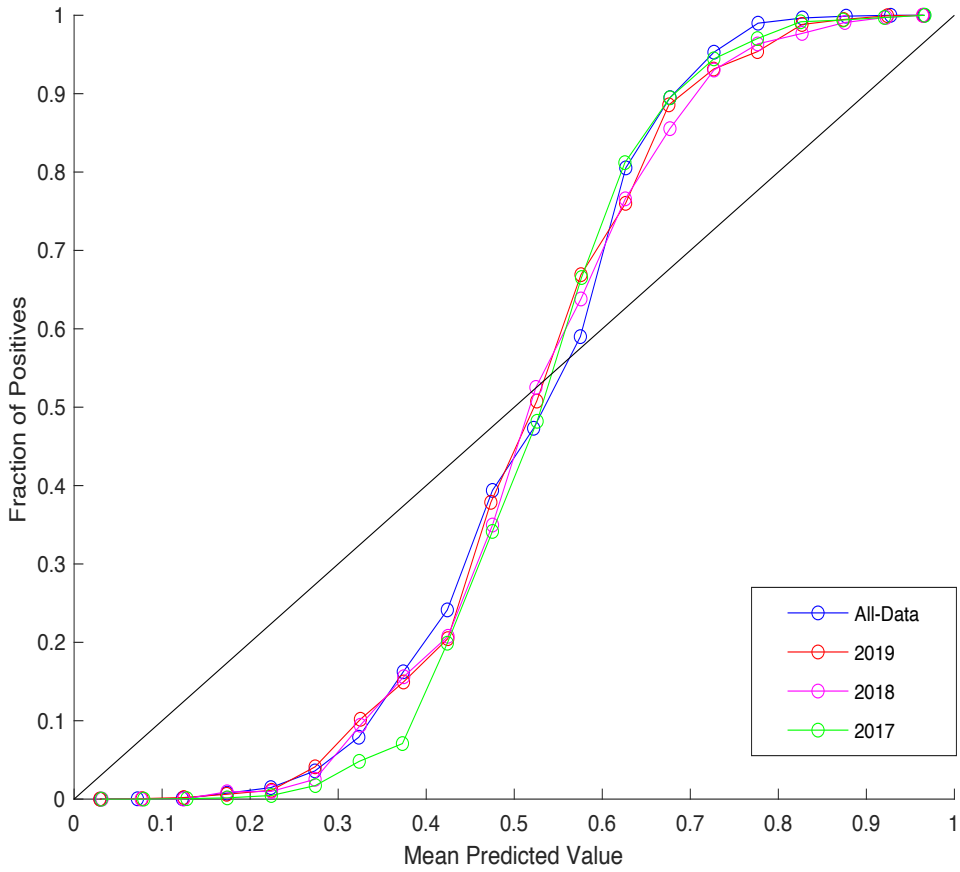

Figure 5.24: Reliability diagram for Cons combiner

5.4. Discussion and Analysis

In this section all traditional combiners' results are analysed along with consensus and fuzzy logic combiners and are compared with single classifiers and LR. The discussion reveals which combiner had the best improvement over single classifier classification and achieved the best performance. Tables 5.9 to 5.12 show all combiners' results for the All-Data dataset, and the datasets for the years 2019, 2018, and 2017 (respectively). It can be seen that the best combiner appears to be Consensus method, which outperformed all traditional and fuzzy logic combiners in terms of average accuracy. Its results show relatively higher stability than other combiners across all years' datasets. This can be explained by the ability of the combiner to allow more contribution of the classifiers which have low uncertainty values to the final classifications.

This was followed by the Weighted Average, which works according to a similar structure to that of the Consensus classifier, but without taking uncertainty values into consideration. Obviously, assigning weights to classifiers when using combining methods has shown an outstanding result in enhancing single classifiers results. Moreover, other traditional combiners (Median, Average, Max, Min, and Majority voting) also show an enhancement on single

classifier performance, with relatively close average accuracy to those achieved by the Consensus and the Weighted Average combiner. However, the Fuzzy Logic combiner achieved the worst average accuracy, with less improvement over single classifiers, in comparison with all other combiners. Despite taking the confidence level and the standard deviation of all single classifiers into consideration when calculating all fuzzy sets, the defuzzification process contributes critically to the final answers, as it uses Centre of Gravity as the defuzzification method, by which results are too sensitive to the threshold.

In terms of accuracy, Consensus again ranked first, followed by the Weighted Average, and the other classifiers had the same ranking as above. The initial accuracy rate was higher for 2019 than for 2018 and 2017, but the differences varied among combiners. The Consensus and Weighted Average classifiers had lower gaps between average accuracy rates than the other combiners and showed more stability over the years. This illustrates the robustness of these combiners to correctly classify failed companies using their financial performance information.

For the All-Data dataset, all combiners achieved relatively good performance in terms of classifying failed companies, reflected in their specificity and Type I Error. The ranking of the combiners' performance based on these parameters was different than for average accuracy. However, the Consensus Combiner still has the highest performance rates among all combiners and ranked first. It shows more capability to classify failed companies than active companies, based on its higher specificity than sensitivity rate. Surprisingly, Max combiner ranked second as in terms of classifying failed companies; it was enhanced by adjusting the threshold, which reflected inversely on its sensitivity rate and Type II Error. This combiner was followed in ranking by Majority Voting, Median, Average, and Weighted Average, respectively. Fuzzy logic was considered to be the worst combiner in terms of classifying failed companies.

Cons had the highest specificity and sensitivity rates among all of the yearly datasets. Based on Brier scores parameter, Consensus, Weighted Average, and Majority voting achieved the lowest scores among all years' datasets, and Fuzzy Logic and Max had the highest.

In conclusion, Consensus Combiner achieved outstanding performance, combining all single classifiers' predictions based on all measurements, and it is more accurate than traditional combiners. However, in comparison with LR, all combiners achieved better classification results. This justifies the usefulness of these combiners to classify companies' status based on their financial performance.

Table 5.9: All combiners' All-Data results

	All-Data Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier score	AURC
Fuzzy Combiner	94.7%	5.4%	5.2%	94.6%	94.8%	99.09%	0.0369	5.13%
Minimum	95.5%	3.5%	6.4%	96.5%	94.4%	99.19%	0.043	13.62%
Maximum	96%	4.9%	3.1%	95.1%	96.9%	99.34%	0.0411	13.78%
Average	96.2%	3.5%	4%	95.4%	96%	99.5%	0.0451	20.1%
Maj_Vote	96.3%	4.3%	3.2%	95.7%	96.8%	99.48%	0.0342	16.81%
Median	96.5%	3.7%	3.3%	96.3%	96.7%	99.53%	.0342	17.34%
Weighted_Avg	96.7%	2.5%	4.1%	97.5%	95.9%	99.78%	0.0351	12.1%
Cons	97.7%	2.1%	2.4%	97.5%	97.9%	99.78%	0.0205	12.49%

Table 5.10: All combiners' Year 2019 results

	2019 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier score	AURC
Fuzzy Combiner	90.8%	10.6%	7.7%	92.3%	89.4%	97.7%	0.0604	5.01%
Minimum	94.1%	4.4%	7.3%	92.7%	95.6%	97.51%	0.0724	16.99%
Maximum	94.4%	7.3%	3.8%	96.2%	92.7%	98.54%	0.0664	14.97%
Average	94.5%	6.5%	4.6%	93.4%	93.5%	98.7%	0.0646	19.82%
Maj_Vote	96.3%	3.2%	4.3%	95.1%	96.8%	99.48%	0.0342	16.81%
Median	92.6%	8.1%	6.7%	93.3%	91.9%	98.15%	0.0619	14.58%
Weighted_Avg	95.7%	4.5%	4.5%	95.9%	95.5%	98.7%	0.0346	12.99%
Cons	96.7%	4%	2.5%	97.5%	96%	99.51%	0.0221	12.99%

Table 5.11: All combiners' Year 2018 results

	2018 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier score	AURC
Fuzzy Combiner	90.2%	11%	8.7%	91.3%	89%	97.38%	0.0659	4.16%
Minimum	93%	5.5%	9.5%	91.5%	94.5%	97.29%	0.0781	15.34%
Maximum	93%	5.5%	9.5%	91.5%	94.5%	97.29%	0.0781	15.34%
Average	94.1%	7.4%	4.3%	95.75	92.6%	98.55%	0.0662	19.56%
Maj_Vote	93.8%	7.5%	4.9%	95.1%	92.5%	97.94%	0.487	16.82%
Median	92.5%	8.8%	6.7%	93.7%	91.2%	98.07%	0.0623	13.995%
Weighted_Avg	95.1%	5%	4.5%	95%	95.1%	98.55%	0.0362	11.54%
Cons	96.1%	4.9%	2.9%	97.1%	95.1%	99.3%	0.0361	12.99%

Table 5.12: All combiners' Year 2017 results

	2017 Dataset							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier score	AURC
Fuzzy Combiner	89.3%	12.5%	8.9%	91.1%	87.5%	97%	0.0698	4.77%
Minimum	93.8%	8.4%	4.1%	95.9%	91.6%	98.17%	0.0736	13.37%
Maximum	92.7%	6.6%	8%	92%	93.4%	97.21%	0.0837	16.42%
Average	94%	8.1%	3.8%	96.2%	91.9%	98.51%	0.0770	22.1%
Maj_Vote	92.6%	9.4%	5.4%	94.6%	90.6%	97.21%	0.0576	12.09%
Median	91.5%	10.1%	6.8%	93.2%	89.9%	97.72%	0.0743	18.52%
Weighted_Avg	94.1%	6.9%	2.9%	95.1%	93.1%	98.51%	0.0370	12.1%
Cons	96.1%	5.5%	2.3%	97.7%	94.5%	99.27%	0.0318	12.18%

5.5. Statistical Significance Testing

This section shows Friedman statistical test results for all implemented models, to prove that Cons classifier is the best classification method not only for all UK year dataset in this study; there is a high probability that this applied for all datasets with a similar structure to those used in this study. Bonferroni-Dunn test is conducted to rank all classification methods from best to worst, and they are divided based on a critical value at a certain alpha level into two groups: (1) a group of classifiers the best classifiers (Cons and its potential rivals – WAVG, Median, DPL, ENS-DT, and DT); and (2) a group of classifiers that are definitely worse than Cons. The former (n = 6) were used for comparative purposes.

The Friedman test performed over all datasets for all classifiers and for the six best-ranking classifiers, accompanied with a pairwise comparison for these classifiers. The reason behind using first six ranked classifiers is that including all classifier results in the test would be render it very complex without contributing to the demonstration of whether the Cons method is the best classifier. Table 5.13 demonstrates the results of Friedman test of these classifiers over all datasets.

Table 5.13: Friedman test – all classifiers and best six

Dataset	All-Data	2019	2018	2017
Friedman χ^2 (all classifiers)	49592.44	7308.05	17184.974	17320.741
Friedman χ^2 (best six classifiers)	1303.306	255.175	690.397	645.058

According to the statistical testing explained in section 3.6, a null-hypothesis means that there is no difference between the 6 classifiers ranking. Based on critical value from Chi-Square distribution table with *C*-1 degrees of freedom, A null-hypothesis is accepted with significance levels of:

- 0.05 if the Friedman test statistic F < $(x^2_{0.05}(5) = 11.07)$.
- 0.10 if the Friedman test statistic F < $(x^2_{0.10}(5) = 9.24)$.

After testing the hypothesis on 0.10 and 0.05 significance levels, pairwise comparison results for the six best classifiers are demonstrated in Tables 5.14 to 5.17 across all datasets.

Table 5.14: Friedman test – comparison of best six classifiers (All-Data)

Friedman χ^2 =	**Accuracy**	**WAVG**	**Median**	**DPL**	**ENS-DT**	**DT**
Cons	97.7%	0	0.065	0	0	0
WAVG	96.7%	-	0	0	0	0
Median	96.5%	-	-	0	0.002	0
DPL	97.2%	-	-	-	0	0.011
ENS-DT	96.2%	-		-	-	0.008
DT	95.3%	-	-	-	-	-

Table 5.15: Friedman test – comparison of best six classifiers (2019)

Friedman χ^2 =	**Accuracy**	**WAVG**	**Median**	**DPL**	**ENS-DT**	**DT**
Cons	96.7%	0	0.479	0.015	0.015	0
WAVG	95.7%	-	0	0	0	0
Median	92.6%	-	-	0.083	0.002	0
DPL	96.3%	-	-	-	0	0
ENS-DT	94.8%	-	-	-	-	0
DT	94.2%	-	-	-	-	-

Table 5.16: Friedman test – comparison of best six classifiers (2018)

Friedman χ^2 =	**Accuracy**	**WAVG**	**Median**	**DPL**	**ENS-DT**	**DT**
Cons	96.1%	0	0.391	.0865	0	0
WAVG	95.1%	-	0	0	0	0
Median	92.5%	-	-	0.304	0	0
DPL	95.3%	-	-	-	0	0
ENS-DT	94.4%	-	-	-	-	0
DT	93.6%	-	-	-	-	-

Table 5.17: Friedman test – comparison of best six classifiers (2017)

Friedman χ^2 =	**Accuracy**	**WAVG**	**Median**	**DPL**	**ENS-DT**	**DT**
Cons	96.1%	0	0.12	0	0	0
WAVG	94.1%	-	0	0	0	0
Median	91.5%	-	-	0	0	0
DPL	95.3%	-	-	-	0	0
ENS-DT	94%	-	-	-	-	0.094
DT	93%	-	-	-	-	-

Based on the results of the Friedman significance test for the six best classifiers, the null-hypothesis is rejected for all datasets at both significance levels (0.05 and 0.10). Pairwise t-test results show if a pair of classifiers has performed in the similar way. The obtained data shows low p-value for all pairs of the six classifiers across all datasets, and hence the performance of each classifier is distinct and proportional to its accuracy.

To evaluate the ranking of all classifiers, Bonferroni-Dunn two tailed test was conducted based on the significance levels 0.05 and 0.10 using the following equation:

$$CD = q_a \sqrt{\frac{k(k+1)}{6N}} \tag{5.21}$$

where k =17 represents number of classifiers, N = 4 represents the number of datasets, q_a represents the studentised statistic calculated based on confidence level *a / (k-1)* which in this case is a / 17, divided by √2. The obtained results gave values of $q_{0.05}$ = 3.5036 and $q_{0.10}$ = 3.1747, resulting in $CD_{0.05}$ =12.510 and $CD_{0.10}$ = 11.335.

These calculated values at each significance level, plus the lowest rank, are represented in the two horizontal lines in Figure 5.25, which shows the height that represents the threshold for the best classifiers. The results clearly show that Cons classifier is better than all of the individual classifiers and traditional combiners. DPL classifier holds the second rank for all datasets, with good and stable results. Although Fuzzy, MIN, and MAX combiners show good performance, they are ranked as the worst combining methods. However, based on the evaluated critical value at the significant levels a = 0.05 and a = 0.10, it can be concluded that

NB, LR, KNN, and LD are significantly worse than Cons method classifier, and SVM is worse only at the significant level a = 0.10.

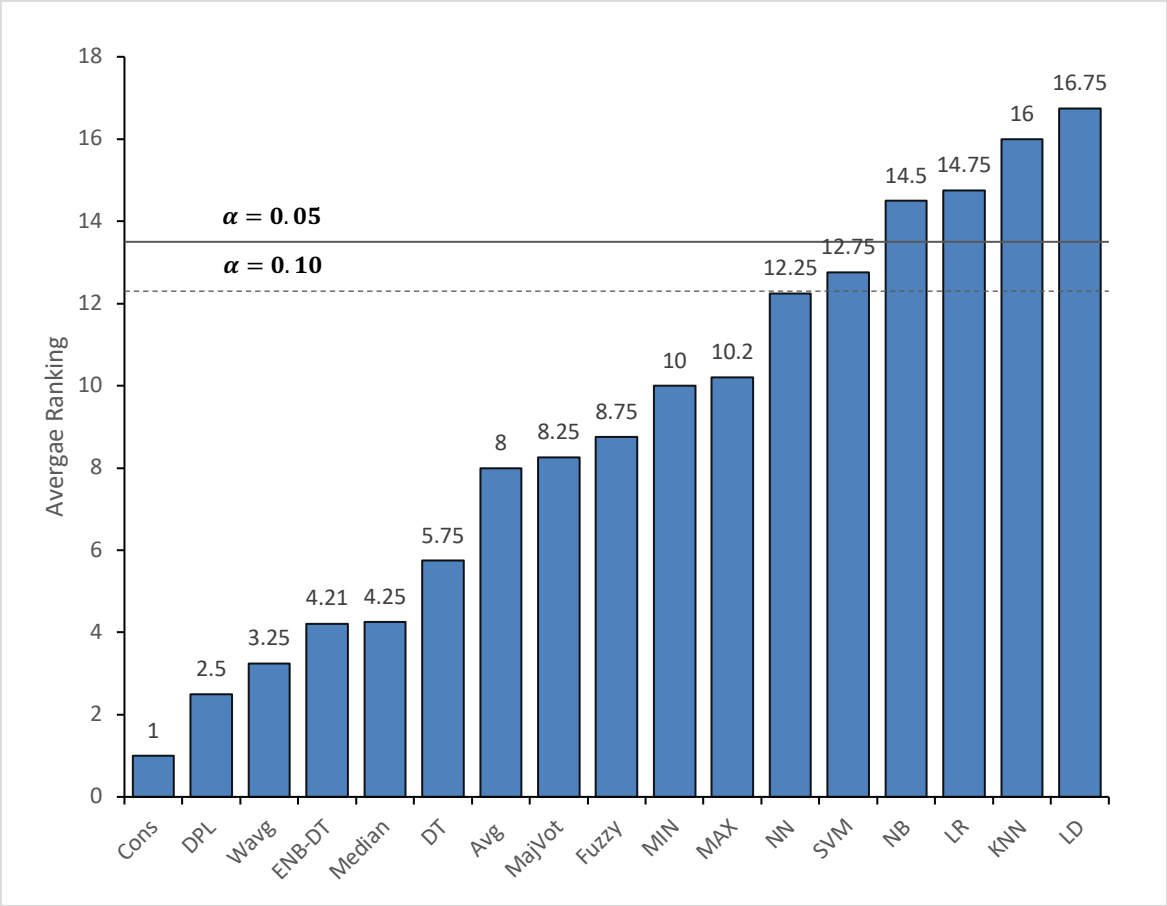

Figure 5.25: Significance Ranking for the Bonferroni-Dunn two-tailed test with a=0.05 and a=0.10

5.6. Classification Model Computational Time

According to Tables 5.18 and 5.19, the computational training time required to train all the models are shown. As can be seen from the tables, the computational time are displayed in seconds. For the individual classifiers, DPL classifier required a longer tunning time in comparison with all single classifiers and the other combiner method used in the study. This can be justified by its greater construction complexity and its ability to ignore insignificant attributes on the data. However, LDA and DT classifiers have completed the training process more quickly than other individual classifiers, as it completed the process in 17.41, and 12.17 seconds for All-Data dataset. Other individual classifiers took more time to process the datasets.

As the final stage is to evaluate the computational time for the combining methods used, it is worth mentioning that its computational time are computed after computing all base individual classifiers. As seen in Table 5.5, traditional combining methods took less training time than other individual classifiers,

due its simplicity of calculation. The Cons model relatively took more training time than other combiner used due to its training complexity. However, it was selected the best classification method among both individual and combining classifiers, as it can expediently offer new business failure classification in a relatively short time.

Table 5.18 Training time for single models in seconds

Dataset	Year 2019	Year 2018	Year 2017	All-Data
LR	12.313	12.942	12.062	42.46
LDA	7.792	7.2802	7.3601	17.41
ANN	20.1	23.49	12.4	82.92
K-NN	595.99	582.24	593.34	859.5
NB	345.11	405.14	385.94	566.9
SVM	2397.3	1938.32	2265.41	3426.5
DT	8.0062	9.61	9.6362	12.17
BEC	647.371	589.181	635.86	987.7
DPL	3694	3964	3793.35	5800

Table 5.19 Training time for committee combiner model in seconds

Dataset	Year 2019	Year 2018	Year 2017	All-Data
MIN	7.71	7.6	7.73	10.34
MAX	7.35	7.64	7.57	10.42
MED	6.1	6.4	6.7	8.1
MajVot	5.86	5.93	6.27	9.1
AVG	7.3	7.1	7.39	8.46
WAVG	75.5	75.3	76.8	121.71
FC	133	134.5	127.89	280.32
CON	640.6	642.8	619.84	859.1

5.7. Summary

Based on the results of this chapter, it has been proven that combining single classifiers predictions can produce more accurate classification of firms' status. Clearly the Consensus combiner is the best method, due to its computational capabilities, which take into consideration single classifiers' uncertainty levels about their answers.

The combiners used in this chapter could be considered as robust methods, producing more stable results for companies across all years' datasets. An advantage of using traditional

combiners is their simple mathematical computation structures, despite the need to adjust thresholds for some combiners (Min and Max). Compared with single classifiers, combination methods achieved more stable results and provided enhancement over individual classifier performance.

It is sufficient to use combiner methods to enhance companies' status classification based on their overall performance on all measurements. The next chapter focuses on predicting companies' status one step ahead, using dynamic time series classifiers.

Chapter 6
Development of Insolvency Prediction Model

6.1. Introduction

Chapters 4 and 5 demonstrated that committee machine combiners have enhanced individual model classification performance. The combination method using Cons model achieved the best classification result for business failure for the studied UK datasets. However, these classifiers, both individual and combiners, used static dataset to classify business activity, which can be called a static classifier. This chapter proposes a new modelling approach using dynamic classifiers to predict business failure one step ahead (before it happens). The modelling techniques adopted in this chapter are DPL-SA, NARX, and NAR. 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 aim is to predict business failure one year before it happens using time series data for UK firms for the year 2019. The dataset used to develop the models consists of the 2019 dataset with an additional four datasets for the four consecutive years before business failure. The model is considered an early warning classifier that can provide users with early information about on-going businesses. The model was trained and tested on the datasets to validate its prediction performance. After modelling and testing, each of the prediction classifier results was compared with the results achieved by the best static individual classifier in this study (DPL). Finally, significance testing is presented to evaluate the best model.

The next section illustrates the dynamic modelling development of each prediction classifier, followed by the experimental results and then the discussion of the outcomes.

6.2. Data Clustering

Data clustering is defined as a grouping process in which a large number of data in a dataset are divided into a certain number of groups, in order to make data points in the same groups more similar than those in other groups (Hammouda and Karray, 2000). In simple words, the aim is to put data with similar traits in a single group and assign them into clusters. The clustering method used in this study is the Fuzzy C-means, based on its fast implementation and credibility (Suganya and Shanthi, 2012). To fit the large number of data into the dynamic modelling techniques, each one of the two classes of the dataset were clustered into 200

representative input data clusters using the Fuzzy C-means method. The result is a dataset consisting of 400 input clusters.

6.3. Nonlinear Autoregressive with Exogenous Input (NARX)

Recurrent neural network is widely used in different fields as a nonlinear dynamic approach to predict the next step in time series data. It is a useful tool for modelling the input dynamical order, number of neurons, and number of delays. A suitable training and testing algorithm has to be selected to enhance the performance of the prediction, and the final (output) layer consists of the predicted target output, which in this case companies' status y(t). Time series data outputs are fed to the model to predict the next output values of the data. In this case, it is assumed to be the first time this method has been used to predict companies' status one year before failure in order to investigate early warning potential of models for companies' health.

NARX is a dynamic recurrent neural network used to predict next value of the target output by regression of the latest values of the dependent variable n_y over the latest value of the independent variables n_x. These values represent the dynamic order of both the dependent and independent variables in the model, where n_y is the companies' status for four years and n_x are the financial variables related to the same companies. The following nonlinear mathematical function is used for NARX modelling.

$$y_{t+1} = f\left(y_t, y_{(t-1)}, \dots, y_{(t-3)}, x_t, x_{(t-1)}, \dots, x_{(t-3)}\right)$$

Similar to neural network topography, the NARX network consists of three main layers, the first of which is the input layer, followed by the hidden layer and the final output layer. The input layer includes the input data (financial variables) and the current output (companies' status). These inputs are fed into the hidden layer, which consists of a certain number of neurons and number of delays, in order to map input data using nonlinear function, which assigns weights and biases to these data in an open loop network. Figure 6.1 displays the NARX view command function.

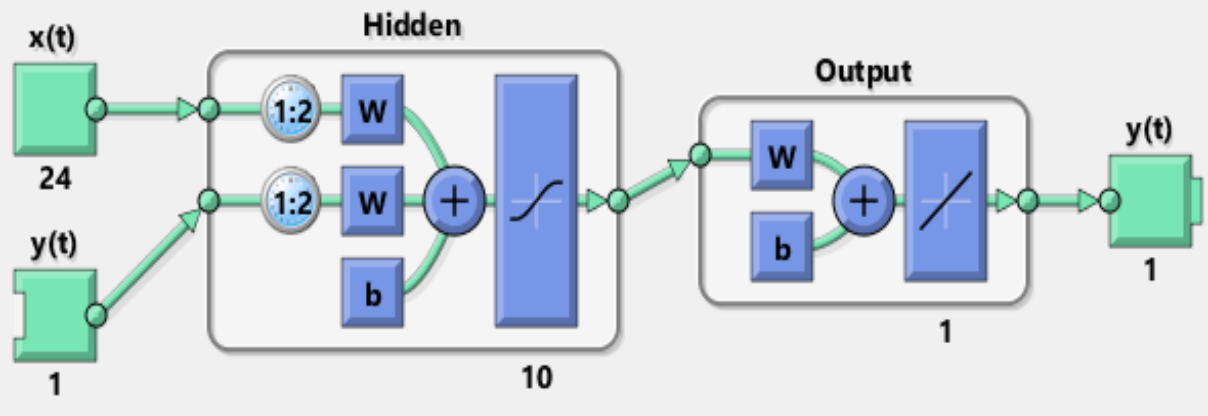

Figure 6.1: NARX view command

After training the network in an open loop based on training, testing, and validating data segmentations, the network is converted into a closed loop one. Here the output of the open loop network is fed back again to the network to produce multi-step ahead predictions of the data. In a closed loop network, the function CLOSELOOP replaces the feedback input with a direct connection from the output layer, as shown in Figure 6.2. To perform the multi-step prediction, it is crucial to simulate a network in open-loop form as long as there is known output data, and then switch to closed-loop form while providing only the external inputs. Next, the network and its final states are converted to closed loop form, to make multi-step predictions with only the inputs provided.

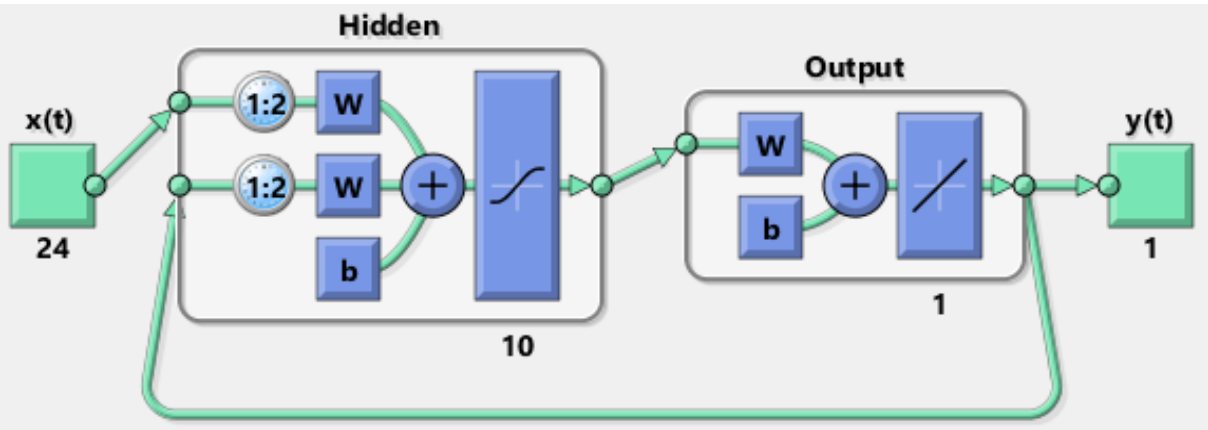

Figure 6.2: NARX close loop

Finally, to generate one step ahead predictions of companies' status as an early alert of its next-year situation, the network returns the predicted *y(t+1)* at the same time it is given *y(t+1)*. It is always beneficial for decision making when early predictions about companies' health are

available. As shown in Figure 6.3, the network can be made to return its output a timestep early, and provide important information by removing one delay (to be zero) where outputs are shifted one more timestep.

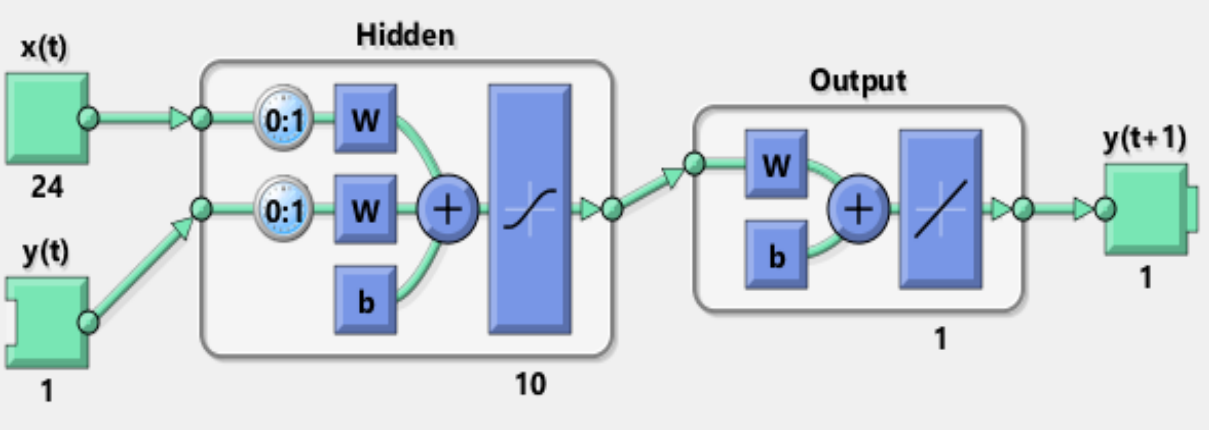

Figure 6.3: NARX predict one step ahead

A NARX network can thus provide early information or prediction about companies' status when provided with data on its previous financial performance, which is considered a substantial contribution in the decision-making process for users.

6.4. Nonlinear Autoregressive Neural Network (NAR)

For most cases, time series applications are characterised by variation in transient periods. Therefore, it is difficult to use linear models for time series predictions, and nonlinear approach are more suitable in such cases. A NAR can be applied to predict one step ahead in time series forecasting using nonlinear autoregressive model as follows:

$$y(t) = f(y(t-1), y(t-2), I, y(t-p)) + \varepsilon(t) \qquad (6.1)$$

This formula shows how NAR network uses the *p* past values of *y* to predict the next step ahead value of *y* at time *t*, *y(t)*, where y represents the financial variables for each company for a time series of four years. The function *f(.)* is set through the training of the neural network, determined by means of the optimisation of the network weights and neuron bias. *ε(t)* is the error of the approximation of the values of the financial variables at time *t+1*.

To solve Autoregression time-series prediction using NAR network, there are substantial steps that the network should go through. First, the variable is defined as the 24 financial ratios of

companies for four years. The second step is choosing the training function from a list of function, *trainlm*, *trainbr*, and *trainscg*. To build the network, *trainlm* (Levenberg-Marquardt backpropagation) is selected, as it is the most commonly used training rule for the NAR network, and it is the fastest learning function to train a backpropagation model. The network is created with feedback 1:2, 10 hidden layers, and an open-loop network form. It is crucial to use the optimal number of neurons, since increasing the number to a high level makes the network more complex, while lower numbers may restrict the computing power of the system and make it less generalised.

The next step is to prepare the data for training and simulation using the function PREPARETS, which minimises time shifts to fill input states and layer states. This maintains the original time series data unchanged, while customising it for networks with different numbers of delays in both open and closed loop feedback modes.

Figure 6.4 shows the topology of the view command. To train and test the network, the data is divided into training, and testing sets, with the ratios 80%, and 20%, respectively. The performance function is selected as mean squared error.

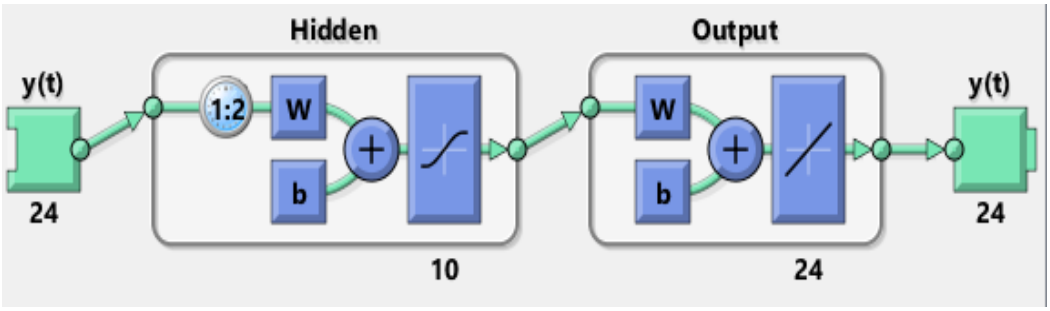

Figure 6.4: NAR view command

After training the network in an open-loop form, the next step is to convert the network to a closed-loop form, to do multi-step predictions, as shown in Figure 6.5. Here the function closed loop replaces the feedback input with a direct connection from the output layer. In multi-step prediction, the network is first simulated using the open-loop form, as long as there is known data for the variables, then the network switches to closed-loop to perform multistep predictions.

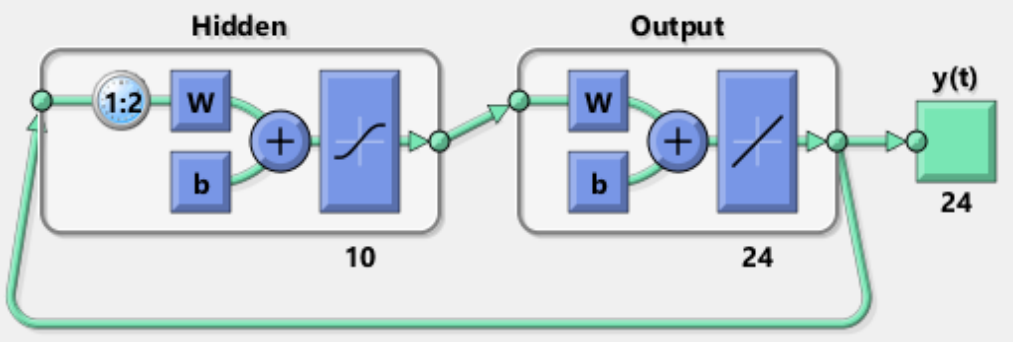

Figure 6.5: NAR close loop

As the aim of the algorithm is to get the predictions of companies' status a timestep early, whereby the original network returns predicted *y(t+1)* at the same time it is given *y(t+1)*. It would be helpful to predict *y(t+1)* once *y(t)* is available, but before the actual *y(t+1)* occurs. According to Figure 6.6, the network can be made to return its output a timestep early by removing one delay, so that its minimal tap delay is now 0 instead of 1. The network returns the same outputs as the original network, but outputs are shifted left one timestep, so the output predicts companies' variables one year ahead.

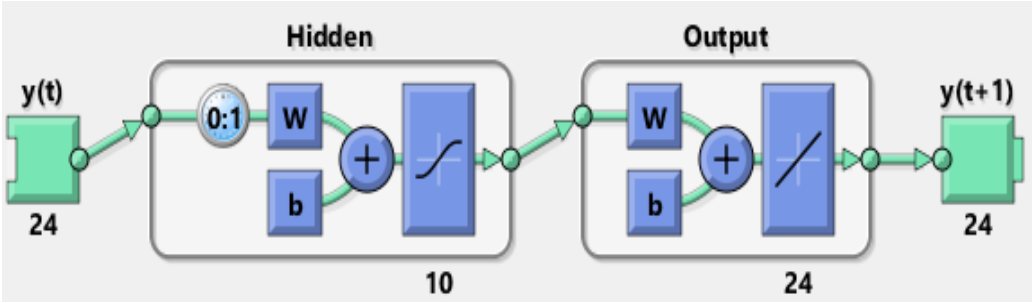

Figure 6.6: NAR one step ahead prediction

Unlike NARX network, where the final output is the prediction of companies' status, the output predictions of the NAR network is all financial variables values one year ahead. As a final step, these new variables are used as an input in a DPL single classifier model to measure how these predictions can be informative about companies' classification. The aim is to compare the performance of these new variables with actual financial variables achieved by the companies in the dataset for year five.

6.5. Deep Learning Time Series

A robust forecasting one step ahead algorithm can be built using deep learning to solve binary classification using multiple layers that progressively extract information from the raw input

data. Deep learning methods are commonly used to train data using supervised learning when providing input data to predict and forecast binary classification. The architecture of the model consists of building a layer-by-layer model. The model is built using LSTM, in which the core components of the network are a sequence input layer. For network creation, a layer containing a sequence input layer is implemented, followed by an LSTM layer fully connected to a Softmax layer, linked to a classification output layer. SofMax activation function is selected because of its ability to handle multiple classes and its usefulness for output neurons.

The input size is set as 24, representing the number of features of the input data used to feed the sequence input layer in the network. The number of hidden units is set as 24, and the number of classes is set as 2, representing the two classes of the target output. The maximum epochs are set as 200, and the minimum batch size is 100. After creating the optimal model structure, the model is trained and tested using a training and testing dataset extracted from the original dataset, with a percentage of 80% and 20%, respectively. The Predict and Update State functions are used to predict the next time step using the observed value of the previous time step. The forecasted output is then compared with year five actual companies' status in order to measure forecasting accuracy. The DL time series framework is displayed in Figure 6.7.

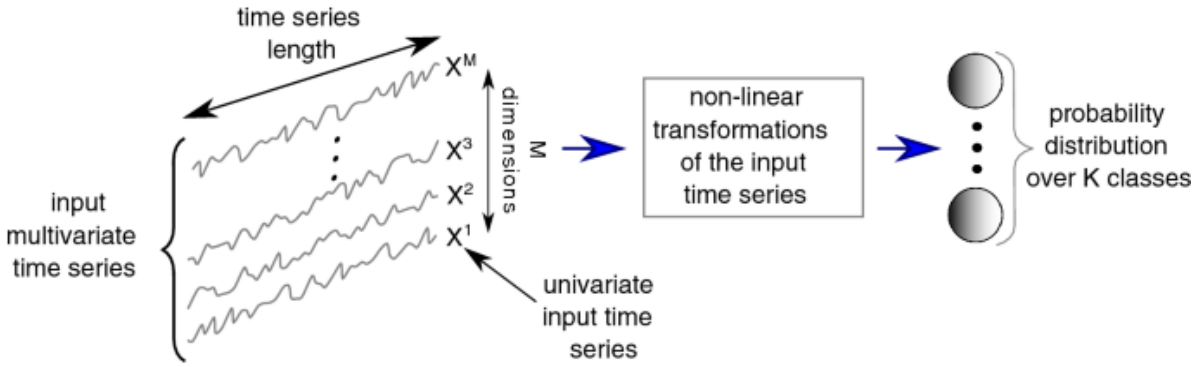

Figure 6.7: DL time series framework

Source: Fawaz *et al.* (2019)

6.6. Experimental Results

6.6.1. NARX Results

Table 6.1 demonstrates NARX results in comparison with Cons single classifier results for the year 2019 based on all performance measures. NARX predictions, for one-year step ahead, achieved 83.6% average accuracy, while DPL single classifier achieved 96.3%. Type I and Type II Error are significantly higher than for Cons, but with a lower gap between the two measurements. The specificity rate for NARX is 82.5%, which is approximately 13% lower

than DPL. The AUC measurement for NARX shows a relatively lower value in comparison with Cons, as this measurement has a positive relationship with average accuracy. Brier score is higher, indicating higher error in predicting companies' status using previous data rather than the same year dataset.

Table 6.1: NARX results

	NARX	**DPL**
Average Accuracy	83.6%	96.3%
Type I Error	17.5%	2.8%
Type II Error	15.3%	4.5%
Sensitivity	84.7%	97.2%
Specificity	82.5%	95.5%
AUC	87.85%	99.35%
Brier Score	0.0728	0.0282
Area Under Reliability Curve	0.1427	0.0405

Figure 6.8 shows the ROC curve for NARX and compares it with the Cons combiner curve. The NARX curve indicates worse performance than the single classifier for the same year dataset. The curve shifts down, resulting in a lower AUC value than Cons.

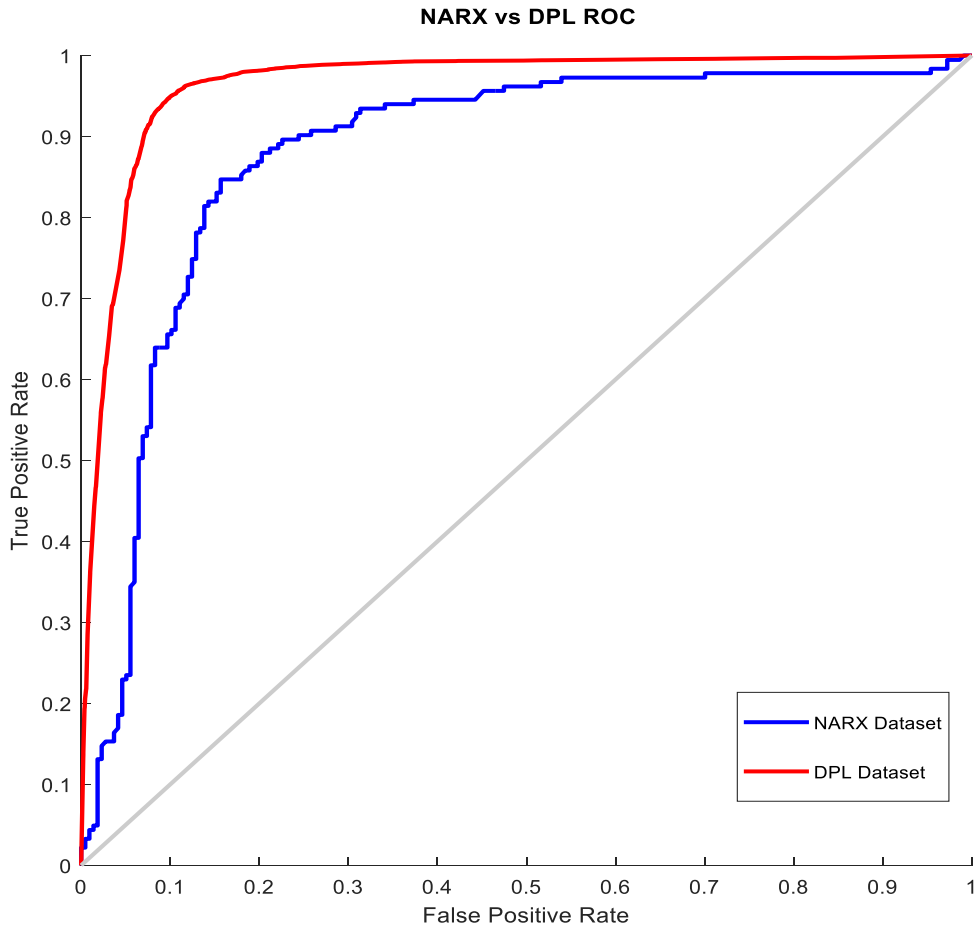

Figure 6.8: ROC curve for NARX classifier

Figure 6.9 shows the reliability diagram for the NARX compared with Cons combiner. The shape of the diagram indicates that NARX is less reliable, due to its fluctuating line.

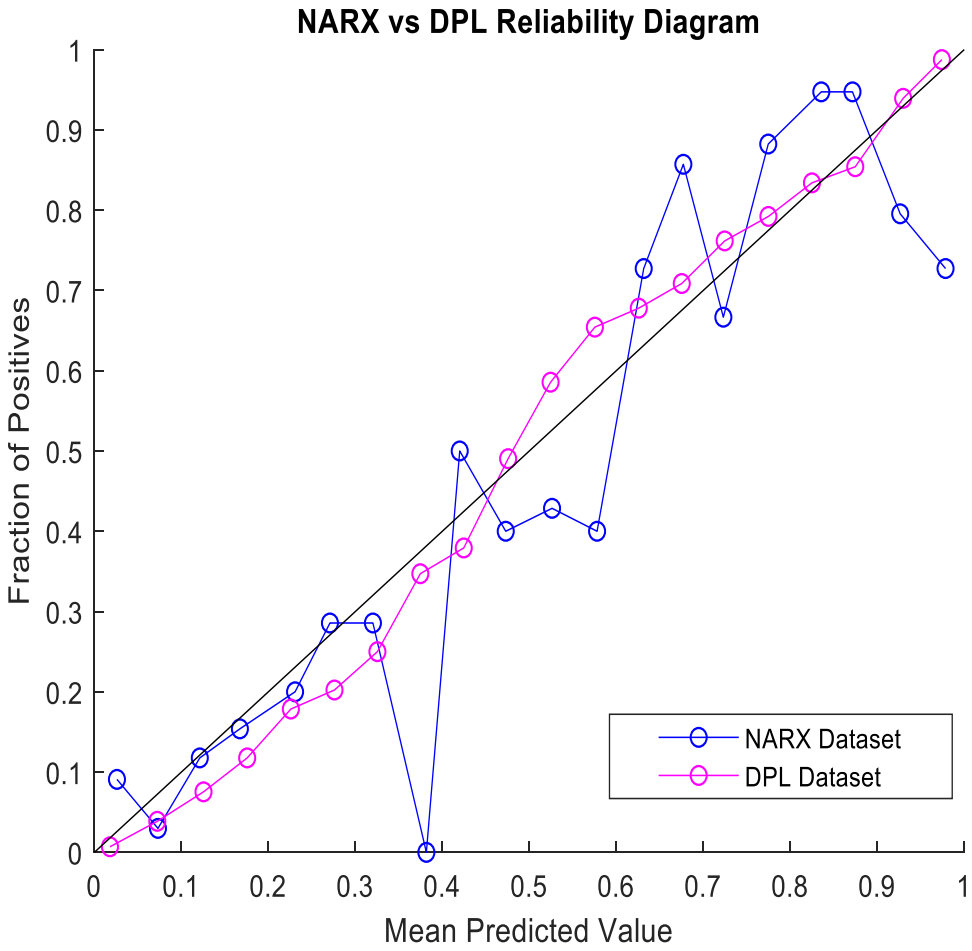

Figure 6.9: Reliability diagram for NARX classifier

6.6.2. NAR Results

The one step ahead results shown in Table 6.2 indicate that NAR shows a substantial improvement for all performance measurements in comparison with NARX classifier. Average accuracy increased to 89.15%, reducing the gap between the actual year data classifier DPL. Sensitivity and specificity rate were enhanced to 91% and 87.3%, respectively. Both Tyoe I and Type II Error decreased, showing more capability of the classifier to predict companies' status more accurately than NARX, albeit it is still outperformed by the Cons combiner. Brier score results are better than NARX, with a value of 0.0628, and are higher than the Cons result. The AUC value increased to 92.48%, showing an improvement in classifier prediction performance over the NARX.

Table 6.2: NAR results

	NAR	**DPL**
Average Accuracy	89.15%	96.3%
Type I Error	12.7%	2.8%
Type II Error	9%	4.5%
Sensitivity	91%	97.2%
Specificity	87.3%	95.5%
AUC	92.48%	99.35%
Brier Score	0.0628	0.0282
Area Under Reliability Curve	0.2144	0.0405

Figure 6.10 shows the ROC curve for the NAR classifier compared to DPL single classifier curve. The NAR curve shows relatively lower performance than the Cons curve, but better performance than the NARX one.

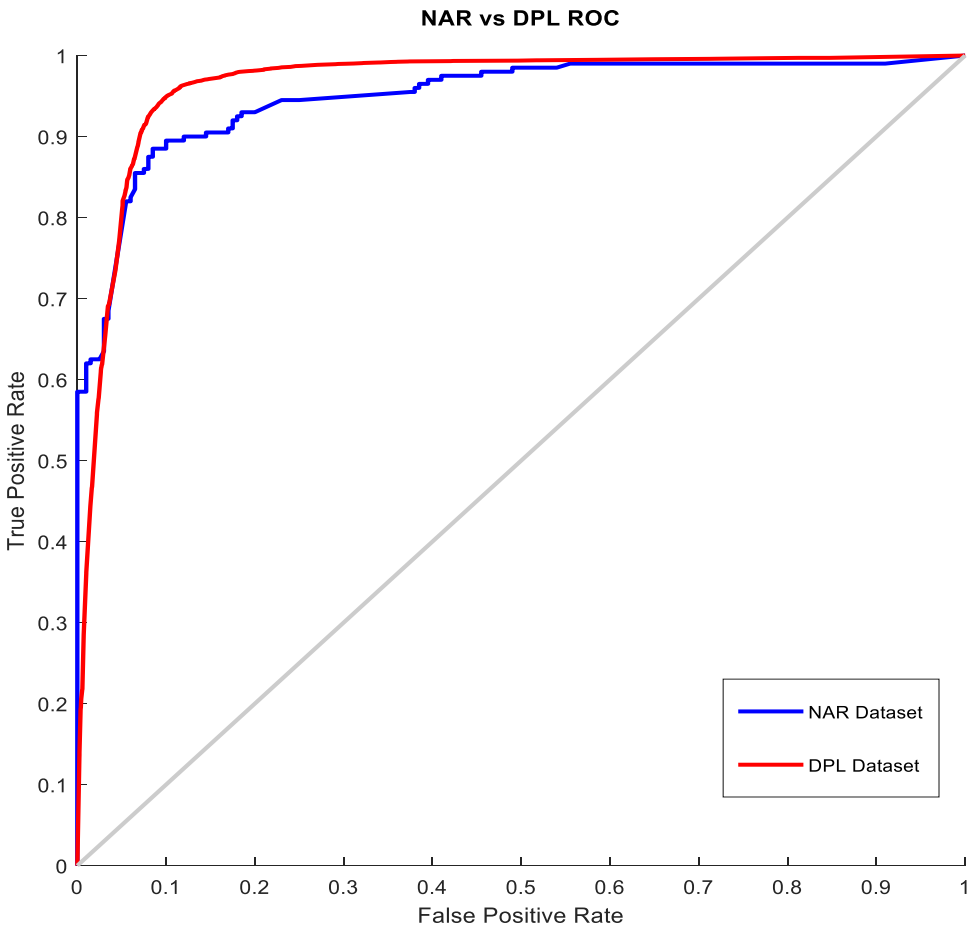

Figure 6.10: ROC curve for NAR classifier

Figure 6.11 shows the reliability diagram for NAR compared with DPL single classifier. The diagram shows bad performance in comparison with DPL.

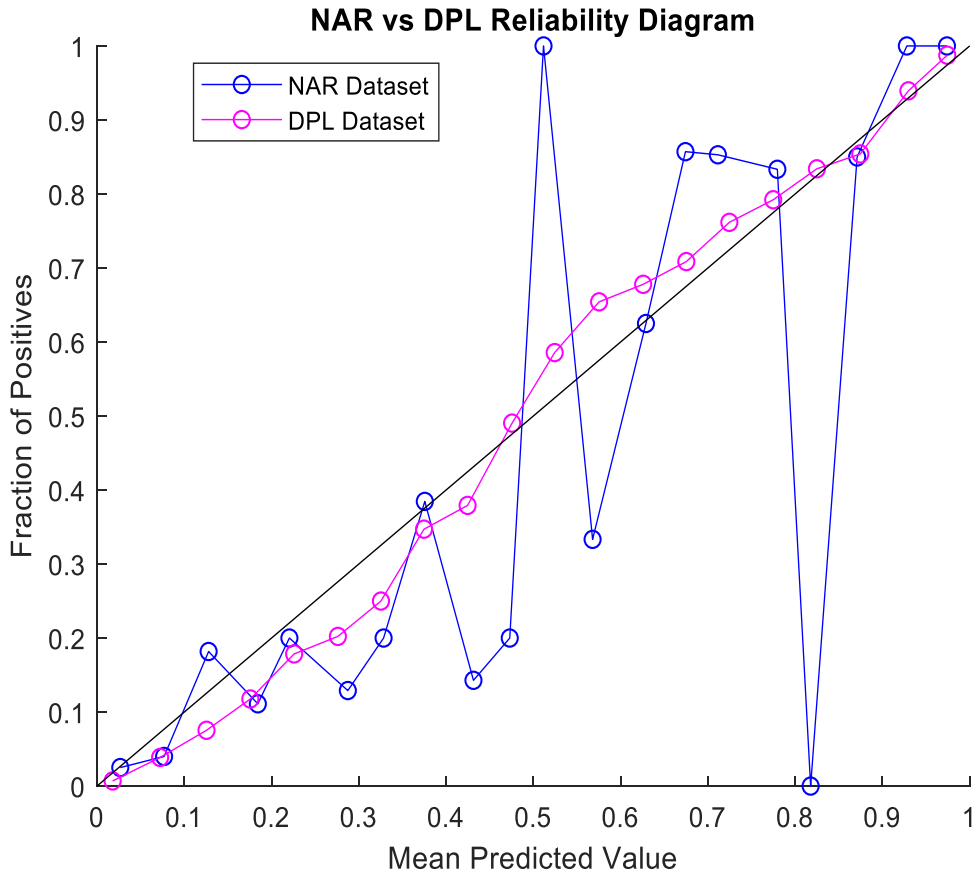

Figure 6.11: Reliability diagram for NAR classifier

6.6.3. DPL-SA Results

Based on Table 6.3, DPL-SA achieved better results than NAR and NARX, and the average accuracy rate of 91.35% is the highest among the three approaches. According to the results, a major enhancement to the classifier performance resulted from its higher sensitivity rate, with a value of 94.7%. Moreover, the specificity rate of 88% indicates reduced Type I Error. The Brier score of 0.0534 and AUC value of 93.43% indicate how DPL-SA has enhanced prediction results when predicting one step ahead, but it is still outperformed by Cons results using same year dataset.

Table 6.3: DPL-SA results

	DPL-SA	**DPL**
Average Accuracy	91.35%	96.3%
Type I Error	12%	2.8%
Type II Error	5.3%	4.5%
Sensitivity	94.7%	97.2%
Specificity	88%	95.5%
AUC	93.43%	99.35%
Brier Score	0.0534	0.0282
Area Under Reliability Curve	0.1402	0.0405

Figure 6.12 shows the ROC curve for the DPL-SA classifier compared to DPL single classifier curve. The curve outperformed both NAR and NARX based on AUC values, as it achieved the highest value. However, single classifier DPL curve has better performance in predicting company status than DPL-SA using the same year dataset (instead of predicting one year ahead).

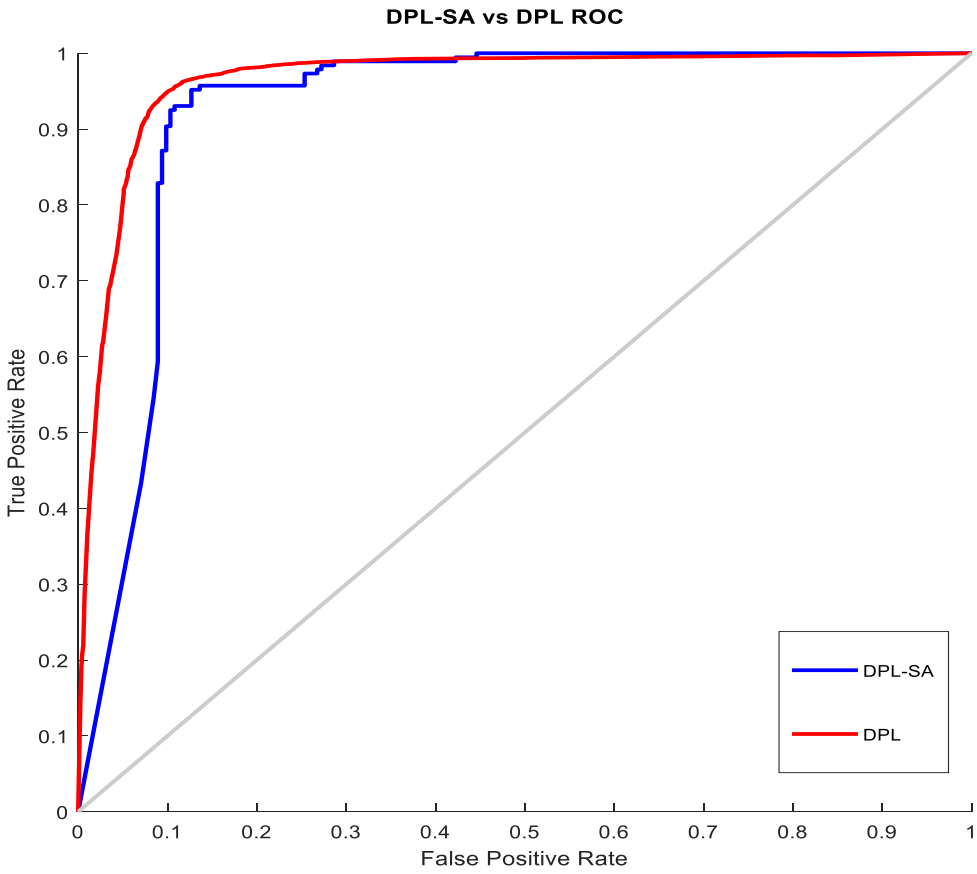

Figure 6.12: ROC curve for DPL-SA classifier

Figure 6.13 shows the reliability diagram for DPL-SA compared to DPL classifier. DPL-SA has relatively better performance than NAR and NARX, but it is still worse than the DPL single classifier.

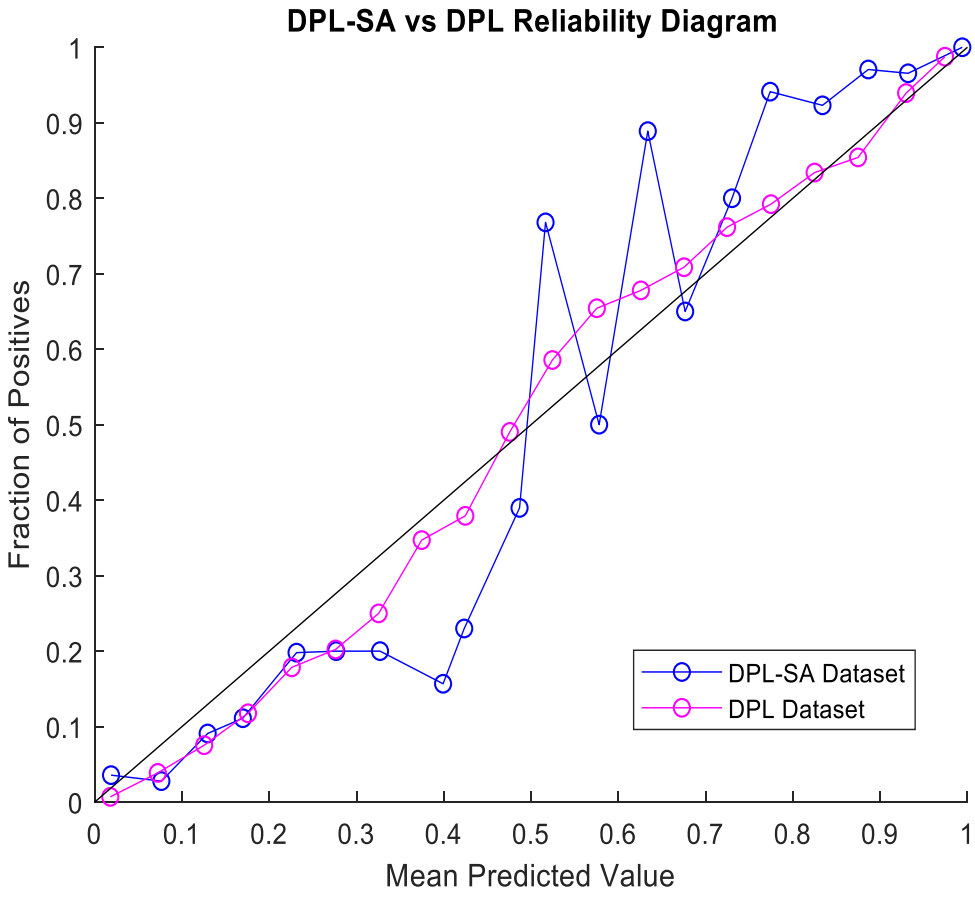

Figure 6.13: Reliability diagram for DPL-SA classifier

6.7. Discussion and Analysis

In this section the results of each classifier for one step ahead prediction are analysed and discussed based on all performance results, in addition to the reliability diagrams and ROC curves of each classifier.

Average accuracy rate is considered to be the most obvious measurement to compare classifiers' performance, which shows how accurately the classifier is capable to classify companies' status based on their financial inputs. Whilst conducting the experiment to predict one year ahead about companies' status, the average accuracy rate declined in comparison with the best single classifier, Cons. Using actual input data in year five, Cons classifier achieved 96.7% average accuracy, while step ahead models achieved 83.6%, 89.5%, and 91.35% using NARX, NAR, and DPL-SA, respectively. The difference between DPL single classifier and NARX is the largest, and it is considered to be the worst performer in this regard. However, NARX average accuracy is considered to be acceptable for a model predicting companies' status one year in advance. Both NAR and DPL-SA outperformed NARX based on all

performance measures, and showed higher average accuracy rates. Here, it should be noted that the NAR model output is different from that of NARX and DPL-SA, as it predicts the next year's financial variables as an output. These financial variables are used as inputs in a DPL single classifier model to classify companies' status, which revealed enhanced prediction performance compared to NARX. Obviously, DPL-SA achieved the highest average accuracy and provided an outstanding capability to accurately classify companies one year in advance.

Based on specificity and sensitivity rates, DPL-SA is ranked first, with the best performance results of 88% and 94.7%, respectively. NARX method achieved the worst results, with 82.5% specificity rate and 84.7% sensitivity rate while NAR fell in the middle, with 87.3% and 91% (respectively). Obviously, it is more efficient to use NAR approach than NARX when predicting companies' status in advance, as it shows substantial improvements in correctly classifying failed firms. It can be seen from the sensitivity rates that using DPL-SA improved active companies' prediction by 3.7% compared to the NAR method. However, the specificity rate is only enhanced by 0.7% by DPL-SA, which means that both methods have similar capability of correctly predicting failed companies.

DPL-SA has the lowest Type I and Type II Error rates as a step ahead prediction model, with 5.3% and 12% (respectively), and it considered to be the closest to the actual year data result achieved by the DPL classifier. DPL-SA outperformed NAR and NARX based on Brier score, AUC, and AURC values, and showed itself to be a more reliable classifier. Its values considered to be the closest to the actual yearly data results.

Table 6.4: One step ahead classifiers' results.

	One Step Ahead							
	Aver Acc.	Type II Err	Type I Err	Sensitivity	Specificity	AUC	Brier score	AURC
NARX	83.6%	15.3%	17.5%	84.7%	82.5%	87.85%	0.0958	14.27%
NAR	89.5%	9%	12.7%	91%	87.3%	92.48%	0.0876	21.44%
DPL-SA	91.35%	5.3%	12%	94.7%	88%	93.4%	0.0789	14.02%
DPL	94.1%	3.6%	8.2%	96.4%	91.8%	98%	0.0664	13.34%

6.8. Dynamic Modelling Training Time

Table 6.5 shows the computational time required for each dynamic classifier to train the dataset. It can be seen from the table that DPL-SA has achieved the higher training time due to its computational complexity at 1060.18 seconds. NARX model had the lowest training time at 56.98, indicating that this

is the fastest training model for classifying business failure one step ahead. However, due to higher average accuracy rate achieved by DPL-ST, it still considered the best prediction method despite its higher training time.

Table 6.5: Dynamic model training times in seconds

Dataset	Training time (seconds)
DPL-LSTM	1060.18
NARX	56.98
NAR	95.37

6.9. Summary

In summary, the DPL-SA shows the best performance in predicting business status one year ahead. It achieved the best prediction results based on all performance measurements. On the other hand, the performance measurements indicate that NARX model achieved the lowest average accuracy rate in predicting business failure a year ahead. The model had relatively low performance in predicting both classes of business activity, with high rates of Type I and Type II Error. The NAR model was more accurate, with better prediction capability, than the NARX model, based on average accuracy. The model was able to predict business status 4.8% more accurately than NARX.

DPL-SA as a modelling technique has revealed a substantial capability of predicting business failure and has outperformed NARX and NAR based on all performance measurements. The model result is relatively close to the results of the static DPL classifier, indicating reliable prediction performance. The model can reliably be deployed on real time series data to help users of financial statement from different backgrounds to have early signs about firms' health and operational status as on-going concerns.

Chapter 7
Conclusions and Future Work

7.1. Conclusions

The main aim of this thesis was to explore the deployment of multiple data mining modelling techniques to solve business failure prediction problems for UK firms by investigating which modelling techniques can achieve the best prediction performance based on specific measurements. The modelling process started by collecting financial data of businesses in the UK for both failed and active firms. Individual base classifiers were then built using both machine learning and statistical modelling approach as a benchmark. More advanced data mining techniques were used to determine the extent to which machine learning models can outperform traditional methods in predicting business failure in the UK. To investigate improvements for single classifiers' results, ensemble classifiers were developed by applying traditional combiners. Finally, time series modelling techniques were proposed for the first time in the field of business failure prediction to predict failure one step ahead (before it occurs). All proposed models were validated using eight performance measurements that reflect classifiers' prediction capability. After all proposed classifiers were developed and tested, their results were statistically tested to investigate their significance against other classifiers to determine the best performer.

Chapter 2 explained the theoretical background of business failure analysis, including definitions and pertinent issues. It reviewed the data sources used to construct and develop prediction models, which in most cases comprise financial information related to each business included in the dataset. Another step was reviewing related literature on business failure modelling techniques and algorithms used to achieve a good prediction performance. The literature was analysed thoroughly, and several findings were drawn that helped investigate how to improve business failure prediction for UK firms.

Chapter 3 focused on the methodological approach used in this study, explaining the main stages of the experimental design adopted on business failure modelling. Several issues were considered in developing the proposed model, such as:

- Modelling techniques used to build the classification and prediction model.
- The database collection process, focusing on the size and type of data.

- Data pre-processing and partitioning techniques.
- Performance measurements used to assess model's prediction results, accompanied with statistical significance testing.

These considerations were necessary to demonstrate the stages that highly affect the prediction performance of the proposed models adopted in this thesis, to help in developing an appropriate and comprehensive business failure model for the UK dataset.

Chapter 4 expounded on the individual classifying approaches used for the UK dataset. Initially, LR classifier was implemented and applied as a benchmark classifier against which all other data mining classifiers were compared. The analysed data mining classifiers were: NN, DT, SVM, NB, KNN, DPL, and ENS-DT. Each classifier behaves differently and has its own strengths and weaknesses in classifying and predicting business failure in the UK. In general, the best classification performance was achieved by DPL, followed by ENS-DT and then DT. It was proven that data mining techniques outperformed statistical methods (LDA and LR) in terms of all performance measurements.

Chapter 5 demonstrated several experimental approaches using traditional committee machine combiners in an attempt to enhance individual classifiers' classification performance. The traditional combining methods adopted were MIN, MAX, AVG, MajVot, and WAVG. For the first time in business failure modelling, two new combiners were used to enhance prediction performance: Consensus (Cons) and fuzzy logic. The experimental results showed that the best traditional combiners were WAVG and AVG combiner, which outperformed ENS-DT and DT individual classifiers, as well as combining methods using fuzzy logic method, but they could not outperform DPL. The main contribution of the chapter was to demonstrate the superior performance of the Cons classifier, which achieved the best results in correctly classifying failed firms.

Chapter 6 focused on developing dynamic prediction models capable to predict business failure one step ahead. Three main modelling techniques were adopted in this thesis: NARX, DPL-SA, and NAR. The first two methods focused on predicting business status one year ahead (before failure), while NAR was used to forecast the next values of the financial variables used to predict firm's status, then these new variables were fed to a DPL single classifier to produce business predictions. Based on the classification results of each classifier, the DPL-SA method outperformed the NARX classifier, but the NAR classifier achieved the best results when used to predict the next values of the financial variables used to classify businesses.

In conclusion, it has been proven that machine learning classification methods have outperformed traditional statistical techniques in terms of all performance measurements. LDA and LR as statistical classifiers, have achieved the lowest average accuracy rates for all data dataset with only 81.8%, and 75%, respectively. Where the best single classifiers, DT, ENS-DT, and DPL are considered as the best individual classifiers among all other classifiers used in the study and have shown higher average accuracy rate of 95.3%, 96.2%, and 97.2%, respectively. Therefore, it can be seen that deep learning methodology for classifying business failure is considered the best method can be used for UK datasets. This advantage of this methods relies on the capability of the classifier to assign more weights to the best features on the datasets where it forgets the features with the lowest classification capability.

For the combining methods that used all single classifiers output to enhance the final classification result of business failure, traditional mathematical combiners (MIN, MAX, AVG, Maj_vote, Median, and WAVG) have shown good results in term of accuracy rates. However, the best method among these combiners, which is the WAVG, still has achieved lower accuracy than the individual deep learning classifier with 96.7% accuracy rate. Therefore, the fuzzy logic and the consensus (Cons) methods were developed to achieve better results.

The Cons model of business failure is considered the best static classification method in this study, and has the advantage over traditional combiners as it uses a fusion of individual classifiers' predictions rather than combining these predictions using logical, arithmetical or other mathematical functions. The model mimics the behaviour of a real expert group who they are constantly exchanging their opinions and adjusting their estimates of possible outcomes based on the advice of other experts. However, just as experts sometimes cannot agree on a decision, sometimes the Cons model cannot converge. In order to overcome this obstacle, the least square error methodology was adopted in the model instead of iteration procedure. In conclusion, consensus combiner has improved business failure classification performance in terms of accuracy and achieved the highest classification rate with 97.7%. Moreover, due its lower type I error of 2.4%, that's reflects its ability to correctly classifies failure businesses, the method is considered the best among all classifiers in this study.

The model was tested on three UK datasets, for the years 2019, 2018, and 2017, with the aim of correctly classify business failure, in addition to the All-Data dataset, which included the three individual yearly datasets. It was found from all classifiers' results that Big Data can

enhance classifiers' predictions and classifications. Based on all performance measurements, Cons outperformed the alternative models, with good specificity rates and Type I Error, reflecting the best failure classification performance.

Another substantial contribution is the deployment of the dynamic model to predict business failure one step ahead, before it occurs. DPL-SA as a modelling technique revealed a substantial capability of predicting business failure, and it outperformed NARX and NAR based on all performance measurements. The model results are relatively close to the result of the static DPL classifier results, indicating reliable prediction performance. The model can reliably be deployed on real time series data to help users of financial statement and stakeholders from various background to have early signs about firms' health and operational status as on-going concerns.

7.2. Limitations

Similar to every research, this research comes with some limitations. Perpetrators in most cases fully understand bankruptcy laws, and the responsibility to declare bankruptcy when their financial variables indicate insolvency business. Thus, the inclusion of insolvent businesses in the data would be wrong and could affect the classification accuracy performance of the classification model.

The best combiner method Cons has limitations mostly with processor duration needed to train all single classifiers and the processor time needed to adjusting Cons combiner parameters to properly fit the data.

In the case of the dynamic prediction classifier, the original data has to be clustered and reduce to a small representative dataset. In which the classifier could not separately predict all instances included in the original dataset.

7.3. Future Works

In order to improve model's classification performance for bankrupt businesses, future research should take into consideration the inclusion of qualitative variables in the data, which can be developed to serve as an indicator of possible insolvency. Such documentation can be completed annually by firms, similar to financial reporting systems. Such data can be combined with financial variables to develop more reliable and robust prediction models. Moreover, qualitative data based on a questionnaire based on theoretical factors in insolvency can be

collect from failed companies to be added to the financial data, in order to further improve prediction capabilities.

Also, the proposed model could be enhanced by:

- Considering the combination of the top three classifiers (Cons, DPL, ENS-DT) to produce better classification results.
- Investigating the extent to which Cons classification can change when combining different homogenous classifiers or heterogeneous classifiers.
- Applying new pre-processing techniques, such as new feature-selection or data filtering, and accordingly investigating how these methods could affect Cons results.

References

Abraham, A. (2004) ‘A model of financial performance analysis adapted for non-profit organisations’.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O. and Bilal, M. (2018) ‘Systematic review of bankruptcy prediction models: Towards a framework for tool selection’, *Expert Systems with Applications*, 94, pp. 164-184. doi: 10.1016/j.eswa.2017.10.040.

Alasadi, S. A. and Bhaya, W. S. (2017) ‘Review of data pre-processing 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.

Aljawazneh, H., Mora, A. M., Garcia-Sanchez, P. and Castillo-Valdivieso, P. A. (2021) ‘Comparing the performance of deep learning methods to predict companies’ financial failure’, *IEEE Access*, 9, pp. 97010-97038. doi: 10.1109/ACCESS.2021.3093461.

Alles, M. and Gray, G. L. (2015) *The pros and cons of using Big Data in auditing: A synthesis of the literature and a research agenda.* Available at: http://jebcl.com/symposium/wp-content/uploads/2015/09/The-Pros-and-Cons-of-Using-Big-Data-in-Auditing-A-Synthesis-of-the-Literature-UWCISA-Revised.pdf (Accessed: 9 June, 2021).

Altman, E.I. (1968) ‘Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy’, The Journal of Finance (New York), vol. 23, no. 4, pp. 589-609 ISSN 0022-1082. DOI 10.1111/j.1540-6261.1968.tb00843.x.

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.

Arieshanti, I., Purwananto, Y., Ramadhani, A., Nuha, M. U. and Ulinnuha, N. (2013) ‘Comparative study of bankruptcy prediction models’, *Telkomnika*, 11(3), pp. 591.

Azayite, F. Z. and Achchab, S. (2016) ‘Hybrid discriminant neural networks for bankruptcy prediction and risk scoring’, *Procedia Computer Science*, 83, pp. 670-674.

Aziz, M. A. and Dar, H. A. (2006) 'Predicting corporate bankruptcy: Where we stand?', *Corporate Governance*, 6(1), pp. 18-33. doi: 10.1108/14720700610649436.

Bai, C., Liu, Q., Lu, J., Song, F.M. and Zhang, J. (2006) 'An empirical study on corporate governance and market valuation in China', Frontiers of Economics in China, vol. 1, no. 1, pp. 83-111 ISSN 1673-3444. DOI 10.1016/j.resp.2005.07.001.

Ball, R. (2006) 'International Financial Reporting Standards (IFRS): Pros and cons for investors', *Accounting and Business Research*, 36(S1), pp. 5-27.

Ball, R., Li, X.I. and Shivakumar, L. (2015) 'Contractibility and Transparency of Financial Statement Information Prepared Under IFRS: Evidence from Debt Contracts Around IFRS Adoption', Journal of Accounting Research, vol. 53, no. 5, pp. 915-963 ISSN 0021-8456. Doi: 10.1111/1475-679X.12095.

Barboza, F., Kimura, H. and Altman, E. (2017) 'Machine learning models and bankruptcy prediction', *Expert Systems with Applications*, 83, pp. 405-417.

Barth, M. E., Landsman, W. R. and Lang, M. H. (2008) 'International accounting standards and accounting quality', *Journal of Accounting Research*, 46(3), pp. 467-498. doi: 10.1111/j.1475-679X.2008.00287.x.

Barua, B., Barua, S. and Rana, R.H. (2018) 'Determining the Financial Performance of Non-Life Insurers: Static and Dynamic Panel Evidence from an Emerging Economy', The Journal of Developing Areas, vol. 52, no. 3, pp. 153-167 ISSN 0022-037X. DOI 10.1353/jda.2018.0043.

Bateni, L. and Asghari, F. (2020) 'Bankruptcy prediction using logit and genetic algorithm models: A comparative analysis', *Computational Economics*, 55(1), pp. 335-348.

Beaver, W.H. (1966) 'Financial ratios as predictors of failure', Journal of Accounting Research, pp. 71-111.

Beaver, W.H. (1968) 'Market prices, financial ratios, and the prediction of failure', Journal of Accounting Research, pp. 179-192.

Behn, B. K., Kaplan, S. E. and Krumwiede, K. R. (2001) 'Further evidence on the auditor's going-concern report: The influence of management plans', *Auditing: A Journal of Practice & Theory*, 20(1), pp. 13-28.

Bell, T. B. and Tabor, R. H. (1991) 'Empirical analysis of audit uncertainty qualifications', *Journal of Accounting Research*, 29(2), pp. 350-370.

Benyoussef, N. and Khan, S. (2017) 'Identifying fraud using restatement information', *Journal of Financial Crime*.

Bertsimas, D. and Kallus, N. (2014) *From predictive to prescriptive analytics*. Available at: https://arxiv.org/pdf/1402.5481.pdf (Accessed: 9 June, 2021).

Bešlić Obradović, D., Jakšić, D., Bešlić Rupić, I. and Andrić, M. (2018) 'Insolvency prediction model of the company: The case of the Republic of Serbia', *Economic Research-Ekonomska Istraživanja*, 31(1), pp. 139-157.

Bhargava, M., Bhardwaj, A. and Rathore, A.P.S. (2017) 'Prediction model for telecom postpaid customer churn using Six-Sigma methodology', International Journal of Manufacturing Technology and Management, vol. 31, no. 5, pp. 387-401 ISSN 1368-2148. Doi: 10.1504/IJMTM.2017.088448.

Bischl, B., Mersmann, O., Trautmann, H. and Weihs, C. (2012) 'Resampling methods for meta-model validation with recommendations for evolutionary computation', *Evolutionary Computation*, 20(2), pp. 249-275.

Boritz, J. E. and Kennedy, D. B. (1995) 'Effectiveness of neural network types for prediction of business failure', *Expert Systems with Applications*, 9(4), pp. 503-512.

Boritz, J. E., Kennedy, D. B. and Albuquerque, Augusto De Miranda E (1995) 'Predicting corporate failure using a neural network approach', *Intelligent Systems in Accounting, Finance and Management*, 4(2), pp. 95-111.

Bozsik, J. (2010) 'Artificial neural networks in default forecast', *Proceedings of the 8th International Conference on Applied Informatics*, 1, pp. 31–39. Available at: http://icai.ektf.hu/pdf/ICAI2010-vol1-pp31-39.pdf (Accessed: 9 June, 2021).

Brier, G. W. (1950) 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, 78(1), pp. 1-3.

Brown, I. and Mues, C. (2012) 'An experimental comparison of classification algorithms for imbalanced credit scoring data sets', *Expert Systems with Applications*, 39(3), pp. 3446-3453.

Charitou, A., Neophytou, E. and Charalambous, C. (2004) 'Predicting corporate failure: Empirical evidence for the UK', *European Accounting Review*, 13(3), pp. 465-497. doi: 10.1080/0963818042000216811.

Chava, S. and Jarrow, R.A. (2004) 'Bankruptcy prediction with industry effects', Review of Finance, vol. 8, no. 4, pp. 537-569.

Chen, H., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S. and Liu, D. (2011) 'A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method', *Knowledge-Based Systems*, 24(8), pp. 1348-1359.

Chen, K. C. and Church, B. K. (1992) 'Default on debt obligations and the issuance of going-concern opinions', *Auditing*, 11(2), pp. 30.

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J. and Neves, J. C. (2011) 'A genetic algorithm-based approach to cost-sensitive bankruptcy prediction', *Expert Systems with Applications*, 38(10), pp. 12939-12945.

Chen, S. (2016) 'Detection of fraudulent financial statements using the hybrid data mining approach', *SpringerPlus*, 5(1), pp. 1-16. doi: 10.1186/s40064-016-1707-6.

Chiu, C., Ku, Y., Lie, T. and Chen, Y. (2011) 'Internet auction fraud detection using social network analysis and classification tree approaches', *International Journal of Electronic Commerce*, 15(3), pp. 123-147. doi: 10.2753/JEC1086-4415150306.

Cho, S., Hong, H. and Ha, B. (2010) 'A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction', *Expert Systems with Applications*, 37(4), pp. 3482-3488.

Choi, H., Son, H. and Kim, C. (2018) 'Predicting financial distress of contractors in the construction industry using ensemble learning', *Expert Systems with Applications*, 110, pp. 1-10.

Clarkson, P. M., Li, Y., Richardson, G. D. and Vasvari, F. P. (2011) 'Does it really pay to be green? Determinants and consequences of proactive environmental strategies', *Journal of Accounting and Public Policy*, 30(2), pp. 122-144. doi: 10.1016/j.jaccpubpol.2010.09.013.

Coats, P.K. and Fant, L.F. (1991) 'A neural network approach to forecasting financial distress', The Journal of Business Forecasting, vol. 10, no. 4, pp. 9.

Dakovic, R., Czado, C. and Berg, D. (2010) ‘Bankruptcy prediction in Norway: a comparison study’, Applied Economics Letters, vol. 17, no. 17, pp. 1739-1746.

Dalnial, H., Kamaluddin, A., Sanusi, Z. M. and Khairuddin, K. S. (2014) ‘Detecting fraudulent financial reporting through financial statement analysis’, *Journal of Advanced Management Science*, 2, pp. 17-22.

De Andrés, J., Lorca, P., De Cos Juez, Francisco Javier and Sánchez-Lasheras, F. (2011) ‘Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS)’, *Expert Systems with Applications*, 38(3), pp. 1866-1875.

De Bock, K. W. (2017) ‘The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles’, *Expert Systems with Applications*, 90, pp. 23-39. doi: 10.1016/j.eswa.2017.07.036.

De Bock, K.W. (2017) ‘The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles’, Available from:http://www.sciencedirect.com.ezproxy.brunel.ac.uk/science/article/pii/S0957417417305122.

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

Diebold, F. X. (2012) *On the origin(s) and development of the term ‘Big Data’*. PIER Working Paper No. 12-037. doi: 10.2139/ssrn.2152421.

Du Jardin, P. (2010) ‘Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy’, *Neurocomputing (Amsterdam)*, 73(10), pp. 2047-2060. doi: 10.1016/j.neucom.2009.11.034.

Du Jardin, P. (2015) ‘Bankruptcy prediction using terminal failure processes’, *European Journal of Operational Research*, 242(1), pp. 286-303.

Du Jardin, P. (2021) ‘Forecasting bankruptcy using biclustering and neural network-based ensembles’, *Annals of Operations Research*, 299(1), pp. 531-566.

Du Jardin, P. and Séverin, E. (2012) ‘Forecasting financial failure using a Kohonen map: A comparative study to improve model stability over time’, *European Journal of Operational Research*, 221(2), pp. 378-396.

Eng, L.L., Tian, X. and Robert Yu, T. (2018) 'Financial statement analysis: Evidence from Chinese firms', Review of Pacific Basin Financial Markets and Policies, vol. 21, no. 4, pp. 1850027 ISSN 0219-0915. DOI 10.1142/S0219091518500273.

Falangis, K. and Glen, J. J. (2010) 'Heuristics for feature selection in mathematical programming discriminant analysis models', *Journal of the Operational Research Society*, 61(5), pp. 804-812.

Falavigna, G. and Ippoliti, R. (2018) 'Industrial spatial dynamics, financial health and bankruptcy: Evidence from Italian manufacturing industry', *Economia E Politica Industriale*, 45(4), pp. 533-554.

FAME (2019) Forecasting Analysis and Modelling Environment. Available at: https://fame.bvdinfo.com/version-20211216/fame/1/Companies/Search.

Fan, S., Liu, G. and Chen, Z. (2017) 'Anomaly detection methods for bankruptcy prediction', *4th International Conference on Systems and Informatics (ICSAI)*, 17, pp. 1456-1460. doi: 10.1109/ICSAI.2017.8248515.

Fan, W. and Bifet, A. (2013) 'Mining Big Data: Current status, and forecast to the future', *ACM SIGKDD Explorations Newsletter*, 14(2), pp. 1-5.

Fitzpatrick, P.J. (1932) 'A comparison of the ratios of successful industrial enterprises with those of failed companies'.

Fletcher, D. and Goss, E. (1993) 'Forecasting with neural networks: An application using bankruptcy data', *Information & Management*, 24(3), pp. 159-167.

Freund, Y. and Schapire, R. E. (1997) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences*, 55(1), pp. 119-139.

Friedman, M. (1940) 'A comparison of alternative tests of significance for the problem of m rankings', *Annals of Mathematical Statistics*, 11(1), pp. 86-92.

Gaeremynck, A. and Willekens, M. (2003) 'The endogenous relationship between audit-report type and business termination: Evidence on private firms in a non-litigious environment', *Accounting and Business Research*, 33(1), pp. 65-79.

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.

Garrison, R.H., Noreen, E.W., Brewer, P.C. and Mcgowan, A. (2010) 'Managerial Accounting', Issues in Accounting Education, vol. 25, no. 4, pp. 792-793 ISSN 0739-3172. DOI 10.2308/iace.2010.25.4.792.

Geng, R., Bose, I. and Chen, X. (2015) 'Prediction of financial distress: An empirical study of listed Chinese companies using data mining', *European Journal of Operational Research*, 241(1), pp. 236-247.

Goo, Y. J., Chi, D. and Shen, Z. (2016) 'Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques', *SpringerPlus*, 5(1), pp. 1-18.

Gordini, N. (2014) 'A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy', *Expert Systems with Applications*, 41(14), pp. 6433-6445.

Gov.uk. (2019) *Applying to become bankrupt.* Available at: https://www.gov.uk/bankruptcy (Accessed: 7 June, 2021).

Gray, G. L. and Debreceny, R. S. (2014) '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.

Guang-Bin Huang, Qin-Yu Zhu and Chee-Kheong Siew (2004) 'Extreme learning machine: a new learning scheme of feedforward neural networks Anonymous', IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2004.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, 3, no', Mar, pp. 1157-1182.

Hajek, P. and Henriques, R. (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.

Härdle, W.K., Prastyo, D. and Hafner, C. (2012) 'Support vector machines with evolutionary feature selection for default prediction'.

Heo, J. and Yang, J. Y. (2014) 'AdaBoost based bankruptcy forecasting of Korean construction companies', *Applied Soft Computing*, 24, pp. 494-499.

Hitzler, P. and Janowicz, K. (2013) 'Linked data, Big Data, and the 4th Paradigm', *Semantic Web*, 4(3), pp. 233-235.

Holsapple, C., Lee-Post, A. and Pakath, R. (2014) 'A unified foundation for business analytics', *Decision Support Systems*, 64, pp. 130-141.

Huang, S., Tsaih, R. and Yu, F. (2014) '*Topological pattern discovery and feature extraction for fraudulent financial reporting*', *Expert Systems with Applications*, 41(9), pp. 4360-4372. doi: 10.1016/j.eswa.2014.01.012.

Huang, Y. and Yen, M. (2019) 'A new perspective of performance comparison among machine learning algorithms for financial distress prediction', *Applied Soft Computing*, 83, pp. 105663. doi: 10.1016/j.asoc.2019.105663.

Huynh, T. L. D., Wu, J. and Duong, A. T. (2020) 'Information asymmetry and firm value: Is Vietnam different?', *Journal of Economic Asymmetries*, 21, pp. e00147. doi: 10.1016/j.jeca.2019.e00147.

Iatridis, G. (2010a) 'IFRS adoption and financial statement effects: The UK case', *International Research Journal of Finance and Economics*, 38, pp. 165-172.

Iatridis, G. (2010b) 'International Financial Reporting Standards and the quality of financial statement information', *International Review of Financial Analysis*, 19(3), pp. 193-204.

Ikpefan, O. A. and Akande, A. O. (2012) 'International Financial Reporting Standards (IFRS): Benefits, obstacles and intrigues for implementation in Nigeria', *Business Intelligence Journal*, 5(2), pp. 299-307.

Iturriaga, F. J. L. and Sanz, I. P. (2015) 'Bankruptcy visualization and prediction using neural networks: A study of US commercial banks', *Expert Systems with Applications*, 42(6), pp. 2857-2869.

Jabeur, S. B., Gharib, C., Mefteh-Wali, S. and Arfi, W. B. (2021) 'CatBoost model and artificial intelligence techniques for corporate failure prediction', *Technological Forecasting and Social Change*, 166, pp. 120658. doi: 10.1016/j.techfore.2021.120658.

Jacobs, A. (2009) ‘The pathologies of Big Data: Scale up your datasets enough and all your apps will come undone – What are the typical problems and where do the bottlenecks generally surface?’, *Queue*, 7(6), pp. 10-19.

Jan, C. L. (2021) ‘Using deep learning algorithms for CPAs’ going concern prediction’, *Information*, 12(2), pp. 73. doi: 10.3390/info12020073.

Jang, Y., Jeong, I., Cho, Y. K. and Ahn, Y. (2019) ‘Predicting business failure of construction contractors using long short-term memory recurrent neural network’, *Journal of Construction Engineering and Management*, 145(11), pp. 04019067.

Jeong, C., Min, J. H. and Kim, M. S. (2012) ‘A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction’, *Expert Systems with Applications*, 39(3), pp. 3650-3658.

Jing, Z. and Fang, Y. (2018) ‘Predicting US bank failures: A comparison of logit and data mining models’, *Journal of Forecasting*, 37(2), pp. 235-256.

Jones, S. (2017) ‘Corporate bankruptcy prediction: A high dimensional analysis’, *Review of Accounting Studies*, 22(3), pp. 1366-1422.

Joy, M.O. (1975) ‘On the Financial Applications of Discriminant Analysis’, Journal of Financial and Quantitative Analysis (December 1975), pp. 723-739.

Kanapickienė, R. and Grundienė, Ž. (2015) ‘The Model of Fraud Detection in Financial Statements by Means of Financial Ratio’, Procedia - Social and Behavioral Sciences, vol. 213, pp. 321-327 ISSN 1877-0428. DOI 10.1016/j.sbspro.2015.11.545.

Karas, M. and Režňáková, M. (2014) ‘A parametric or nonparametric approach for creating a new bankruptcy prediction model: The Evidence from the Czech Republic’, *International Journal of Mathematical Models and Methods in Applied Sciences*, 8(1), pp. 214-223.

Kasgari, A. A., Divsalar, M., Javid, M. R. and Ebrahimian, S. J. (2013) ‘Prediction of bankruptcy Iranian corporations through artificial neural network and Probit-based analyses’, *Neural Computing and Applications*, 23(3), pp. 927-936.

Kristóf, T. and Virág, M. (2012) ‘Data reduction and univariate splitting: Do they together provide better corporate bankruptcy prediction?’, *Acta Oeconomica*, 62(2), pp. 205-228. doi: 10.1556/AOecon.62.2012.2.4.

Kukuk, M. and Rönnberg, M. (2013) 'Corporate credit default models: A mixed logit approach', *Review of Quantitative Finance and Accounting*, 40(3), pp. 467-483.

Kulustayeva, A., Jondelbayeva, A., Nurmagambetova, A., Dossayeva, A. and Bikteubayeva, A. (2020) 'Financial data reporting analysis of the factors influencing on profitability for insurance companies', Entrepreneurship and Sustainability, Issues, vol. 7, no. 3, pp. 2394-2406 ISSN 2345-0282. DOI 10.9770/jesi.2020.7.3(62).

Kuruppu, N., Laswad, F. and Oyelere, P. (2003) 'The efficacy of liquidation and bankruptcy prediction models for assessing going concern', *Managerial Auditing Journal*, 18(6/7), pp. 577-590. doi: 10.1108/02686900310482713.

Laitinen, E. K. (2007) 'Classification accuracy and correlation: LDA in failure prediction', *European Journal of Operational Research*, 183(1), pp. 210-225. doi: 10.1016/j.ejor.2006.09.054.

Larivière, B. and Van Den Poel, D. (2005) 'Predicting customer retention and profitability by using random forests and regression forests techniques', *Expert Systems with Applications*, 29(2), pp. 472-484.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S. and Kruschwitz, N. (2011) 'Big Data, analytics and the path from insights to value', *MIT Sloan Management Review*, 52(2), pp. 21-32.

Lee, S. and Choi, W. S. (2013) 'A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis', *Expert Systems with Applications*, 40(8), pp. 2941-2946. doi: 10.1016/j.eswa.2012.12.009.

Lee, T. and Chen, I. (2005) 'A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines', *Expert Systems with Applications*, 28(4), pp. 743-752.

Lessmann, S., Baesens, B., Seow, H. and Thomas, L. C. (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.

Leuz, C. and Wysocki, P.D. (2016) 'The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research', Journal of Accounting Research, vol. 54, no. 2, pp. 525-622 ISSN 0021-8456. DOI 10.1111/1475-679X.12115.

Levitan, A. S. and Knoblett, J. A. (1985) 'Indicators of exceptions to the going concern assumption', *Auditing*, 5(1), pp. 26-39.

Lewellen, J. (2004) 'Predicting returns with financial ratios', Journal of Financial Economics, vol. 74, no. 2, pp. 209-235 ISSN 0304-405X. Doi: 10.1016/j.jfineco.2002.11.002.

Li, H. and Sun, J. (2009) 'Forecasting business failure in China using hybrid case-based reasoning', Journal of Forecasting, pp. n/a ISSN 0277-6693. DOI 10.1002/for.1149.

Li, H. and Sun, J. (2011) 'Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study', *Expert Systems with Applications*, 38(10), pp. 12747-12756. doi: 10.1016/j.eswa.2011.04.064.

Li, M. L. and Miu, P. (2010) 'A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach', *Journal of Empirical Finance*, 17(4), pp. 818-833.

Liang, D., Lu, C., Tsai, C. and Shih, G. (2016) 'Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study', *European Journal of Operational Research*, 252(2), pp. 561-572. doi: 10.1016/j.ejor.2016.01.012.

Lin, J. Y. (2008) 'The impact of the financial crisis on developing countries', *SSRN Electronic Journal*, 1(13S), pp. 7-14. doi: 10.2139/ssrn.1523363.

Lisic, L. L., Silveri, S., Song, Y. and Wang, K. (2015) 'Accounting fraud, auditing, and the role of government sanctions in China', *Journal of Business Research*, 68(6), pp. 1186-1195. doi: 10.1016/j.jbusres.2014.11.013.

Mackenzie, B., Coetsee, D., Njikizana, T., Chamboko, R., Colyvas, B. and Hanekom, B. (2012) *Wiley IFRS 2013: Interpretation and application of international financial reporting standards*. London: John Wiley & Sons.

Martens, D., Bruynseels, L., Baesens, B., Willekens, M. and Vanthienen, J. (2008) 'Predicting going concern opinion with data mining', *Decision Support Systems*, 45(4), pp. 765-777.

Matin, R., Hansen, C., Hansen, C. and Mølgaard, P. (2019) 'Predicting distresses using deep learning of text segments in annual reports', *Expert Systems with Applications*, 132, pp. 199-208.

Mbona, R.M. and Yusheng, K. (2019) 'Financial statement analysis: Principal component analysis (PCA) approach case study on China telecoms industry', AJAR (Asian Journal of Accounting Research) (Online), vol. 4, no. 2, pp. 233-245 ISSN 2443-4175. DOI 10.1108/AJAR-05-2019-0037.

Mccrum-Gardner, E. (2008) 'Which is the correct statistical test to use?', *British Journal of Oral and Maxillofacial Surgery*, 46(1), pp. 38-41.

Menon, K. and Schwartz, K. B. (1987) 'An empirical investigation of audit qualification decisions in the presence of going concern uncertainties', *Contemporary Accounting Research*, 3(2), pp. 302-315.

Mesak, D. (2019) 'Financial Ratio Analysis in Predicting Financial Conditions Distress IN Indonesia Stock Exchange', Russian Journal of Agricultural and Socio-Economic Sciences, vol. 86, no. 2, pp. 155-165 ISSN 2226-1184. DOI 10.18551/rjoas.2019-02.18.

Messier, W. F. Jr., and Hansen, J. V. (1988) 'Inducing rules for expert system development: An example using default and bankruptcy data', *Management Science*, 34(12), pp. 1403-1415. doi: 10.1287/mnsc.34.12.1403.

Mutchler, J. F. (1985) 'A multivariate analysis of the auditor's going-concern opinion decision', *Journal of Accounting Research*, 23(2), pp. 668-682. doi: 10.2307/2490832.

Mutchler, J. F., Hopwood, W. and Mckeown, J. M. (1997) 'The influence of contrary information and mitigating factors on audit opinion decisions on bankrupt companies', *Journal of Accounting Research*, 35(2), pp. 295-310.

Nam, C.W., Kim, T.S., Park, N.J. and Lee, H.K. (2008) 'Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies' Journal of Forecasting, vol. 27, no. 6, pp. 493-506.

Noreen, E.W., Brewer, P.C. and Garrison, R.H. (2011) 'Managerial accounting for manager', Includes index.

Odom, M. D. and Sharda, R. (1990) 'A neural network model for bankruptcy prediction', *IJCNN International Joint Conference on Neural Networks*, 2, pp. 163-168.

Ohlhorst, F. J. (2012) *Big Data analytics: Turning Big Data into big money*. London: John Wiley & Sons.

Ohlson, J.A. (1980) 'Financial ratios and the probabilistic prediction of bankruptcy', Journal of Accounting Research, pp. 109-131.

Olson, D. L., Delen, D. and Meng, Y. (2012) 'Comparative analysis of data mining methods for bankruptcy prediction', *Decision Support Systems*, 52(2), pp. 464-473.

Prati, R. C., Batista, G. E. A. P. A. and Monard, M. C. (2011) 'A survey on graphical methods for classification predictive performance evaluation', *IEEE Transactions on Knowledge and Data Engineering*, 23(11), pp. 1601-1618. doi: 10.1109/TKDE.2011.59.

Priego, A. M., Lizano, M. M. and Madrid, E. M. (2014) 'Business failure: Incidence of stakeholders' behavior', *Academia Revista Latinoamericana De Administración*, 27(1), pp. 75-91. doi: 10.1108/ARLA-12-2013-0188.

Prusak, B. (2018) 'Review of research into enterprise bankruptcy prediction in selected central and eastern European countries', *International Journal of Financial Studies*, 6(3), pp. 60. doi: 10.3390/ijfs6030060.

Rafiei, F. M., Manzari, S. M. 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.

Raghunandan, K. and Rama, D. V. (1995) 'Audit reports for companies in financial distress: Before and after SAS No. 59', *Auditing*, 14(1), pp. 50-63.

Rashid, C. A. (2018) 'Efficiency of financial ratios analysis for evaluating companies' liquidity', *International Journal of Social Sciences & Educational Studies*, 4(4), pp. 11. doi: 10.23918/ijsses.v4i4p110.

Richardson, F.M. and Davidson, L.F. (1984) 'On linear discrimination with accounting ratios', Journal of Business Finance & Accounting, vol. 11, no. 4, pp. 511-525.

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. 63-79.

Rita W. Y. Yip and Young, D. (2012) 'Does mandatory IFRS adoption improve information comparability?', *Accounting Review*, 87(5), pp. 1767-1789. doi: 10.2308/accr-50192.

Salehi, M. and Fard, F. Z. (2013) 'Data mining approach using practical swarm optimization (PSO) to predicting going concern: Evidence from Iranian Companies', *Journal of Distribution Science*, 11(3), pp. 5-11.

Salehi, M., Shiri, M. M. and Pasikhani, M. B. (2016) 'Predicting corporate financial distress using data mining techniques: An application in Tehran Stock Exchange', *International Journal of Law and Management*, 58(2), pp. 216-230. doi: 10.1108/IJLMA-06-2015-0028.

Samman, H.A. and Al-Jafari, M.K. (2015) 'Trading volume and stock returns volatility: Evidence from industrial firms of Oman', Asian Social Science, vol. 11, no. 24, pp. 139-146 ISSN 1911-2017. DOI 10.5539/ass.v11n24p139.

Samuel, A.L. (1959) 'Some studies in machine learning using the game of checkers', IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229.

Serrano-Cinca, C. (1997) 'Feedforward neural networks in the classification of financial information', The European Journal of Finance, vol. 3, no. 3, pp. 183-202.

Sharma, H. and Kumar, S. (2016) 'A survey on decision tree algorithms of classification in data mining', *International Journal of Science and Research (IJSR)*, 5(4), pp. 2094-2097.

Shin, K. and Lee, Y. (2002) 'A genetic algorithm application in bankruptcy prediction modelling', Expert Systems with Applications, vol. 23, no. 3, pp. 321-328.

Shin, K., Lee, T. S. and Kim, H. (2005) 'An application of support vector machines in bankruptcy prediction model', *Expert Systems with Applications*, 28(1), pp. 127-135.

Shumway, T. (2001) 'Forecasting bankruptcy more accurately: A simple hazard model', The Journal of Business, vol. 74, no. 1, pp. 101-124.

Shumway, T. (2001) 'Forecasting bankruptcy more accurately: A simple hazard model', The Journal of Business, vol. 74, no. 1, pp. 101-124.

Singh, D. and Singh, B. (2020) 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing*, 97, pp. 105524.

Smith, R.F. (1935) 'Changes in the financial structure of unsuccessful corporations', University of Illinois.

Smiti, S. and Soui, M. (2020) 'Bankruptcy prediction using deep learning approach based on borderline SMOTE', *Information Systems Frontiers*, 22(5), pp. 1067-1083.

Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management*, 45(4), pp. 427-437. doi: 10.1016/j.ipm.2009.03.002.

Son, H., Hyun, C., Phan, D. and Hwang, H. J. (2019) 'Data analytic approach for bankruptcy prediction', *Expert Systems with Applications*, 138, pp. 112816. doi: 10.1016/j.eswa.2019.07.033.

Suthaharan, S. (2014) 'Big Data classification: Problems and challenges in network intrusion prediction with machine learning', *ACM SIGMETRICS Performance Evaluation Review*, 41(4), pp. 70-73.

Taffler, R.J. (1982) 'Forecasting company failure in the UK using discriminant analysis and financial ratio data', Journal of the Royal Statistical Society: Series A (General), vol. 145, no. 3, pp. 342-358.

Talia, D. (2013) 'Clouds for scalable Big Data analytics', *Computer*, 46(05), pp. 98-101.

Tam, K. Y. and Kiang, M. Y. (1992) 'Managerial applications of neural networks: The case of bank failure predictions', *Management Science*, 38(7), pp. 926-947.

Teng, S., Du, H., Wu, N., Zhang, W. and Su, J. (2010) 'A cooperative network intrusion detection based on fuzzy SVMs', *Journal of Networks*, 5(4), pp. 475-484.

Tian, S., Yu, Y. and Guo, H. (2015) 'Variable selection and corporate bankruptcy forecasts', Journal of Banking & Finance, vol. 52, pp. 89-100.

Traczynski, J. (2017) 'Firm default prediction: A Bayesian model-averaging approach', Journal of Financial and Quantitative Analysis, vol. 52, no. 3, pp. 1211-1245.

Tsai, C. and Cheng, K. (2012) 'Simple instance selection for bankruptcy prediction', *Knowledge-Based Systems*, 27, pp. 333-342. doi: 10.1016/j.knosys.2011.09.017.

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

Tseng, F. and Hu, Y. (2010) 'Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks', *Expert Systems with Applications*, 37(3), pp. 1846-1853.

Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S. K. (2020) 'Financial crisis prediction model using ant colony optimization', *International Journal of Information Management*, 50, pp. 538-556.

Veganzones, D. and Séverin, E. (2018) 'An investigation of bankruptcy prediction in imbalanced datasets', *Decision Support Systems*, 112, pp. 111-124.

Vuran, B. (2009) 'Prediction of business failure: A comparison of discriminant and logistic regression analyses. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 38(1), pp. 47-63.

Wang, G., Ma, J. and Yang, S. (2014) 'An improved boosting based on feature selection for corporate bankruptcy prediction', *Expert Systems with Applications*, 41(5), pp. 2353-2361.

Watson, J. and Everett, J. (1993) 'Defining small business failure', *International Small Business Journal*, 11(3), pp. 35-48.

Whitehead, J. (1980) 'Fitting Cox's regression model to survival data using GLIM', Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 29, no. 3, pp. 268-275.

Williams, D. A. (2016) 'Can neural networks predict business failure? Evidence from small high tech firms in the UK', *Journal of Developmental Entrepreneurship*, 21(1), pp. 1650005. doi: 10.1142/S1084946716500059.

Wilson, R. L. and Sharda, R. (1994) 'Bankruptcy prediction using neural networks', *Decision Support Systems*, 11(5), pp. 545-557.

Wruck, K.H. (1990) 'Financial distress, reorganization, and organizational efficiency', Journal of Financial Economics, vol. 27, no. 2, pp. 419-444.

Yang, Z. R., Platt, M. B. and Platt, H. D. (1999) 'Probabilistic neural networks in bankruptcy prediction', *Journal of Business Research*, 44(2), pp. 67-74.

Yeh, C., Chi, D. and Lin, Y. (2014) 'Going-concern prediction using hybrid random forests and rough set approach', *Information Sciences*, 254, pp. 98-110.

Yoon, J. S. and Kwon, Y. S. (2010) 'A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information', *Expert Systems with Applications*, 37(5), pp. 3624-3629. doi: 10.1016/j.eswa.2009.10.029.

Youn, H. and Gu, Z. (2010) 'Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model', *International Journal of Hospitality Management*, 29(1), pp. 120-127. doi: 10.1016/j.ijhm.2009.06.007.

Zavgren, C. V. (1985) 'Assessing the vulnerability to failure of American industrial firms: A logistic analysis', *Journal of Business Finance & Accounting*, 12(1), pp. 19-45.

Zhang, Y.D. and L.N. WU. (2011) 'Bankruptcy prediction by genetic ant colony algorithm Anonymous Advanced Materials Research'.

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

Zhou, L., Lai, K. K. and Yen, J. (2014) 'Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation', *International Journal of Systems Science*, 45(3), pp. 241-253. doi: 10.1080/00207721.2012.720293.

Zięba, M., Tomczak, S. K. and Tomczak, J. M. (2016) 'Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction', *Expert Systems with Applications*, 58, pp. 93-101.

Zikopoulos, P. and Eaton, C. (2011) *Understanding Big Data: Analytics for enterprise class hadoop and streaming data*. New York: McGraw-Hill Osborne Media.