International Series in
Operations Research & Management Science

Mohammad Zoynul Abedin
Petr Hajek *Editors*

# Novel Financial Applications of Machine Learning and Deep Learning

## Algorithms, Product Modeling, and Applications

Springer

# International Series in Operations Research & Management Science


**Founding Editor**

Frederick S. Hillier, Stanford University, Stanford, CA, USA


Volume 336


**Series Editor**

Camille C. Price, Department of Computer Science, Stephen F. Austin State University, Nacogdoches, TX, USA

**Editorial Board Members**

Emanuele Borgonovo, Department of Decision Sciences, Bocconi University, Milan, Italy

Barry L. Nelson, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, USA

Bruce W. Patty, Veritec Solutions, Mill Valley, CA, USA

Michael Pinedo, Stern School of Business, New York University, New York, NY, USA

Robert J. Vanderbei, Princeton University, Princeton, NJ, USA

**Associate Editor**

Joe Zhu, Foisie Business School, Worcester Polytechnic Institute, Worcester, MA, USA

The book series **International Series in Operations Research and Management Science** encompasses the various areas of operations research and management science. Both theoretical and applied books are included. It describes current advances anywhere in the world that are at the cutting edge of the field. The series is aimed especially at researchers, advanced graduate students, and sophisticated practitioners.

The series features three types of books:

• Advanced expository books that extend and unify our understanding of particular areas.

• Research monographs that make substantial contributions to knowledge.

• Handbooks that define the new state of the art in particular areas. Each handbook will be edited by a leading authority in the area who will organize a team of experts on various aspects of the topic to write individual chapters. A handbook may emphasize expository surveys or completely new advances (either research or applications) or a combination of both.

The series emphasizes the following four areas:

**Mathematical Programming:** Including linear programming, integer programming, nonlinear programming, interior point methods, game theory, network optimization models, combinatorics, equilibrium programming, complementarity theory, multiobjective optimization, dynamic programming, stochastic programming, complexity theory, etc.

**Applied Probability:** Including queuing theory, simulation, renewal theory, Brownian motion and diffusion processes, decision analysis, Markov decision processes, reliability theory, forecasting, other stochastic processes motivated by applications, etc.

**Production and Operations Management:** Including inventory theory, production scheduling, capacity planning, facility location, supply chain management, distribution systems, materials requirements planning, just-in-time systems, flexible manufacturing systems, design of production lines, logistical planning, strategic issues, etc.

**Applications of Operations Research and Management Science:** Including telecommunications, health care, capital budgeting and finance, economics, marketing, public policy, military operations research, humanitarian relief and disaster mitigation, service operations, transportation systems, etc.

This book series is indexed in Scopus.

Mohammad Zoynul Abedin • Petr Hajek
Editors

# Novel Financial Applications of Machine Learning and Deep Learning

## Algorithms, Product Modeling, and Applications

Springer

*Editors*
Mohammad Zoynul Abedin
Department of Finance, Performance and Marketing
Teesside University International Business School, Teesside University
Middlesbrough, UK

Petr Hajek
Faculty of Economics and Administration
University of Pardubice
Pardubice, Czech Republic

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-031-18551-9 ISBN 978-3-031-18552-6 (eBook)
https://doi.org/10.1007/978-3-031-18552-6

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The *Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modelling, and Applications* presents the state of the art of the application of machine learning (ML) and deep learning (DL) in the domain of finance. We will present a combination of empirical evidence to diverse fields of finance so that this book is useful to academics, practitioners, and policymakers who are looking to train novel and the most advanced machine learning classifiers. Thus, the purpose of this book is to provide a broad area of applications to different financial assets and markets. Furthermore, from an extensive literature assessment, it is evident that there are no existing textbooks that narrate ML and DL to unlike areas of finance or to an extensive range of products and markets.

Business risk and uncertainty certainly are the toughest challenge in the finance domain faced by many researchers and managers. Such uncertainty thereby initiates an unavoidable risk factor, which is a fundamental element of financial theory. To the best of our knowledge, the financial domain has not been a focused subject-matter for good ML related books. There is also a scarcity of information about how financial enterprises supervise crisis events and achieve turnaround. In order to fix the multifarious nature of the financial problem, this edited book advocates interdisciplinary approaches based on machine learning.

Machine learning is involved in the analysis of large and multiple feature instances. It principally refers to acquiring knowledge and intelligence (by a computer program) from a processed training example for generating predictions. It deals with computationally intensive techniques, such as cluster analysis, dimensionality reduction, and support vector analysis. It is principally the area of computer science and is already frequently applied in social sciences, finance and banking, marketing research, operations research, and applied sciences. Moreover, computational finance is a domain of applied computer science that is concerned with practical issues in finance. It may be characterized as the study of features, instances, and learning algorithms applied in finance. It is an interdisciplinary area that integrates computational tools with numerical finance. Furthermore, computational finance applies arithmetical proofs that can be fitted to economic experiments, thereby contributing to the advancement of financial data modeling techniques and systems. These computational techniques are utilized in financial risk management, corporate bankruptcy prediction, stock price prediction, and portfolio management. Finally, this proposed textbook could play an important role in financial data learning.

Besides, this volume will be a basis for empirical and theoretical practices. The empirical experiments aim to minimize financial risk and uncertainty by covering and fitting the most advanced and novel machine learning algorithms. Moreover, it generates academic literature as well as financial product and finance modeling inferences toward customer credit risk assessment, data mining, pattern recognition, bankruptcy prediction, and so on. To be specific, the volume is broadly divided into three parts, with the first set of chapters focusing on the recent trend and issues of financial technology (FinTech). The second set of chapters comprises empirical essays on the prediction and forecasting financial risk by applying ML and DL tools and techniques. The third set of chapters combines empirical evidence of financial time-series data forecasting. The volume ends with a set of emerging technologies in financial education and healthcare and their empirical applications.

## Part 1: Recent Developments in FinTech

The first part presents four chapters on recent development in FinTech.

Chapter "FinTech Risk Management and Monitoring" focuses on risk management and monitoring in FinTech. The recent emergence of financial technology innovations in the financial services and some significant risks are investigated using the qualitative research method. Additionally, the appropriate way to mitigate the risk is discussed in this chapter. Besides this objective, this chapter discusses the major risk behind the rapid development of fintech and the steps for fintech risk management. The four key regulatory techniques that have important applications in FinTech management and monitoring are added, and, finally, the chapter summarizes the main challenges of FinTech risk management.

Chapter "Digital Transformation of Supply Chain with Supportive Culture in Blockchain Environment" explores the influence of blockchain on the digital transformation of Supply Chain Management (SCM). This chapter is also aimed to determine the importance of supportive culture in the adoption of blockchain in supply chains. The study findings indicate that the digitalization of supply chain management by adopting blockchain technology is positively correlated with organizational prosperity. The chapter also indicates that supportive culture is crucial to practicing blockchain technology. This study suggests that policymakers and stakeholders ensure a supportive culture to establish a traceable, efficient, and effective supply chain.

Chapter "Integration of Artificial Intelligence Technology in Management Accounting Information System: An Empirical Study" conducts an empirical study on the integration of artificial intelligence technology in management accounting information systems. This study established an artificial neural network-based model to predict management information and verify the accuracy of the model using some real data. Five dimensions are considered to develop the model, accounting analysis management system, accounting decision support system, performance management information system, risk management information system, and environment management information system.

The essentiality to analyze big data in accounting and finance is discussed in Chap. “The Impact of Big Data on Accounting Practices: Empirical Evidence from Africa”. Evidence indicates that big data significantly impact accounting and auditing accounting, utilizing the diversity of data volume, data variety, and data velocity. Chapter “The Impact of Big Data on Accounting Practices: Empirical Evidence from Africa” shows the impact of big data on accounting practices, and the study area is Africa. The main goal of this chapter is to explore the impacts of big data on accounting using accountants in Nigeria. Multiple regression is used for 151 responses, and samples are collected using the random sampling method. This study proves that big data positively and significantly affect financial reporting, performance measurement, corporate budgeting, audit evidence, risk management, and fraud management. This study helps accountants, prospective accountants, and accounting graduates in their studies.

## Part 2: Financial Risk Prediction Using Machine Learning

The second part contains four chapters that discuss the applications of ML and DL approaches to predict and forecast financial risk.

Chapter “Using Outlier Modification Rule for Improvement of the Performance of Classification Algorithms in the Case of Financial Data” discusses how to improve classifier performance by mining and modifying outliers of financial datasets. This chapter offers insights into the Financial Decision Support System for financial decision makers. This study employs four distinct classification algorithms such as linear discriminant analysis, $k$-nearest neighbor, naïve Bayes, and support vector machine for both original and modified datasets to detect credit card fraud. The study’s findings show that the classifiers perform better on modified datasets than on original credit card datasets.

Chapter “Default Risk Prediction Based on Support Vector Machine and Logit Support Vector Machine” is a predictive analysis of the machine learning algorithm for default risk prediction. This study proposes a LogitSVM model that hybridized the traditional support vector machine with popular logistic regression to assess the credit default risk. The authors use real-world credit databases to validate the probability and value of the proposed model. Type I error, type II error, and root mean square error (RMSE) are used to evaluate the performance of the regressors. Empirical findings show that the proposed hybrid model is superior to maximize accuracy and minimize RMSE. This chapter helps stockholders develop a wide variety of approaches to predict the credit customers’ default risk.

Chapter "Predicting Corporate Failure Using Ensemble Extreme Learning Machine" shows the corporate failure prediction using the Ensemble Extreme Learning Machine. The claim is that the early-stage prediction of corporate failure is essential for banks and financial institutions to solve financial decision-making problems. Newly developed artificial intelligence technique Extreme Learning Machine has an extremely fast learning classifier. To prove the superiority of this method, the authors compare the result with four benchmark ensemble methods, namely multiple classifiers, bagging, boosting, and random subspace. Experimental results on French firms indicated that bagged and boosted extreme learning machines showed the best-improved performance.

Chapter "Assessing and Predicting Small Enterprises' Credit Ratings: A Multicriteria Approach" focuses on small enterprises; it assigns and predicts the small enterprise's credit rating using a multicriteria approach. In reality, small enterprises have made it difficult for financial institutions such as commercial banks to accurately determine the credit risk, creating salient loan difficulties due to short time, high frequency, urgent demand for credit, and a small number of their loans. To solve this issue, the chapter develops a new approach for assessing credit risk in small enterprises by combining high-dimensional attribute reduction methods with fuzzy *C*-means to grade the credit ratings of enterprises requesting loans.

## Part 3: Financial Time-Series Forecasting

The third part contains two chapters that explore empirical evidence of time-series data modeling.

Chapter "An Ensemble LGBM (Light Gradient Boosting Machine) Approach for Crude Oil Price Prediction" is on the prediction of crude oil prices. Every second counts when governments, businesses, and individuals need to know what the future of the crude oil market will bring in terms of pricing. Estimating the future cost of crude oil is a crucial step toward building an economy that can last. In order to effectively predict future crude market prices, this research will use machine learning and ensemble learning techniques. The model using light gradient boosting (LGBM) is proposed by the authors to predict the price of crude oil. By analyzing and modeling the Brent time-series crude oil data, the accuracy and precision of our predictors can be improved. The LBGM forecast is compared to the lasso regression, random forest regression, and decision tree regression methods. The results achieved by the suggested model are quite similar to and better than those obtained by the baseline model when measured using RMSE, mean absolute percentage error (MAPE), mean squared error (MSE), and mean absolute error (MAE).

Chapter "Model Development for Predicting the Crude Oil Price: Comparative Evaluation of Ensemble and Machine Learning Methods" also shows the prediction of crude oil prices using different methods. This study shows a comparative study of ensemble algorithms and machine learning algorithms to find the best forecasting model. This research uses machine learning and an ensemble algorithm to forecast crude oil prices, and it compares the efficacy of three different regression models—AdaBoost, Bagging Lasso, and Support Vector Regression—to conclude which is the most suitable. Time-series data on crude oil prices are analyzed and used to validate the forecasting model. The results of the various algorithms are compared using an actual vs. anticipated curve. According to the results, the ensemble AdaBoost method has superior performance. The mean square error, mean absolute error, root mean square error, mean absolute percentage error, variance score, and R2 are used to verify the outcome. This research will help those with a stake in the crude oil industry decide and craft policies based on projected future prices.

## Part 4: Emerging Technologies in Financial Education and Healthcare

The fourth part contains three chapters that explore the financial education and healthcare issues and their emerging trends.

Chapter "Discovering the Role of M-Learning Among Finance Students: The Future of Online Education" investigates the role of m-learning among finance students and the future of online higher education. This study aims to find the hidden issues of m-learning in finance studies. This study is mainly a qualitative approach, and the findings show that digitalized education provides the opportunity for major finance students to access financial markets using the Internet and gain personal and professional knowledge in a better way rather than traditional learning. It also shows that m-learning has a significant positive relationship with the effectiveness of online education. This analysis has a significant implication for education policymakers and practitioners.

Chapter "Exploring the Role of Mobile Technologies in Higher Education: The Impact of Online Teaching on Traditional Learning" demonstrates how technological evolvements derive the conduction of higher education, especially mobile technology. This study also intended to detect the factors that attract pupils who do not adopt an online education system. A qualitative approach is used to determine the pros and cons of the technology-based education system in universities. The authors reveal that the adoption of mobile technologies in academic education enables students to access valuable resources free of cost and effortlessly, which in turn helps them to develop strong knowledge and understanding of their study contents. This study opens up a new arena for research scholars to discover the importance of online education systems.

Chapter "Knowledge Mining from Health Data: Application of Feature Selection Approaches" assessed the performance of feature selection techniques in knowledge mining of health datasets. This study compared seven popular knowledge mining approaches on six popular Affymetrix and cDNA datasets. Employing a support vector machine classifier, the study determined the knowledge miners' accuracy and area under the curve values. The finding of this chapter informs that the simple lasso knowledge mining algorithm performs well on Affymetrix datasets while random forest performs well on cDNA datasets. This chapter contributes to the existing literature by mentioning the state-of-the-art knowledge mining approaches in health informatics.

To conclude, this edited volume would provide both practical and managerial implications of financial and managerial decision support systems that capture a wide range of financial data traits. It would guide the execution of risk-adjusted financial product pricing systems, supplemented with a significant add up to the financial literacy of the investigated study. Furthermore, the book could show a roadmap to master's degree students and Ph.D. researchers for financial data analysis. In a wider sense, this specific volume targets an extensive audience, including academic and professional financial analysts. The contents of this book are expected to be useful to a wide audience involved in forecasting, modeling, trading, risk management, economics, credit risk, and portfolio management.

Middlesbrough, UK — Mohammad Zoynul Abedin
Pardubice, Czech Republic — Petr Hajek

# Contents

# Part I
# Recent Developments in FinTech

# FinTech Risk Management and Monitoring

**Morshadul Hasan and Ariful Hoque**

**Abstract** The recent emergence of financial technology innovations in the financial services industry also faces many challenges due to some significant risks. This chapter aims to identify specific fintech risks and appropriate ways to manage the risks. A qualitative research method is used to explore the objectives of this study. The findings of this study include the major risks behind the rapid development of fintech, and the fintech risk management steps. Also, this study identifies four key regulatory techniques that have important applications in managing and monitoring fintech risks. Finally, the findings summarize the main challenges of fintech risk management.

**Keywords** Financial technology · FinTech · Risk management · Risk monitoring

## 1 Introduction

In recent years, substantial development of financial technology (Fintech), such as artificial intelligence (AI), big data, machine learning (ML), cloud storage, blockchain, and other technologies, continues to promote the digital transformation of financial institutions (Deloitte, 2019; Hasan et al., 2020a; Wang et al., 2021). The application of financial products and tools is becoming more abundant, and the efficiency and inclusiveness of financial services have significantly improved. For example, the popularity of electronic payments, especially mobile payments, increases the coverage of basic financial services. The promotion and application of fintech have (i) increased the breadth, depth, and speed of financial services, (ii) brought benefits and convenience to users, (iii) helped financial institutions achieve quality and efficiency improvements, and (iv) improved the availability of financial services under the new crown epidemic (Hasan et al., 2020b). Given the

M. Hasan (✉) · A. Hoque
Murdoch Business School, Murdoch University, Perth, Australia
e-mail: mohammad.hasan@murdoch.edu.au; a.hoque@murdoch.edu.au

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_1

importance of Fintech, most of the positive effects of the rapid development of consumer finance in recent years can be attributed to fintech. Such improvements include enhancing the breadth and depth of encompassing financing and industry's overall efficiency (Hasan et al., 2022; Long, 2016). Traditional financial institutions have found new directions for financial service transformation. Fintech transformation can also play a role in reforming the future economic structure and improving efficiency. At the same time, fintech development carries significant downside risks. For example, the rapid growth of fintech also creates new problems as it solves the shortcomings of traditional financial services. These downside risks often make things very challenging for the policymakers to enable new opportunities and safeguard traditional weaknesses. Also, risks impact fintech companies' strategic goals. Thereby, managing the risks involved in fintech services is one of the essential jobs of fintech institutions. Fintech institutions usually measure, manage, and monitor fintech risks in different ways. The details of the fintech risk management and monitoring process are given in the following section of this chapter.

## 2 Definition of FinTech

The word Fintech is a synthesis of finance (Fin) and technology (Tech) (Hasan et al., 2020b, 2021). Fintech is a technology-oriented financial innovation that transforms or innovates financial products and business models using the results of modern science and technology to promote the quality and efficiency of financial services (Aggarwal, 2014; Gai et al., 2018; Gomber et al., 2017). Fintech refers to financial innovations provided by technologies, especially AI, Blockchain, big data analytics, cloud computing, and other means to redesign traditional financial products, processes, models, and organizational structures (Goldstein et al., 2019; Hasan et al., 2020a). Fintech services include digital payment, digital investment, crowdlending, crowdfunding, and online banking. The rise of financial technology on a global scale has significantly improved the service level. Also, the operating efficiency of banks has fundamentally changed the banking industry's original competitive environment. In response to the rapidly changing competitive environment, banks have already started their journey with financial technology. Banks can use mobile Internet, biometrics, big data, AI, and other technologies to broaden service channels, reduce manual services, improve financial institutions' full-process risk management and control capabilities, and reduce compliance and additional operating costs.

## 3 What Is FinTech Risk?

The rise of new Fintech firms also means some unknown challenges and risks must be addressed appropriately. Significant innovation poses challenges not only for financial institutions but also for regulators. Fintech risk is a threat that arises during consumer financial transactions and dealing through fintech technologies. In other words, the danger posed by technological innovations when using financial services. Also, fintech risk can be defined as any potential failures, shortcomings, and misuse of technology that disrupt consumers' financial dealings. Fintech risk includes many hidden risks and contagious issues that are discussed in the following sections. In addition, some of the risks and challenges are caused by the improper use of financial technology and some problems with financial technology itself.

## 4 Importance of Maintaining FinTech Securities

The upgrade risk supervision and the emergence of new technologies make a big difference between the future development of risk management and current well-known risk management capabilities. It is highly essential for financial institutions to reconsider and leverage emerging technologies to change their existing risk management methods to improve risk management quality and efficiency. Also, financial institutions need to consider financial technology risks to make the risk management approaches more dynamic and capable of responding quickly to new development trends. However, maintaining financial security is a fundamental strategic issue related to one country's overall economic and social development. And the accurate judgment of hidden risks is a prerequisite for ensuring financial security. Therefore, for building financial power, it is essential to pay attention to and maintain financial security while promoting financial innovation and strengthening the prevention of financial risks. In this aspect, using emerging risk management technologies is also important to improve the quality of risk management.

## 5 Risks Behind the Rapid Development of FinTech

This study thinks that technology is neutral; the key difference is how and who uses it. Due to the specifics of financial technology, financial institutions have to address aseries of new risks while dealing with financial services. Some significant fintech risks are discussed in the following sections. Barefoot (2020) classified fintech risk into different categories. These are loss of privacy, rising risks of fraud and scams, compromised data security, harmful manipulation of consumer behavior, uses of data that are non-transparent to both consumers and regulators, and discriminatory and unfair uses of data and data analytics. Also, Fintech companies entering financial or regulatory sectors lack sufficient knowledge, stability, and operational efficiency. Deloitte has also identified some of the most significant risks involved in financial services after using technology, such as strategic risk, cyber security risk, information technology (IT) vendor risk, IT resiliency and continuity risk, data management risk, response risk, third-party risk, technology operations risk, risk of ineffective risk management, and IT program execution risk.[1] Zhentao, (2021, July 28) added market risk, operational risk, liquidity risk, legal risk, regulatory risk, and credit risk. Risk may represent itself in various forms; however, this study points to fintech risks that need to be considered in fintech operations. These risks are cyber-attack, data privacy risk, data misuse and quality, technical risk, credit risk, market risk, liquidity risk, and regulatory risk. The stated risks are discussed in the following section.

### 5.1 Cyberattack

One of the most known risks for financial technology services is the threat of cyber-attacks, network intrusions, email phishing, malware, and other hazards (Alhayani et al., 2021; Khan et al., 2022b; Miao et al., 2022). Different malware and ransomware can easily corrupt data, disrupt and shut down computing processes, and cause significant financial and reputational damage (Ankita & Rani, 2021; Sharma et al., 2021).

### 5.2 Data Privacy Risk

Data privacy is one of the most critical concern for fintech industry. Data privacy risks primarily focuses on customer data theft, which is one of the most burning issues nowadays. Due to hacking of customer data, such as personal identity information, bank accounts, and card information, both fintech users and companies are continuously losing money. Due to the booming expansion of fintech companies, data privacy concerns are also booming.[2]

### 5.3 Data Misuse and Quality

Fintech services deal with millions of data every day. Thus, dealing with big data creates significant risks of data misuse and poor data quality (Clarke, 2016). Also, due to the absence of proper data regulatory standards, in some cases, fintech

[1] https://www2.deloitte.com/us/en/pages/center-for-board-effectiveness/articles/information-technology-risks-financial-services.html

[2] https://www.idx.us/knowledge-center/data-privacy-concerns-in-booming-fintech-industry

companies process poor quality data; thereby, the poor quality data raises important threat to the effective decision-making process (Barefoot, 2020). The misuse of data is also considered as a breach of data privacy that ultimately damages a financial institution's reputation regarding data privacy concerns and undermines institution's business interest.

### 5.4 Technical Risks

Since financial technology has not yet achieved effective breakthroughs in security technology, the technical deficiencies of fintech and its dependence on information system will reduce fintech's security performance and expand the scope of security challenges. Also, the application of new technology has not received the necessary risk assessment. As a result, some organizations blindly pursue the so-called subversive technologies without rigorous testing and risk assessment.

### 5.5 Operational Risk

The Fintech sector integrates the financial industry, technology companies, and market infrastructure operators. In this aspect, its' operation is complex compared to other sectors. Thereby, in any case of a high concentration of different industries, financial risks may also arise once a risk arises in any of the sectors.

### 5.6 Credit Risk

Online credit or loan is one of the most popular fintech services. The online loan business easily causes credit risks or default of borrowers (Bussmann et al., 2020; Santoso et al., 2020). Traditional financial institutions are exposed to the risks posed by financial technology companies. The cooperation between financial institutions and P2P online loans, third-party payments, and crowdfunding have been continuously strengthened. Any irregular cooperation, violations, and inadequate supervision can easily lead to cause a rise in credit risk. Also, there is a risk of a lack of borrowers' information compared to traditional banks (Bussmann et al., 2020).

### 5.7 Market Risk

Fintech has broken through the temporal and spatial barriers that exist between traditional financial institutions, financial institutions and non-financial institutions, and between economic entities. When a risk breaks out, it spreads faster and has a more significant impact on financial institutions. For example, commercial banks face unexpected changes due to continuous market transformation. This continuous market transformation also increases the risk of bankruptcy of the commercial bank (Yao & Song, 2021b). Also, the return from financial technology products is not stable, and the high-yield model that attracts investors is not sustainable in some cases. These market risks always impact the financial stability and performance of fintech service providers (Li, 2021; Yao & Song, 2021a).

## 5.8 Liquidity Risk

The cooperation of financial institutions with P2P lending, alternative financing, Internet wealth management, third-party money transfer services, and Internet banks can easily cause liquidity risks. The market failures cause systematic liquidity risk in the financial market infrastructure (Avgouleas & Kiayias, 2019). Liquidity risk may occur in different aspects, such as when in P2P online loans use high-interest rates, it creates unfair market competition. The unfair market competition also influences cash management of traditional banks. The unfair competition also influences banks' capital chain, thus also causing liquidity risks. Also, once the financial industry experiences major instability, it will cause large-scale difficulties in cashing out funds, which will initiate liquidity risks and interest rate risks (Lee & Shin, 2018). As a result, this kind of P2P and other online financial products with the characteristics of popularization and network externalities will lead to unpredictable losses for society.

## 5.9 Regulatory Risk

Fintech came into the market within a very short period of time with complicated business processess, and the industry legal system has not been established yet. Therefore, compliance or regulatory risks are more prominent in fintech services. For example, there are number of blind spots and loopholes in the existing laws, regulations, and supervision rules in the financial industry. The industry's lack of legal treatment and supervision basis leads to some illegal businesses. Institutions use legal loopholes to carry out criminal and unlawful activities, causing economic losses to financial institutions.

# 6 FinTech Risk Management, Monitoring, and Applications

Financial institutions must do a good job of monitoring and managing risk while providing financial transactions, product marketing, business handling, and after-sales service. Handling or managing risks effectively is an important factor in successful fintech services.[3] Considering the importance of effective FinTech risk management, monitoring, and applications, fintech institutions should focus on the following issues.

## 6.1 *FinTech Risks Management*

Usually, the industry should focus first on general risk management practices. Later, they can focus on specific actions or processes that will help manage fintech risks. This study focuses on the necessary steps of fintech risk management in the following section.

### 6.1.1 Identify and Categorize Fintech Risks

Risk management teams use different tools, such as AI, ML algorithms, and other technology, to identify fintech risks. Risk analysts should identify when, where, why, and how fintech risks can occur. Also, it needs to be recognized by both the internal and external parties involved in the risks. Besides, risk analysts should identify the parties who might be affected if any risk occurs.[4] Identifying risks is the basic ongoing risk management process.

### 6.1.2 Risks Measurement

Risk measurement refers to determining the probability of risk occurrence and the likely impact of such risks on the institution. After identifying major and influential risks, all the risks should be categorized and placed on a priority list to sort out which risks ranked first and need urgent solution. The responsible team should have a good understanding of financial data analytics techniques to identify and categorize risks. Risk measurement is one of the most important stages of analyzing risks with qualitative and quantitative tools (Alvarez-dionisi, 2020).

[3] Stoneburner, G., Goguen, A., & Feringa, A. (2002). Risk management guide for information technology systems. *Nist special publication*, *800*(30), 800–30.

[4] https://www.business.qld.gov.au/running-business/protecting-business/risk-management/preparing-plan/identify

### 6.1.3 Risk Mitigation Plan Focused on Anti-Fraud Methods and Technological Model

Financial institutions need to develop effective risk mitigation plans and procedures in the third stage. One of the most vital issues for financial institutions is to design effective anti-fraud methods (Fang et al., 2021) based on product characteristics to prevent application fraud, transaction fraud, and marketing fraud. Additionally, in order to track external risk situations such as emerging cybercrime or illicit property trends, financial institutions should be prepared with effective risk mitigation plans and respond on time when risks arise. Also, financial institutions need to specify and build their own technological model that will work to mitigate different risks.

### 6.1.4 Analysis and Mitigation

Before mitigating the risks, the risk management team analyzes the risks and their impacts (Ward, 1999). After analyzing the risks, the team will proceed to the risk mitigation stage. At this stage, the risk management team determines the probable solution to prevent or manage the risk and implements the technological models and other effective ways to mitigate the risks. The team should work with the top priorities and risks that would have the greatest impact compared to others. In some cases, the team implements immediate action to prevent the risks from occurring proactively.

### 6.1.5 Monitor and Supervision the Performance of Models

It is necessary to continuously monitor the risk of the external participant, including the risk monitoring of the participant itself and the abnormal behavior of the participant. Also, financial institutions must monitor the performance of models that were built to mitigate the risks. Fintech products often involve big data and AI models, and some models or algorithms have a problem during rapid execution. Therefore, continuous monitoring of the model performance is required, such as carrying out model verification in time to check functional efficiency to manage institutional risk.

## *6.2 Key Regulatory Technology and Applications*

The development of financial technology supervision is critical. More attention should be paid to the development of supervision technology in the regulatory process. There are a number of supervision technologies that have been widely used in the supervision of banking, securities, insurance, Internet finance, and other fields. Those regulatory technologies are expected to move towards the full-chain application of financial supervision. The industry calls for attention to the development of the following regulatory technologies to strict guard against unknown risks in the development of financial technology.

### 6.2.1 New Encryption Technology

The new encryption security technology is an emerging security tool that can effectively protect the privacy and ensure the data security of financial institution information. Kaspersky defined data encryption as *"Encryption in cyber security is the conversion of data from a readable format into an encoded format. Encrypted data can only be read or processed after it's been decrypted"*.[5] Even in large data sets, the new encryption technologies can map data objects to a common data platforms through access control, assisting the regulatory authorities in overcoming data security issues, and enabling data to be shared with the regulatory authorities.

### 6.2.2 Blockchain Technology

The powerful function of this technology is manifested in different aspects. It brings nearly real-time transaction data through smart monitoring (Masuda et al., 2020; Yang et al., 2022), which allows regulators to more accurately analyze systemic risks and improve the efficiency of on-site and off-site inspections. Also, the transparent design of blockchain can provide the supervisory authority with direct, instant and completely transparent, and trustworthy supervisory information (Khan et al., 2022a) and effectively enhance the supervisory authority's ability to deal with financial market emergencies.

### 6.2.3 Machine Learning Technology

Machine learning (ML) technologies provide different services, such as risk prediction, monitoring, and supervision (Abedin et al., 2021a, b; Jordan & Mitchell, 2015; Mantere et al., 2012). ML tools can use historical data to effectively identify possible fraud and can be used in the anti-money laundering field. It has a unique ability to stimulate language and text. Once a transaction deviates from compliance requirements is found, the system will automatically issue an early warning to financial institutions and regulatory agencies to monitor their transaction (Awoyemi et al., 2017; Goy et al., 2019; Sunny et al., 2022).

[5] https://www.kaspersky.com.au/resource-center/definitions/encryption

### 6.2.4 Big Data Technology

Big data technology can reorganize and analyze various types of data, obtain valuable information, and reveal the essential attributes of things. With the aid of effective analysis and discovery tools, big data allows regulators to briefly see what has been and is happening in the financial market. It can also accurately determine the probability of upcoming risks, which enhances the supervisor's ability to allocate supervisory resources dynamically (Khan et al., 2022c).

## *6.3 Main Applications of Regulatory Technology*

Blockchain, machine learning, big data, and other risk regulatory tools help the financial institution in different aspects, such as smart supervision, fraud detection and prevention, data management, transaction monitoring, and so on. The major applications of key regulatory technologies are discussed in the following sections.

### 6.3.1 Smart Supervision

Regulatory technology uses ML and cloud computing technology to enable the system to consciously track supervision, identify compliance requirements, provide targeted response solutions, manage compliance workflows, build data reporting platforms, open up different supervision reports, and other supervision activities. The Internet generates massive amounts of user data that are difficult to model manually every day. ML can solve the problem of slow manual model iteration. For the supervision of financial risks, the ML model can efficiently and quickly self-iterate by monitoring the characteristics and performance of the model, loan groups, and business feedback.

### 6.3.2 Fraud Prediction and Prevention

Big data helps to find clues to illegal activities based on data analysis. For online transactions, both senders and receivers of the transaction cannot visit physically. Therefore, this online connection opens room for the applicant for material fraud. In this case, big data technology can compare the information provided by the applicant with the authentic and accurate information that has been stored, discover the difference between the before and after dispatch information and provide evidence to prevent fraud and crack down on illegal and criminal activities in time. For example, big data tracks people's daily trajectories and accurately locates them based on geographic location. When the applicant's home address does not match the registered address or the information, such as the transaction address, is different from the stored information, the big data system automatically compares and issues an early warning.

### 6.3.3 Data Management

The establishment and use of big data technology, cloud computing, and other platforms are inseparable from data. Raw data is increasingly vital for the accuracy of risk prediction results. Data management covers using raw data to forecast all kinds of risk modeling, situation analysis and stress testing, scientific research and judgment on various financial risks, and formulating solutions. A high-quality database is needed to accomplish the above things. Therefore, data accuracy, completeness, and credibility significantly impact risk management and improve risk management performance. With the improvement of data quality requirements, the operating costs of risk databases also increase accordingly, which puts forward new requirements for the ability to select data.

### 6.3.4 Transaction Monitoring

Transaction monitoring is designed to detect unusual behavior that may indicate the occurrence of other financial crimes, such as terrorist financing and money laundering.[6] Real-time payment transaction monitoring has systemic problems, such as inaccurate data monitoring, which provides space for money laundering and other illegal activities. In this aspect, supervisory technology has the characteristics of intelligent, efficient, and automatic solution generation, which provides the possibility to discover system defects and eliminate illegal activities. Financial regulatory authorities use different applications in finance to improve regulatory efficiency and combat against financial crime. Those monitoring and managing applications prohibit financial market's false transactions and irregularities, and enhance risk management efficiency. Also, regulatory technologies guarantee the compliance and transparency of transactions and can improve transaction efficiency.

## 7 Challenges of FinTech Risk Management

Today's business environment is changing rapidly, and risks are also rapidly evolving. The financial industry also faces evolving challenges, such as continuous regulatory changes, growing awareness of third-party risk, lack of technology expertise, evolving data governance standards, increasing operational resilience demands, increasing cybersecurity threats, and other security and data privacy

[6]https://sanctionscanner.com/blog/biggest-transaction-monitoring-challenges-626

issues.[7] As a result, fintech firms face complicated risks and compliance challenges. For example, integrating big data and AI technologies is challenging to implement. It requires exceptional and high engineering skills and constant costly maintenance.

In some cases, technological integrations are changing and reshaping the operations of the financial industry.[8] It is evident that attempting to address these risks through manual techniques only increases risks, such as the inability to adapt to regulatory changes, poor data governance, and greater cyber risk. Instead, fintech organizations may consider taking a more strategic approach to successfully tackle these difficulties.

## 8 Conclusion

Risk in the fintech industry is a highly concerning issue at present time. Robust and very effective risk management techniques and strategies are highly demanding. A sound risk management system makes an organization more dynamic and responds quickly to emerging threats. This study is one of the first to explore hidden risks and appropriate risk management approaches in the FinTech industry. In addition, this paper discusses risk monitoring and oversight techniques and their applications to support the risk management processes. Overall, this research will have a significant implications on the risk management operations of fintech firms and make a substantial contribution to the fintech literature.

## References

Abedin, M. Z., Hassan, M. K., Hajek, P., & Uddin, M. M. (2021a). Machine learning in finance and accounting. In *The essentials of machine learning in finance and accounting* (1st edn). Springer.

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2021b). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research, 2021*, 1. https://doi.org/10.1007/s10479-021-04420-6

Aggarwal, D. V. K. (2014). Financial inclusion in India – Opinion. *International Journal of Commerce, Business and Management (IJCBM), 3*(6), 841–849.

Alhayani, B., Abbas, S. T., Khutar, D. Z., & Mohammed, H. J. (2021). Best ways computation intelligent of face cyber attacks. *Materials Today: Proceedings, xxxx*. https://doi.org/10.1016/j.matpr.2021.02.557

Alvarez-dionisi, L. E. (2020). A fintech risk assessment model. *ISSCA Journal, 3*, 1–6.

Ankita, A., & Rani, S. (2021). Machine learning and deep learning for malware and ransomware attacks in 6G network. In *Proceedings - 2021 4th international conference on computational intelligence and communication technologies, CCICT 2021*, pp. 39–44. https://doi.org/10.1109/CCICT53244.2021.00019

[7] https://www.protechtgroup.com/blog/top-5-risk-management-challenges-for-fintechs

[8] https://www.mobindustry.net/blog/7-key-challenges-fintech-startup-faces-and-their-solutions/

Avgouleas, E., & Kiayias, A. (2019). The promise of Blockchain Technology for global securities and derivatives markets: The new financial ecosystem and the 'holy grail' of systemic risk containment. *European Business Organization Law Review, 20*(1), 81–110. https://doi.org/10.1007/s40804-019-00133-3

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, January). Credit card fraud detection using machine learning techniques: A comparative analysis. In *Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017*, pp. 1–9. https://doi.org/10.1109/ICCNI.2017.8123782

Barefoot, J. A. (2020). Digital technology risks for finance: Dangers embedded in Fintech and Regtech. In *M-RCBG Associate Working Paper Series* (p. 151). https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp151

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence, 3*(April). https://doi.org/10.3389/frai.2020.00026

Clarke, R. (2016). Big data, big risks. *Information Systems Journal, 26*(1), 77–90. https://doi.org/10.1111/isj.12088

Deloitte. (2019). *Fintech risk and compliance management - A framework to empower the organization*.

Fang, W., Li, X., Zhou, P., Yan, J., Jiang, D., & Zhou, T. (2021). Deep learning anti-fraud model for internet loan: Where we are going. *IEEE Access, 9*, 9777–9784. https://doi.org/10.1109/ACCESS.2021.3051079

Gai, K., Qiu, M., & Sun, X. (2018). A survey on FinTech. *Journal of Network and Computer Applications, 103*, 262–273. https://doi.org/10.1016/j.jnca.2017.10.011

Goldstein, I., Jiang, W., & Karolyi, G. A. (2019). To FinTech and beyond. *Review of Financial Studies, 32*(5), 1647–1661. https://doi.org/10.1093/rfs/hhz025

Gomber, P., Koch, J.-A., & Siering, M. (2017). Digital finance and FinTech: Current research and future research directions. *Journal of Business Economics, 87*(5), 537–580. https://doi.org/10.1007/s11573-017-0852-x

Goy, G., Gezer, C., & Gungor, V. C. (2019, March). Credit card fraud detection with machine learning methods. In *UBMK 2019 - Proceedings, 4th International Conference on Computer Science and Engineering*, pp. 350–354. https://doi.org/10.1109/UBMK.2019.8906995

Hasan, M. M., Popp, J., & Oláh, J. (2020a). Current landscape and influence of big data on finance. *Journal of Big Data, 7*(1), 21. https://doi.org/10.1186/s40537-020-00291-z

Hasan, M. M., Yajuan, L., & Khan, S. (2020b). Promoting China's inclusive finance through digital financial services. *Global Business Review, 1–23*. https://doi.org/10.1177/097215091989534

Hasan, M., Le, T., & Hoque, A. (2021). How does financial literacy impact on inclusive finance? *Financial Innovation, 7*(1), 1. https://doi.org/10.1186/s40854-021-00259-9

Hasan, M., Noor, T., Gao, J., Usman, M., & Abedin, M. Z. (2022). Rural consumers' financial literacy and access to FinTech services. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-00936-9

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255.

Khan, K. M., Arshad, J., Iqbal, W., Abdullah, S., & Zaib, H. (2022a). Blockchain-enabled real-time SLA monitoring for cloud-hosted services. *Cluster Computing, 25*(1), 537–559. https://doi.org/10.1007/s10586-021-03416-y

Khan, M. N. R., Ara, J., Yesmin, S., & Abedin, M. Z. (2022b). Machine learning approaches in cybersecurity. *Data Intelligence and Cognitive Informatics*. https://doi.org/10.1007/978-981-16-6460-1_26

Khan, M. N. R., Tasnim, F., Yesmin, S., & Abedin, M. Z. (2022c). Review of city pricing system analysis based on big data. *Data Intelligence and Cognitive Informatics*. https://doi.org/10.1007/978-981-16-6460-1_25

Lee, I., & Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons, 61*(1), 35–46. https://doi.org/10.1016/j.bushor.2017.09.003

Li, C. (2021). Quantitative measurement and analysis of FinTech risk in China. *Economic Research-Ekonomska Istrazivanja, 35*, 2596. https://doi.org/10.1080/1331677X.2021.1970606

Long, C. (2016). From Fintech to Finlife: The case of Fintech development in China. *China Economic Journal, 9*(3), 225–239. https://doi.org/10.1080/17538963.2016.1215057

Mantere, M., Uusitalo, I., Sailio, M., & Noponen, S. (2012). Challenges of machine learning based monitoring for industrial control system networks. In *Proceedings - 26th IEEE International Conference on Advanced Information Networking and Applications Workshops, WAINA 2012*, pp. 968–972. https://doi.org/10.1109/WAINA.2012.135

Masuda, D., Shinkuma, R., Inagaki, Y., & Oki, E. (2020). Blockchain framework for real-time streaming data generated in image sensor networks for smart monitoring. In *2020 2nd Conference on Blockchain Research and Applications for Innovative Networks and Services, BRAINS 2020*, pp. 217–221. https://doi.org/10.1109/BRAINS49436.2020.9223311

Miao, Y., Chen, C., Pan, L., Han, Q. L., Zhang, J., & Xiang, Y. (2022). Machine learning based cyber attacks targeting on controlled information: A survey. *ACM Computing Surveys, 54*(7), 1. https://doi.org/10.1145/3465171

Santoso, W., Trinugroho, I., & Risfandy, T. (2020). What determine loan rate and default status in financial technology online direct lending? Evidence from Indonesia. *Emerging Markets Finance and Trade, 56*(2), 351–369. https://doi.org/10.1080/1540496X.2019.1605595

Sharma, S., Kumar, R., & Rama Krishna, C. (2021). A survey on analysis and detection of Android ransomware. *Concurrency and Computation: Practice and Experience, 33*(16), 1–24. https://doi.org/10.1002/cpe.6272

Sunny, F. A., Khan, M. I., Satu, M. S., & Abedin, M. Z. (2022). Investigating external audit records to detect fraudulent firms employing various machine learning methods. *Proceedings of the Seventh International Conference on Mathematics and Computing*. https://doi.org/10.1007/978-981-16-6890-6_38

Wang, R., Liu, J., & Luo, H. (2021). Fintech development and bank risk taking in China. *European Journal of Finance, 27*(4–5), 397–418. https://doi.org/10.1080/1351847X.2020.1805782

Ward, S. (1999). Assessing and managing important risks. *International Journal of Project Management, 17*(6), 331–336. https://doi.org/10.1016/S0263-7863(98)00051-9

Yang, F., Qiao, Y., Abedin, M. Z., & Huang, C. (2022). Privacy-preserved credit data sharing integrating Blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/TII.2022.3151917

Yao, T., & Song, L. (2021a). Examining the differences in the impact of Fintech on the economic capital of commercial banks' market risk: Evidence from a panel system GMM analysis. *Applied Economics, 53*(23), 2647–2660. https://doi.org/10.1080/00036846.2020.1864275

Yao, T., & Song, L. (2021b). Fintech and the economic capital of Chinese commercial bank's risk: Based on theory and evidence. *International Journal of Finance and Economics*. https://doi.org/10.1002/ijfe.2528

Zhentao, Y. (2021, July 28). *Risks and countermeasures brought about by the rapid development of financial technology*. Qindianstar

# Digital Transformation of Supply Chain with Supportive Culture in Blockchain Environment

**Shakila Akter, Mohammad Samiul Haque, Ashrafuzzaman Sohag, Md. Jahangir Alam Siddikee, and Mohammad Zoynul Abedin**

**Abstract** This study aims to evaluate the effect and usefulness of digital transformation of supply chain management (SCM) on blockchain with a supportive culture. This paper explores the effect of blockchain on SCM under consideration of automated controls with smart contracts, fundamental attributes, cooperation, supportive culture, transparency and identification, and trust building. Here, this study finds that the supportive culture has great potential to boost the transformation of SCM rapidly and successfully. Blockchain technology has the potential to transmit the supply chain. Finally, this current study indicates that the transformation of SCM in blockchain with supportive culture has a positive impact on the success of organizations. Therefore, this study inspires policymakers and stakeholders to ensure a supportive environment to build a robust sustainable supply chain that will be traceable, more effective, and efficient.

**Keywords** Digital transformation · Supportive culture · Supply chain management · Blockchain

## 1 Introduction

Technological or digital transformation is one of the trends that shape the business world and changes in the work environment. To cope with technological transformations and utilize opportunities that arise from digital technologies, the SCM of the company faces numerous pressures, such as lack of supportive culture, industry-specific guidelines, digital skills, etc. (Agrawal et al., 2020). Digital transformation

S. Akter · M. S. Haque · A. Sohag · M. J. A. Siddikee
Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

M. Z. Abedin (✉)
Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_2

(DT) is known as a way of developing a new business model that helps organizations generate relatively greater value (Verhoef et al., 2019). This transformation has an impact on firm schedules, capabilities, and business procedures (Da Xu et al., 2018). DT supports firms in offering better products and services by eliminating the obstacles between final users, businesses, and objects. A supportive culture is essential for every business to enhance and share knowledge, learning, resources, and skills (Bollinger et al., 2002). Organizational culture and environmental sustainability play the driving role in adopting the digital transformation of businesses by bringing out a continuous change in their structure (Isensee et al., 2020; Khan et al., 2022). Supportive culture ensures the situation in which human resources can build a supportive correlation between them, organizational culture, environment, and their working conditions (Karine, 2020).

At present, blockchain technology is practiced in several industries including the finance and accounting industry such as capital markets, international trade, corporate governance, banking, and taxation (Farhana et al., 2022). Blockchain technology accelerates consumer confidence by operating transactions more efficiently, traceably, safely, and transparently (Aste & Matteo, 2017; Kshetri, 2018; Queiroz & Fosso, 2019). Blockchain technology (BCT) represents an appropriately circulated public ledger that covers details about each type of data transaction among network participants (Singh & Kim, 2018; Yang et al., 2022). Traditional supply chain management (SCM) has to face a number of problems such as product tampering, fraud, and, delay, etc. (Petr & Abedin, 2020; Abedin et al., 2020). BCT has the potential to eliminate the aforementioned difficulties through its significant available features, such as anonymity, decentralization, stability, traceability, and transparency (Ali et al., 2020). The adoption of blockchain in SCM helps increase the cooperation between supply chain members, efficiency in the supply chain process, and reduce overall cost. To detect and prevent products fraud, blockchain traceability activities have a significant influence on SCM (Chen, 2018; Sana et al., 2022). Blockchain has the ability to solve composite issues such as accountability and transparency (Kshetri, 2018). Therefore, in the perception of SCM, blockchain is considered as an identical technology (Choi et al., 2020).

Nowadays, digital technology has completely updated how people interact with their surroundings. Individuals use smartphones, smart watches, personal computers, advanced television units, wearable devices, drones, and self-driving cars to access and transfer data that are the reflection of digitalization (Prasitlumkum et al., 2020). These technological innovations have a major impact on each sector, including the supply chain sectors (Abedin et al., 2021). A supply chain is a unified system of organizations, people, and information that involves planning, organizing, controlling, and coordinating the transfer of products and services from the provider to the consumer (Azzi et al., 2019; Shajalal et al., 2021). Digital technology affects every phase of human life as well as the supply chain process (Nasiri et al., 2020). Companies are increasingly aware of these potential developments and strengthen how the digital supply chain (DSC) can add value to them. DSC is a series of interrelated actions that are driven by new technology and involved in the supply chain process (Büyüközkan & Göçer, 2018). DSC can create new forms of revenue and business value for companies by using various innovative technologies such as drones, cloud computing, bar code readers, QR codes, and unmanned aerial vehicles (Bicocchi et al., 2019).

Adoption of DT faces plenty of difficulties, namely lack of vision, insufficient leadership knowledge and skills, financial inadequacy, and lack of a supportive organizational culture (Papagiannidis et al., 2020). Therefore, this empirical study investigates the role of a supportive culture in the adoption of digital transformation, especially blockchain technology in conducting supply chain activities. This study tends to detect the potential impact of the blockchain environment and digital transformation with supportive culture on SCM issues: traceability, transparency, security, and efficiency play. This study contributes to the existing literature on digital supply chain management and organizational supportive culture. This study extends the existing domains by identifying the effect of supportive culture in SCM considering digitalization. This paper suggests that stakeholders consider the organizational internal and external environment while adopting new technology to carry out SCM activities.

## 2 Literature Review

By employing a theoretical framework on archival data from case studies, Kshetri (2018) explores that blockchain impacts on SCM objectives like quality, reliability, cost, sustainability, risk minimization, and flexibility. Wang et al. (2019) seek to identify how BCT changes the traditional supply chain practices. For this purpose, their study employs narrative analysis and cognitive mapping. Applying transaction cost theory, the study of Schmidt and Wagner (2019) establishes a preliminary idea of how blockchain affects supply chain relations. In this regard, they consider authority decisions and operation costs. Saurabh and Dey (2020) utilize the conjoint analysis (CA), by developing the theoretical framework, to identify the influential factors that affect the BCT in the grape wine supply chain. To detect the financial and operational advantages of adapting blockchain technology rather than a traditional platform, Giovanni (2020) applies a simple supply chain (SC) model. By combining the Fuzzy Delphy and Best-Worst method (BWM), Ghasemian et al. (2020) generate an integrated method to determine the barriers to blockchain adoption in a humanitarian supply chain management.

The study by Dowty and Wallace (2010) detected the role of organizational culture in disrupting and restoring the supply chain. In the study by Li et al. (2016), they explore the organizational pressure to take on Internet-enabled SCM from the perspective of organizational culture. Conducting survey data from 131 Chinese service and manufacturing firms, their study develops a conceptual framework and hypothesis test. By using the mediating effect of structural equation modeling (SEM), Liou et al. (2012) analyze the institutional commitment in relation to organizational supportive culture and employee job satisfaction. They collect primary data from 210 samples of Taiwanese universities. Lin (2013) identifies the factors for adopting an electronic supply chain management system (e-SCM) from an organizational, environmental, and technological perspective using logistic regression. Their survey collects data from 283 managers from Taiwanese firms. Conducting questionnaire-based data from 418 graduates from Dutch Business School (DBS), the Netherlands, Sok et al. (2014) explore the relationship between work-to-home and organizational culture spillover. Their study utilizes structural equation modeling and confirmatory factor analysis (CFA).

Jabbar et al. (2020) describe the digital transformation of sustainable supply chain management (SSCM) as big data analytics. They applied a systematic literature review (SLR) method. Their study considers the Scopus database as article searches by title, abstract, and keyword. Nasiri et al. (2020) inspect the mediating effect of smart technologies. In their study, they consider 280 Finnish small and medium-sized enterprises to show how the organization's digital transformation affects the relationship performance from the supply chain perspective. Song et al. (2021) seek to clarify the various e-commerce methods of the wholesale market that can update and transform its ecosystem by implementing Information and Communication Technology (ICT). In their study, they collect 24 interviews as primary data, market records, papers, internal reports, as well as different published documents as secondary data from a theoretical point of view. Büyüközkan and Göçer (2018), take into account the Analytic Hierarchy Process (AHP), Additive Ratio Assessment (ARAS), and Interval Valued Intuitionistic Fuzzy (IVIF) sets under the Group Decision Making (GDM) method. Their study initiates a DSC procedure for the selective activities of suppliers.

On the basis of the literature mentioned above, this study determines the following research gap. There are a range of studies dealing with the relationship between blockchain and supply chain, supportive culture and DT, and DT relations with SCM, respectively. That means the existing literature covers the interconnections between corresponding issues, but they cannot reflect the impact of a supportive culture in adopting digital transformation in SCM.

To cover the aforementioned research gap, this study aims to consider the factors simultaneously. This paper sheds a new light on the importance of a supportive culture in the face of digital transformation (BCT) in managing supply chain.

## 3 Methodology of the Study

The method is a description of the manner in which data are collected, analyzed, and interpreted. This study proposes a systematic literature review (SLR) of academic and practitioner literature on the Digital Transformation of Supply Chain and Blockchain technology. This current study conducts several steps of analysis to include a set of articles in the review. First, for the purpose of this paper, we searched the top academic journal databases. Accepted articles include the following keywords blockchain, supply chain transformation, and organizational culture. Second, this study considers the science citation index (SCI), the social science citation index

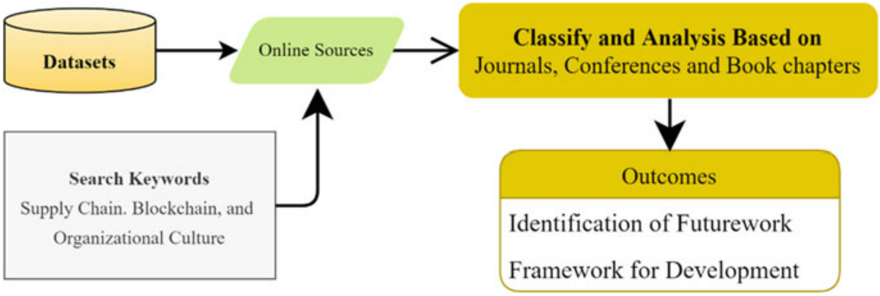



**Fig. 1** Framework of data collection and processing

(SSCI), and social science citation expanded (SSCE) for papers related to blockchain, supply chain transformation, and organizational supportive culture. The time frame of the paper is the data during the 1991–2020 years. But the maximum data is targeted for the past seven years (2013–2020).

Finally, we examine 87 articles including journal article, article in a periodical, conference proceedings, book chapters, and reports (Fig. 1).

# 4 Analysis and Interpretation

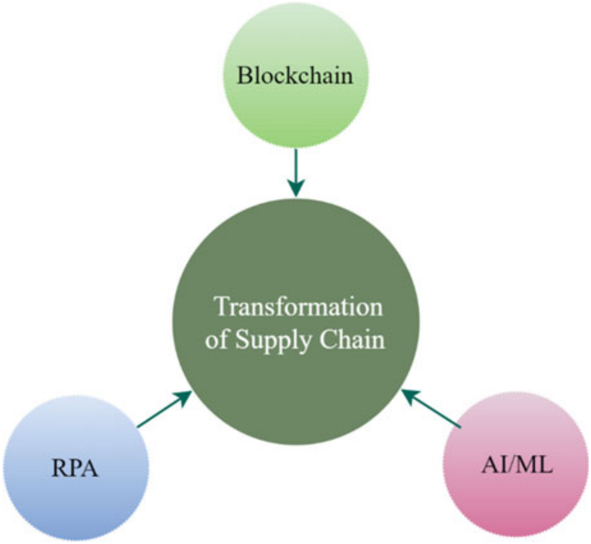
## *4.1 Digital Transformation of Supply Chain*

Supply chain experts think about how to go forward due to the rapid prosperity of modern technology. The supply chain process changes a lot with the touch of digital transformation. To accept these major changes, companies must identify the prospects and barriers made by digital technology. DSC allows firms to recognize the customer's needs, the supplier's challenges, and continue their operation efficiently. Machine learning (ML)/Artificial intelligence (AI), blockchain, and Robotic process automation (RPA) are considered technological assistance to make the firm digitalize (Hartley & Sawaya, 2019); see Fig. 2.

**Robotic Process Automation**

Robotic Process Automation (RPA) can be defined as a developing technology that ensures the utilization of software bots to enable the firm to automate tasks and rule-based business procedures (Kokina & Blanchette, 2019). Evidence has shown that about 60% of supply chain experts apply RPA to automate supply chain processes (APQC, 2018). In DT practice, RPA considers the organizational first phase. Data designs, entry, evaluation, and mining from the Enterprise Resource Planning (ERP) structure are the main activities associated with RPA (Huang & Vasarhelyi, 2019). The supply chain conducts various monotonous tasks automatically, such as collections, operations, and logistics through RPA (Hartley & Sawaya, 2019). Organizations move forward with digital transformation with RPA for different motives. First,

**Fig. 2** Supply chain transformation




setting up RPA application is comparatively easy by applying the software bots. Second, it can be applied faster than a completely reformed process from end to end. In conclusion, on the basis of business requirements, there are options to improve or eliminate the capacity.

**Artificial Intelligence/Machine Learning**
Artificial Intelligence (AI) refers to the potentiality to contribute to engineering and scientific assignments by replicating, broadening, and converting human expressions in an effective and accurate manner (Muthukrishnan et al., 2020). Nowadays, there are many AI applications in the supply chain and the possibilities of AI applications are endless. This study considers machine learning (ML) as a subsection of AI for supply chain operation procedures. ML contains algorithms that can learn compound operations and develop analytical models from test data (Carbonneau et al., 2008). Supply chain applications of ML include scheduling of warehouse pick processes, demand planning, and forecasting, governing the equipment nurturing plans, examining information to advance the transportation supervision, etc. (Toorajipour et al., 2021).

**Blockchain**
Blockchain is an independent digitally Distributed Ledger Technology (DLT) (Di et al., 2020), holding random information, which is not supervised by a sole or a corporation of entities; anybody can access this platform easily (Lafourcade & Lombard-platet, 2020). Although BCT applications were first introduced in the cryptocurrency perspective (Nakamoto, 2008), currently, this technology uses various contexts such as SCM (Karamchandani et al., 2019), health care record management (Shi et al., 2020), electronic voting (Nam et al., 2021), the insurance industry (Kar & Navin, 2021), and so on. Generally, BCT platforms are more secure.

Permitted users have access to include or view particular data. Blockchain adds positive value in the SCM area in a different way such as product traceability, SCM transmission, inventory supervision, and customer affiliation (Jabbour et al., 2020). Daily operations are automated using smart contracts through blockchain (Xuan et al., 2020).

## *4.2 Digital Transformation of Supply Chain in Supportive Culture*

Today's world is changing a lot by technological innovation. That is why the online-based or automated business has taken place rapidly rather than a traditional business. Organizations implement different modern technologies for different reasons, such as meeting customer demand, competitive pressure, and the wide acceptance of technology. The digital transformation of supply chains changes the organizational operation procedure, model, plans, and culture. Digitalization updates current cultures or creates new ones and uses structures, symbols, and digital art around the business (Bounfour, 2016).

Based on previous studies, several factors have an impact on supply chain transformation. Employee engagement and acceptance are considered the most crucial factors to support the transformation progression (Michela & Burke, 2000). The supportive culture ensures a collaborative and human-aligned, friendly, motivating, and trustful workplace (Dowty & Wallace, 2010), and it minimizes the likelihood of negative working experiences for employees by increasing job satisfaction (Liou et al., 2012). A leader is one who supports and understands the feelings of others. To successfully implement digital technology, leaders play a major role (Banks et al., 2019).

Organizational culture refers to the ways in which norms, beliefs, values, and communications help establish an organization's emotional and exceptional social environment (Wu, 2008). Although cultural change is too challenging, any kind of organizational change culture is crucial. To implement digital transformation, it is necessary to change strategy, leadership, and organizational culture (Halpern et al., 2021). Sometimes the chief executive manager and other higher authorities allow the change. Therefore, the transformation depends greatly on the entire staff support of the organization. Combining culture and technology is not an easy job, as both concepts interact with the organizations' subsystems. To adjust the culture in the digital transformation of the supply chain, a supportive approach is needed. Cabrera (2001) concludes that to introduce the technological transformation organizational culture should be considered. Organizational culture positively considers the environment so far, and it also assists the changes (Gordon, 1991). When culture is ignored and supportive approaches are lacking, the digital transformation of the supply chain will fail. Organizations face the challenge of taking a step on digital transformation when they fail to encourage their employees and managers

(Garcia-lorenzo, 2020). So a supportive culture should be maintained or changes should be made if it is required in the transformation of supply chains.

## 4.3 Blockchain and Supply Chain Management

Nowadays, different supply chain issues are solved through the adoption of blockchain, such as smart contracts, traceability, product fraud detection, and trust building (Howson, 2020; Giovanni, 2020; Sunny et al., 2020). Figure 3 clarifies it more specifically. Blockchain has an impact on the traditional supply chain. In this context, blockchain on SCM is explained in the following section.

**Traditional Supply Chain with Blockchain Technology**
Traditional SCM has some common strategic objectives (B. Wang et al., 2020). BCT provides essential assistance to accomplish these objectives efficiently and effectively (Kshetri, 2018).

- Cost reduction: Transaction made through BTC minimize the cost by creating an exclusive code for all transactions. This helps to thoroughly examine the flow of funds throughout the supply chain discipline process.
- Operational speed: BTC can speed up processing by reducing physical interconnection and transmission.
- Sustainability: BTC can support developing meaningful and computable performance metrics to achieve environmental, economic, and social sustainability.
- Risk management: Transactions can only be made when relevant parties agree to transactions by negotiating among themselves within the blockchain network. This process supports controlling the data risk of all supply chain transactions through BCT.

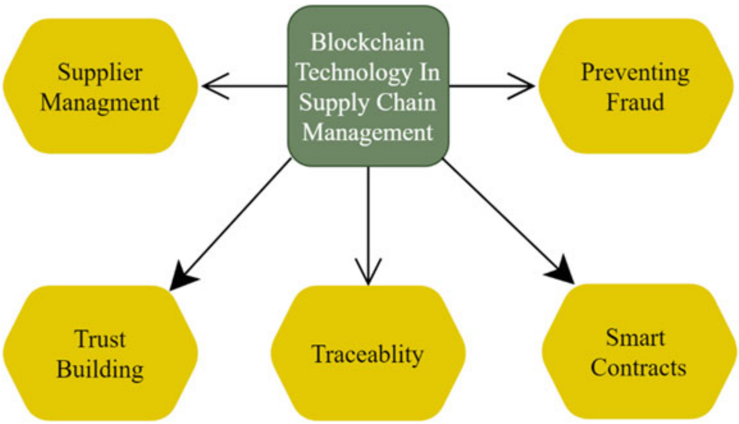



**Fig. 3** Implementing Blockchain in Supply Chain Management

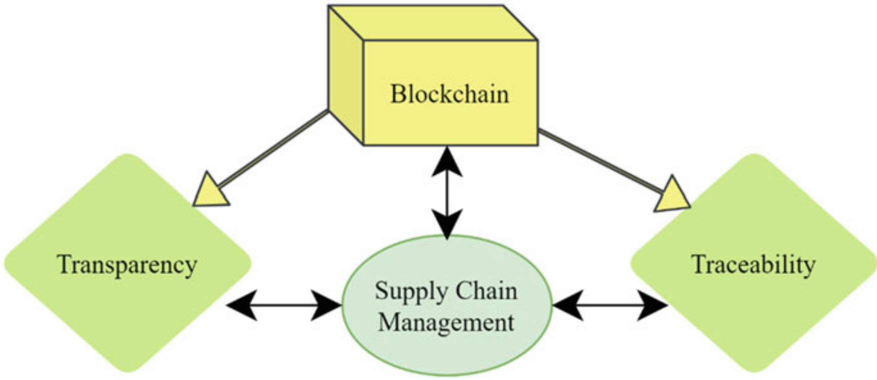



**Fig. 4** Blockchain technology and Supply Chain Management

- Flexibility: BCT can assist customers to locate and track orders from upstream to downstream, allowing customers to easily change, and also the suppliers to adapt to instant changes.**Basic Characteristics of Blockchain**

The characteristics of BCT are explained in this part. BCT establishes visibility, confidence, order, lucidity, and computerization in a disordered environment (Viriyasitavat & Hoonsopon, 2018). Blockchain ensures better visibility and security than traditional supply chain processes. BCT stores specific information on each component and provides it to the individual producer in the manufacturing operation both upstream and downstream (Leary, 2017). Blockchain can be used as an alternative to improving and replacing paper tracing, speeding up data sharing (Brent et al., 2013). These data sharing facilities of BCT strengthen the total capacity to manage the supply chain activities.

Additionally, blockchain keeps a record of business information in a permanent, verifiable, and safe form and keeps track of ownership. That helps the organization minimize the risk of cybercrime, fraud, and hacking. BTC builds hope among participants by committing that each record is noted and saved in numerous locations beyond the whole distributed network. It also increases the skills of supply chain and reduces the difficulty of the system. BTC allows manufacturers and resellers to gain insight into consumer needs and tailor their products and services in view of that (Adams et al., 2017).

**Transparency/Visibility and Traceability**

In a blockchain environment, traceability is defined as the ability to trace and track data (Sunny et al., 2020). Uses of traceability in the supply chain enhance transparency. Although traceability and transparency are two interconnected features of BCT (Wang et al., 2018). The visibility of the supply chain depends a lot on transparency (Hernandez, 2003). Blockchain ensures better transparency by providing all details regarding transactions among all parties involved in the supply chain process (Yasin et al., 2019). Blockchain has a great impact on SCM in traceability and transparency dimensions (Fig. 4).

"*Transparency of supply chain is the area in which all its stakeholders have a shared understanding of, as well as access to, the product-related information that they desire, without delay, noise, loss, and distortion*" (Holland et al., 2017). Product tracking continues from stats to end, whereas tracing generally towards the origin from the endpoint. Customers easily gather information about the material, source, and environmental impact of the product. Manufacturers and distributors benefited by providing new information to the customer and better product tracking.

There are some main areas in transparency/visibility and traceability. The following are quoted:

- Track the origin of the product.
- Fraud prevention beyond the supply chain network.

- Ensure data security.

**Security**

The blockchain uses public keys to enhance security and prevent maliciously. The supply chains of dangerous products should be handled in a very secure manner. Transforming dangerous goods requires advanced care (Berdik et al., 2021). All stakeholders involved in the process of hazardous products find the appropriate information through BCT. Producers make smart contact to transport products with initial information. All parties involved, including the administrative body, can access this information (Thakur & Breslin, 2020). In this way, the blockchain creates security through transparency in the supply chain. BCT is built with secure, "blocks" that store copies of the documents and are oriented to the previous blocks. This makes them secure and challenging to falsify (Bhushan et al., 2020).

**Smart Contracts**

Since blockchain is viewed as a more inherently secure form of technology, there is still a vital role to play for automation. Smart contracts are defined as self-operating and enhancement applications that use software code and a computing framework to activate a specific contract or terms of agreement (Hewa et al., 2020). Smart contract considers as a complement the use of Distribution Ledger Technology (DLT) and a decentralized program in the BC network (Han et al., 2020). It can be executed autonomously in predetermined contexts. The main function of smart contracts is to implement a peer-to-peer approach without central third-party involvement (Hu et al., 2021). There is no central dependence on the availability of services in this system.

## 5 Findings

This empirical study helps enrich the extant literature on SCM, BCT, supportive culture, and DT. The present study improves the understanding of how supportive culture affects supply chain performance in digital transformation. In order to improve SCM performance in numerous aspects, supportive culture and blockchain with smart controls play a vital role is identified in this paper.

Those aspects are quoted below:

- Enhancing transparency and traceability helps build a better relationship.
- Reducing the bullwhip effect by providing symmetric information among partners.
- Detecting fraudulent entries helps to prevent fraud.
- Using smart contracts helps reduce transaction cost and save time.
- By developing a better relationship, providing effective information and preventing fraud, it creates trust and collaboration among partners.

## 6 Discussion

Wang et al. (2019) conclude some probable benefits to implementing blockchain in the supply chain sector, such as increased operational efficiency and supply chain transparency, building mutual trust, and sharing reliable information. The finding of Sahebi et al. (2020) indicates that lack of knowledge, cost of employee training, and vagueness of regulations are the most significant barriers to adopting blockchain. Schmidt and Wagner (2019) concluded that blockchain minimizes operating and governance cost by automating buyer and supplier contracts and a permanent ledger of records. The results of Saurabh and Dey (2020) study noted that traceability, price, consent, faith, dis-intermediation, control, and coordination are the influential supply chain actors for implementing BCT.

Liu et al. (2010) found that the organizational culture has diverse effects on the dimensions of institutional pressures and inter-organizational technological adoption intention. Sok et al. (2014) find that a favorable culture explains the majority of variance in positive work-to-home meddling and strain-based negative work-to-home meddling. Blockchain, the internet of things, and AI have the potential to enrich transparency, faith, and provide substantial assistance by changing national and organizational culture (Kimani et al., 2020).

Lin (2013) shows that the implementation of e-SCM relies on higher authority support, absorptive capacity, and competitive pressure. Kshetri (2018) claims that the supply chain sector is one of the most likely sectors to be transformed into blockchain. The interconnection between relationship performance and digital transformation is fully mediated by smart technologies (Nasiri et al., 2020). Jabbar et al. (2020) imply that applying big data is good for every phase of the triple bottom line in the supply chain. Song et al. (2021) conclude that the introduction of ICT can be both a warning and an avenue for the wholesale market. Furthermore, marketing channels and transaction expenses can reduce the attraction of physical wholesale markets to customers and wholesalers.

# 7 Conclusion, Theoretical Contribution, Policy Implications, and Future Work

## 7.1 Conclusion

Today's world is changing a lot by technological innovation. That is why the online-based or automated business has taken place rapidly instead of a traditional business. Organizations implement different modern technologies for different reasons, such as meeting customer demand, competitor pressure, and the wide acceptance of technology. Typically, a supportive culture seeks to use the flexibility of the operating system to link up the needs of employees, maintain interpersonal relationships, and care for people, thus representing and defending its fundamental beliefs (Sok et al., 2014). For any kind of organizational change, organizational culture is crucial. To implement digital transformation, it is necessary to change strategy, leadership, and organizational culture. The day-by-day organizational culture becomes the basis of digital transformation in the organization.

DT and analytical methods and novel tactics including DSC can illustrate how to use different innovative technologies (IoT, cloud computing) to manage supply chain processes. Blockchain technology is an indicator of digital transformation. In reducing cost and increasing supply chain performance, BCT plays the driving role. Most importantly, practicing BCT is more secure, so that only allowed users can get access the information. That indicates that in facilitating the performance of SCM, adoption of digital technology more specifically, BCT is important.

## 7.2 Theoretical Contribution

This study has an important contribution to supply chain management and organizational supportive culture domains. This paper determines how supportive culture impacts the adoption of modern innovations such as blockchain technology in SCM.

## 7.3 Policy Implications

It appears that the findings should have important implications. Supportive culture is essential for effective transformation. This paper suggests that stakeholders, policymakers, and supply chain managers consider organizational culture while adopting innovative technology. For this reason, the organization has to gain a deep understanding of cultural complexities and transformation barriers. If an organization improves its understanding of the relationship among supportive cultural effects, blockchain adoption and the performance of the supply chain will play an important role in various fields.

## 7.4 Future Work

However, blockchain technology in SCM is currently in its early stages, and further studies are needed to extend the present study. Although BCT is becoming a more widely accepted and recognized topic, there are still many ideas that require future exploration and analysis. Which can be developed through further research that are quoted below:

- The relationship among supportive culture, blockchain, and supply chain performance in various areas.
- Future investigation is required to develop trust among parties involved in the supply chain through BCT.
- The blocks in the area of transformation of SCM in blockchain.
- Identify how cultural elements affect supply chain activities to adapt with new technological changes.
- The operation of smart contracts in SCM should be addressed more in future work.

This work informs academicians that in the near future, the application of blockchain in supply chain management will be a new avenue for investigation. It will be sensational to see what happens over the next decade.

## References

Abedin, M. Z., Chi, G., Uddin, M. M., Shahriare, S., Imran, K., & Petr, H. (2020). Tax default prediction using feature transformation-based machine learning. *IEEE Access, 9*, 19864–19881. https://doi.org/10.1109/ACCESS.2020.3048018

Abedin, M. Z., Hassan, M. K., Hajek, P., & Uddin, M. M. (Eds.). (2021). *The essentials of machine learning in finance and accounting*. Routledge.

Adams, R., Parry, G., Godsiff, P., & Ward, P. (2017). The future of money and further applications of the blockchain. *Strategic Change, 26*(5), 417–422. https://doi.org/10.1002/jsc.2141

Agrawal, P., Narain, R., & Ullah, I. (2020). Analysis of barriers in implementation of digital transformation of supply chain using interpretive structural modelling approach. *Journal of Modelling in Management, 15*(1), 297–317. https://doi.org/10.1108/JM2-03-2019-0066

Ali, O., Ally, M., & Dwivedi, Y. (2020). The state of play of blockchain technology in the financial services sector: A systematic literature review. *International Journal of Information Management, 54*, 102199. https://doi.org/10.1016/j.ijinfomgt.2020.102199

APQC. (2018). Quick poll: Current state of robotic process automation in the supply chain. *Supply Chain Management Review*. Accessed from https://www.scmr.com/article/quick_poll_current_state_of_robotic_process_automation_in_the_supply_chain

Aste, T., & Matteo, T. D. (2017). Blockchain technologies: The foreseeable impact on society and industry. *Computer, 50*(9), 18–28. https://doi.org/10.1109/MC.2017.3571064

Azzi, R., Kilany, R., & Sokhn, M. (2019). The power of a blockchain-based supply chain. *Computers & Industrial Engineering, 135*, 582–592. https://doi.org/10.1016/j.cie.2019.06.042

Banks, G. C., Dionne, S. D., Sayama, H., Schmid, M., & Special, M. (2019). Leadership in the digital era: Social media, big data, virtual reality, computational methods, and deep learning. *The Leadership Quarterly, 30*(3), 3–4. https://doi.org/10.1016/S1048-9843(19)30294-2

Berdik, D., Otoum, S., Schmidt, N., Porter, D., & Jararweh, Y. (2021). A survey on Blockchain for information systems management and security. *Information Processing and Management, 58*(1), 102397. https://doi.org/10.1016/j.ipm.2020.102397

Bhushan, B., Sinha, P., Sagayam, K. M., & Andrew, J. (2020). Untangling blockchain technology: A survey on state of the art, security threats, privacy services, applications and future research directions. *Computers and Electrical Engineering, 90*, 106897. https://doi.org/10.1016/j.compeleceng.2020.106897

Bicocchi, N., Cabri, G., Mandreoli, F., & Mecella, M. (2019). Dynamic digital factories for agile supply chains: An architectural approach. *Journal of Industrial Information Integration, 15*, 111–121. https://doi.org/10.1016/j.jii.2019.02.001

Bollinger, A. S., Smith, R. D., Bollinger, A. S., & Smith, R. D. (2002). Managing organizational knowledge as a strategic asset. *Journal of Knowledge Management, 5*(1), 8–18. https://doi.org/10.1108/13673270110384365

Bounfour, A. (2016). *Digital futures, digital transformation: From lean production to acceluction*. Springer.

Brent, A., Joseph, D. W., Travis, R., & Morgan, T. (2013). Leveraging supply chain visibility for responsiveness: The moderating role of internal integration. *Journal of Operations Management, 31*(7–8), 543–554. https://doi.org/10.1016/j.jom.2013.09.003

Büyüközkan, G., & Göçer, F. (2018). Computers in industry digital supply chain: Literature review and a proposed framework for future research. *Computers in Industry, 97*, 157–177. https://doi.org/10.1016/j.compind.2018.02.010

Cabrera, E. F. (2001). The key role of organizational culture in a multi-system view of technology-driven change. *International Journal of Information Management, 21*(3), 245–261.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research, 184*, 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.004

Chen, Y. (2018). Blockchain tokens and the potential democratization of entrepreneurship and innovation. *Business Horizons, 61*(4), 567–575. https://doi.org/10.1016/j.bushor.2018.03.006

Choi, T., Feng, L., & Li, R. (2020). Information disclosure structure in supply chains with rental service platforms in the Blockchain technology era. *International Journal of Production Economics, 221*, 107473. https://doi.org/10.1016/j.ijpe.2019.08.008

Da Xu, L., Xu, E. L., & Li, L. (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research, 56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

Di, D., Maesa, F., & Mori, P. (2020). Blockchain 3.0 applications survey. *Journal of Parallel and Distributed Computing, 138*, 99–114. https://doi.org/10.1016/j.jpdc.2019.12.019

Dowty, R. A., & Wallace, W. A. (2010). Implications of organizational culture for supply chain disruption and restoration. *International Journal of Production Economics, 126*(1), 57–65. https://doi.org/10.1016/j.ijpe.2009.10.024

Farhana, A., Petr, H., Munk, M., Abedin, M. Z., Shahriare, S., Efat, M. I. A., & Islam, J. (2022). A systematic review of blockchain applications. *IEEE Access*.

Garcia-lorenzo, L. (2020). Organizational remembering as a trigger for cultural change: Exploring the episodic memories of a financial scandal. *Scandinavian Journal of Management, 36*(1), 101089. https://doi.org/10.1016/j.scaman.2019.101089

Ghasemian, I., Masoomi, B., & Ghorbani, S. (2020). Technology in society expert oriented approach for analyzing the blockchain adoption barriers in humanitarian supply chain. *Technology in Society, 63*, 101427. https://doi.org/10.1016/j.techsoc.2020.101427

Giovanni, P. D. (2020). Blockchain and smart contracts in supply chain management: A game theoretic model. *International Journal of Production Economics, 228*, 107855. https://doi.org/10.1016/j.ijpe.2020.107855

Gordon, G. G. (1991). Industry determinants of organizational culture. *Academy of Management Review, 16*(2), 396–415.

Halpern, N., Mwesiumo, D., Suau-sanchez, P., Budd, T., & Bråthen, S. (2021). Ready for digital transformation? The effect of organisational readiness, innovation, airport size and ownership on digital change at airports. *Journal of Air Transport Management, 90*, 101949. https://doi.org/10.1016/j.jairtraman.2020.101949

Han, D., Zhang, C., Ping, J., & Yan, Z. (2020). Smart contract architecture for decentralized energy trading and management based on blockchains. *Energy, 199*, 117417. https://doi.org/10.1016/j.energy.2020.117417

Hartley, J. L., & Sawaya, W. J. (2019). Tortoise, not the hare: Digital transformation of supply chain business processes. *Business Horizons, 62*(6), 707–715. https://doi.org/10.1016/j.bushor.2019.07.006

Hernandez, F. G. H. (2003). Derivatives and the FASB: Visibility and transparency? *Critical Perspectives on Accounting, 14*(8), 777–789. https://doi.org/10.1016/S1045-2354(02)00192-2

Hewa, T., Ylianttila, M., & Liyanage, M. (2020). Survey on blockchain based smart contracts: Applications, opportunities and challenges. *Journal of Network and Computer Applications, 177*, 102857. https://doi.org/10.1016/j.jnca.2020.102857

Holland, D., Krause, A., Provencher, J., & Seltzer, T. (2017). Transparency tested: The influence of message features on public perceptions of organizational transparency. *Public Relations Review, 44*(2), 256–264. https://doi.org/10.1016/j.pubrev.2017.12.002

Howson, P. (2020). Building trust and equity in marine conservation and fisheries supply chain management with blockchain. *Marine Policy, 115*, 103873. https://doi.org/10.1016/j.marpol.2020.103873

Hu, T., Liu, X., Chen, T., Zhang, X., & Huang, X. (2021). Transaction-based classification and detection approach for Ethereum smart contract. *Information Processing & Management, 58*(2), 102462.

Huang, F., & Vasarhelyi, M. A. (2019). Applying robotic process automation (RPA) in auditing: A framework. *International Journal of Accounting Information Systems, 35*, 100433. https://doi.org/10.1016/j.accinf.2019.100433

Isensee, C., Teuteberg, F., Griese, K., & Topi, C. (2020). The relationship between organizational culture, sustainability, and digitalization in SMEs: A systematic review. *Journal of Cleaner Production, 275*, 122944. https://doi.org/10.1016/j.jclepro.2020.122944

Jabbar, S., Lloyd, H., Hammoudeh, M., Adebisi, B., & Raza, U. (2020). Blockchain-enabled supply chain: Analysis, challenges, and future directions. *Multimedia Systems, 27*, 787–806. https://doi.org/10.1007/s00530-020-00687-0

Jabbour, C. J. C., Fiorini, P. D. C., Ndubisi, N. O., Queiroz, M. M., & Piato, É. L. (2020). Digitally-enabled sustainable supply chains in the 21st century: A review and a research agenda. *Science of the Total Environment, 725*, 138177. https://doi.org/10.1016/j.scitotenv.2020.138177

Kar, A. K., & Navin, L. (2021). Telematics and informatics diffusion of blockchain in insurance industry: An analysis through the review of academic and trade literature. *Telematics and Informatics, 58*, 101532. https://doi.org/10.1016/j.tele.2020.101532

Karamchandani, A., Srivastava, S. K., & Srivastava, R. K. (2019). Perception-based model for analyzing the impact of enterprise blockchain adoption on SCM in the Indian service industry. *International Journal of Information Management, 52*, 102019. https://doi.org/10.1016/j.ijinfomgt.2019.10.004

Karine, C. (2020). The relationship between presenteeism and job satisfaction: A mediated moderation model using work engagement and perceived organizational support. *European Management Journal, 39*(2), 270–278. https://doi.org/10.1016/j.emj.2020.09.001

Khan, M., Nasfikur, R., Bala, S., Yesmin, S., & Abedin, M. Z. (2022). Bioinformatics: The importance of data mining techniques. In *Sentimental analysis and deep learning* (pp. 411–422). Springer.

Kimani, D., Adams, K., Attah-Boakye, R., Ullah, S., Frecknall-Hughes, J., & Kim, J. (2020). Blockchain, business and the fourth industrial revolution: Whence, whither, wherefore and how? *Technological Forecasting and Social Change, 161*, 120254.

Kokina, J., & Blanchette, S. (2019). Early evidence of digital labor in accounting: Innovation with robotic process automation. *International Journal of Accounting Information Systems, 35*, 100431. https://doi.org/10.1016/j.accinf.2019.100431

Kshetri, N. (2018). 1 Blockchain's roles in meeting key supply chain management objectives. *International Journal of Information Management, 39*, 80–89. https://doi.org/10.1016/j.ijinfomgt.2017.12.005

Lafourcade, P., & Lombard-platet, M. (2020). About blockchain interoperability. *Information Processing Letters, 1*, 105976. https://doi.org/10.1016/j.ipl.2020.105976

Leary, D. E. O. (2017). Configuring blockchain architectures for transaction information in blockchain consortiums: The case of accounting and supply chain systems. *Intelligent Systems in Accounting, Finance and Management, 24*(4), 138–147. https://doi.org/10.1002/isaf.1417

Li, W., Liu, K., Belitski, M., Ghobadian, A., & Regan, N. O. (2016). Sized enterprises in the digital age. *Journal of Information Technology, 31*(2), 185–206. https://doi.org/10.1057/jit.2016.10

Lin, H. (2013). Understanding the determinants of electronic supply chain management system adoption: Using the technology–organization–environment framework. *Technological Forecasting and Social Change, 86*, 80–92. https://doi.org/10.1016/j.techfore.2013.09.001

Liou, D., Tu, C., & Chang, S. (2012). Mediating effect between supportive culture and job satisfaction in administrative services at higher education institutions. *Human Factors and Ergonomics in Manufacturing & Service Industries, 24*(6), 627–640.

Liu, H., Ke, W., Kee, K., Gu, J., & Chen, H. (2010). The role of institutional pressures and organizational culture in the firm' intention to adopt internet-enabled supply chain management systems. *Journal of Operations Management, 28*(5), 372–384. https://doi.org/10.1016/j.jom.2009.11.010

Michela, J. L., & Burke, W. W. (2000). Organizational culture and climate in transformations for quality and innovation. In *Handbook of organizational culture and climate* (pp. 225–244). Springer.

Muthukrishnan, N., Maleki, F., & Ovens, K. (2020). Brief history of artificial intelligence. *Neuroimaging Clinics, 30*(4), 393–399. https://doi.org/10.1016/j.nic.2020.07.004

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review, 21260*, 1–9.

Nam, K., Dutt, C. S., Chathoth, P., Khan, M. S., & Dutt, C. S. (2021). Blockchain technology for smart city and smart tourism: Latest trends and challenges. *Asia Pacific Journal of Tourism Research, 26*, 454–468. https://doi.org/10.1080/10941665.2019.1585376

Nasiri, M., Ukko, J., Saunila, M., & Rantala, T. (2020). Technovation managing the digital supply chain: The role of smart technologies. *Technovation, 96–97*, 102121. https://doi.org/10.1016/j.technovation.2020.102121

Papagiannidis, S., Harris, J., & Morton, D. (2020). WHO led the digital transformation of your company? A reflection of IT related challenges during the pandemic. *International Journal of Information Management, 55*, 102166.

Petr, H., & Abedin, M. Z. (2020). A profit function-maximizing inventory backorder prediction system using big data analytics. *IEEE Access, 8*, 58982–58994. https://doi.org/10.1109/ACCESS.2020.2983118

Prasitlumkum, N., Cheungpasitporn, W., Chokesuwattanaskul, A., Thangjui, S., Thongprayoon, C., Bathini, T., Vallabhajosyula, S., Kanitsoraphan, C., Leesutipornchai, T., & Chokesuwattanaskul, R. (2020). Diagnostic accuracy of smart gadgets/wearable devices in detecting atrial fibrillation: A systematic review and meta-analysis. *Archives of Cardiovascular Diseases, 114*(1), 4–16. https://doi.org/10.1016/j.acvd.2020.05.015

Queiroz, M. M., & Fosso, S. (2019). Blockchain adoption challenges in supply chain: An empirical investigation of the main drivers in India and the USA. *International Journal of Information Management, 46*, 70–82. https://doi.org/10.1016/j.ijinfomgt.2018.11.021

Sahebi, I. G., Masoomi, B., & Ghorbani, S. (2020). Expert oriented approach for analyzing the blockchain adoption barriers in humanitarian supply chain. *Technology in Society, 63*, 101427.

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). Data transformation based optimized customer churn prediction model for the telecommunication industry. arXiv preprint arXiv:2201.04088.

Saurabh, S., & Dey, K. (2020). Blockchain technology adoption, architecture, and sustainable agri-food supply chains. *Journal of Cleaner Production, 284*, 124731. https://doi.org/10.1016/j.jclepro.2020.124731

Schmidt, C. G., & Wagner, S. M. (2019). Blockchain and supply chain relations: A transaction cost theory perspective. *Journal of Purchasing and Supply Management, 25*(4), 100552. https://doi.org/10.1016/j.pursup.2019.100552

Shajalal, M., Hajek, P., & Abedin, M. Z. (2021). Product backorder prediction using deep neural network on imbalanced data. *International Journal of Production Research, 1–18*, 1. https://doi.org/10.1080/00207543.2021.1901153

Shi, S., He, D., Li, L., & Kumar, N. (2020). Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. *Computers & Security, 97*, 101966. https://doi.org/10.1016/j.cose.2020.101966

Singh, M., & Kim, S. (2018). Branch based blockchain technology in intelligent vehicle. *Computer Networks, 145*, 219–231. https://doi.org/10.1016/j.comnet.2018.08.016

Sok, J., Blomme, R., & Tromp, D. (2014). Positive and negative spillover from work to home: The role of organizational culture and supportive arrangements. *British Journal of Management, 25*(3), 456–472. https://doi.org/10.1111/1467-8551.12058

Song, Y., Escobar, O., Arzubiaga, U., & Massis, A. D. (2021). The digital transformation of a traditional market into an entrepreneurial ecosystem. *Review of Managerial Science, 16*(1), 65–88. https://doi.org/10.1007/s11846-020-00438-5

Sunny, J., Undralla, N., & Pillai, V. M. (2020). Supply chain transparency through blockchain-based traceability: An overview with demonstration. *Computers & Industrial Engineering, 150*, 106895. https://doi.org/10.1016/j.cie.2020.106895

Thakur, S., & Breslin, J. G. (2020). Scalable and secure product serialization for multi-party perishable good supply chains using blockchain. *Internet of Things, 11*, 100253. https://doi.org/10.1016/j.iot.2020.100253

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research, 122*, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi, J., Fabian, N., & Haenlein, M. (2019). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research, 122*, 889–901. https://doi.org/10.1016/j.jbusres.2019.09.022

Viriyasitavat, W., & Hoonsopon, D. (2018). Cultivating product innovation performance through creativity: The impact of organizational agility and flexibility under technological turbulence. *Journal of Industrial Information Integration, 33*(4), 741–762. https://doi.org/10.1016/j.jii.2018.07.004

Wang, X., Fu, D., Fruk, G., Chen, E., & Zhang, X. (2018). Improving quality control and transparency in honey peach export chain by a multi-sensors-managed traceability system. *Food Control, 88*, 169–180. https://doi.org/10.1016/j.foodcont.2018.01.008

Wang, Y., Singgih, M., Wang, J., & Rit, M. (2019). Title page making sense of blockchain technology: How will it transform supply chains? *International Journal of Production Economics, 211*, 221–236. https://doi.org/10.1016/j.ijpe.2019.02.002

Wang, B., Luo, W., Zhang, A., Tian, Z., & Li, Z. (2020). Computers in industry blockchain-enabled circular supply chain management: A system architecture for fast fashion. *Computers in Industry, 123*, 103324. https://doi.org/10.1016/j.compind.2020.103324

Wu, J. Y. (2008). A general behavior model and new definitions of organizational cultures. *The Journal of Socio-Economics, 37*(6), 2535–2545. https://doi.org/10.1016/j.socec.2007.10.002

Xuan, S., Zheng, L., Chung, I., Wang, W., Man, D., Du, X., Yang, W., & Guizani, M. (2020). An incentive mechanism for data sharing based on blockchain with smart contracts R. *Computers and Electrical Engineering, 83*, 106587. https://doi.org/10.1016/j.compeleceng.2020.106587

Yang, F., Qiao, Y., Abedin, M. Z., & Huang, C. (2022). Privacy-preserved credit data sharing integrating Blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics, 2022*, 1. https://doi.org/10.1109/TII.2022.3151917

Yasin, M., Sabir, M., & Ali, H. (2019). CertLedger: A new PKI model with certificate transparency based on blockchain. *Computers & Security, 85*, 333–352. https://doi.org/10.1016/j.cose.2019.05.013

# Integration of Artificial Intelligence Technology in Management Accounting Information System: An Empirical Study

**Emon Kalyan Chowdhury**

**Abstract** At present, most of the business organizations take their management decisions using traditional approach. In the traditional approach, the freedom to be flexible is limited due to numerous assumptions. This paper aims to establish an artificial neural network-based model to predict management information and verify the accuracy of the model using some real data. The proposed model covers five dimensions, namely, accounting analysis management system, accounting decision support system, performance management information system, risk management information system, and environmental management information system. It is observed that the proposed model can predict the management accounting information by 98.83%, which is extremely good and meets the accounting information requirement.

**Keywords** Artificial intelligence · Machine learning · Management accounting · Information system · Neural network

## 1 Introduction

Management accounting provides information to managers who make important decisions in an organization (Garrison et al., 2003). The size and complexity of data is increasing day by day as a result managers are in serious trouble in processing large amount of data (Munim et al., 2020). The success of a decision depends on the quality of the information. Therefore, an efficient management accounting information system where data are processed through artificial intelligence technology plays a vital role in improving the operating efficiency of an organization (Zhang, 2021).

Management enterprises are substantially dependent on advanced information technology to make rational and effective decisions. Among management information systems, the management accounting information system is the most important

E. K. Chowdhury (✉)
CIU Business School, Chittagong Independent University, Chattogram, Bangladesh

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_3

segment (Hutahayan, 2020). The significance of management accounting information system lies in the economic progress, expansion, scale of economies, acquisition, and continuous improvement of strengths through scientific management decisions (Cai et al., 2019).

Practically, the use of management accounting information system is confined to the cost management, preparation of different budgets, and performance management. Smooth functioning of enterprise management is highly dependent on the comprehensive and stable construction of management accounting information systems integrated with other management information systems (Goetz et al., 2015).

The remaining part continues as follows. Section 2 reviews previous studies. Section 3 analyzes different models based on artificial intelligence technology. Section 4 experiments the success rates of prediction capacity of model using authentic management information data, and Sect. 5 concludes the paper.

## 2 Literature Review

Management control systems ensure optimal use of limited resources to achieve the organization's goal. In addition to financial data, an efficient management control system also uses psychological and control variables (Nguyen et al., 2017). The data from multiple sources are collected and fed into the management information system so as to generate various sub-objectives from a single organizational objective. It helps to compare the actual performance with the projected plans from diverse perspectives (Al-Ali et al., 2017). To sustain itself in a competitive and technology-based environment, an organization must strengthen its managerial and supervisory functions by introducing a management control system (Chi et al., 2019; Xin et al., 2018). Out of the different wings of the management information system, the development of the management accounting information system is crucial, as it directly contributes to the organization's financial solvency, internal control system, customer retention, and overall sustainability (Chowdhury, 2019; Ward et al., 2016). Recently, the use of an e-commerce-based accounting information system has increased tremendously among the enterprises to enjoy competitive advantages (Shajalal et al., 2021; Hidayat et al., 2020). Management accounting plays an important role in fulfilling the economic needs of an organization's operation and management with the help of responsibility center. The responsibility center ensures optimum uses of internal accounting control systems and further assists in organizing and delivering other functional internal management systems (Ghasemi et al., 2019). Amershi et al. (2014) observed a significantly positive impact of management accounting on innovation management. Management accounting systems simplify the cost calculation of single and batch products (Rodriguez-Galiano et al., 2015). Cooper et al. (2017) noticed the increasing popularity of using balanced scorecards in organizations to measure the performance of different indicators.

The traditional management accounting system mostly depends on the assumptions rather than versatility of data, which imperatively directs to take fixed decisions. This study finds a gap to explore the possibility of taking dynamic decisions by using alternative models where artificial intelligence technology is used in line with machine learning and data mining algorithms.

## 3 Artificial Neural Network (ANN)

The design of ANN is inspired by the structure of biological neurons such as the human brain. In a human brain, neurons create a network through interconnections. A neuron is known as a cell and executes a single task by responding to an input signal. In an ANN, the nodes are connected to each other and establish a network among themselves. The nodes are designed using artificial intelligence to handle massive amount of data using multiple equations simultaneously. In this network, the equations are established through sequential computations following a trial-and-error approach (Abedin et al., 2021; Chakraborty et al., 2018). The basic structure of ANN is expressed in Fig. 1.

Input neurons $X_1, X_2, \ldots, X_n$ indicate various inputs to the network, synapse weights $W_1, W_2, \ldots, W_n$ signify the weights of connections. The weights are very important in ANN as these represent the strength of each node. The weights that govern the effect of neurons are measured in the numerical parameters, which determines the output by converting the input.

The hidden layer performs the processing task. It applies two operational functions, the summation function and the transfer or activation function. The summation function multiplies each input ($X_i$) with the corresponding weight ($W_i$) and all products ($W_i \times X_i$) result in the summation function $\xi = \sum W_i \times X_i + B$, where $B$ represents the bias value. It controls the output of the neuron in line with the weighted sum of inputs.

The activation function transforms the input signal from the summation function into to output of a node for an ANN model. Each ANN is made up of three components. First, the node character determines inputs and outputs through signal processing. Second, the network topology determines how the nodes are connected

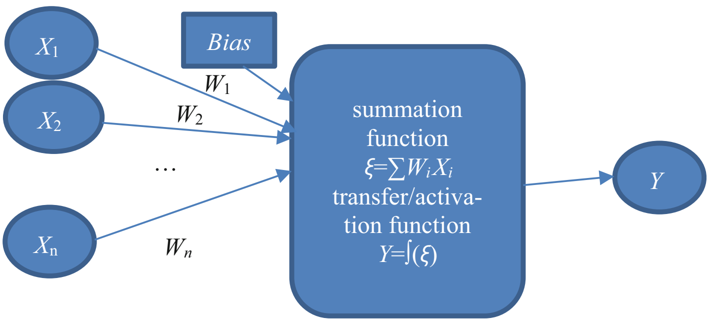



**Fig. 1** Model of an artificial neuron

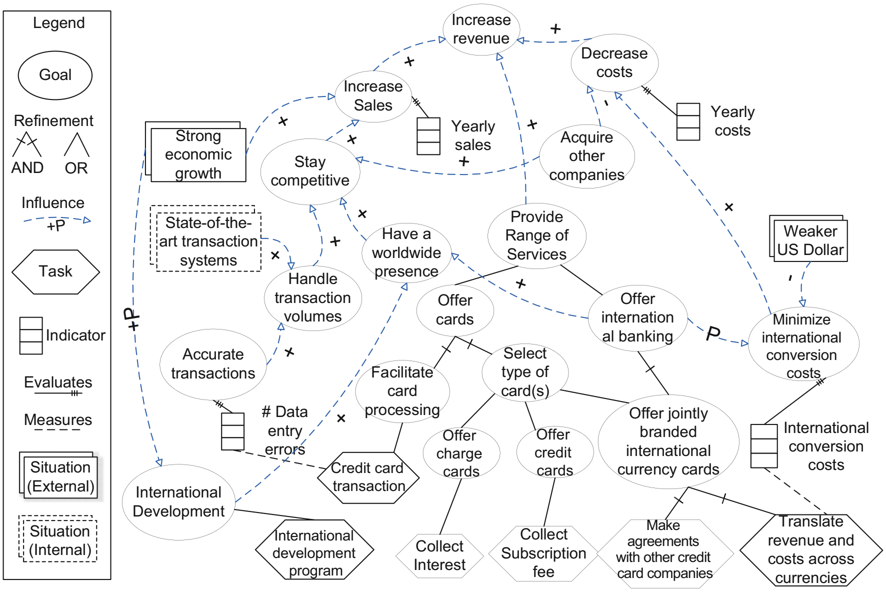



**Fig. 2** Business intelligence model for a credit card company. Source: Horkoff et al. (2012)

and organized. Third, the learning rules create and adjust weights. A few widely used ANN-based models have been discussed below.

**Business Intelligence (BI) Model**

BI helps an organization excel at operational activities in such a way that helps tap the opportunities in the market while overcoming potential threats. It has the capacity to generate effective information to take strategic decisions by processing massive volume of data. BI establishes a network between an organization and the external environment with the support of different reasoning techniques that controls influences, situations, and the indicators (Fig. 2). The reasoning tools for this model are "what if" a bottom-up approach and "is it possible?" a top-down approach (Horkoff et al., 2012).

**Three-Tier Data Model**

The three-tier data model is widely used in the data warehouse management of an organization. It provides subject-wise analytical environment in the global context (Abedin et al., 2020; Lau et al., 2018). The three tiers have been outlined below:

(a) Conceptual model: This is the top level of the model which is expressed by topics. Topics are derived from the dimensions and measures. Dimension refers to a perspective through which people observe the world, and measurement is related to data information of the dimension. For example, sales volume.
(b) Logical model: Logical model may be classified into two models such as the star model and the snowflake model. The star model includes the fact table and dimension model, while at the same time, they are connected to each other. The

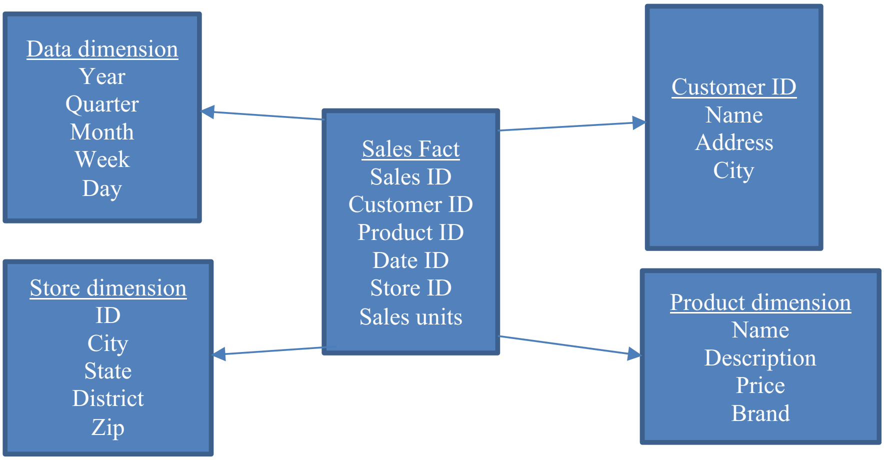



**Fig. 3** Star model

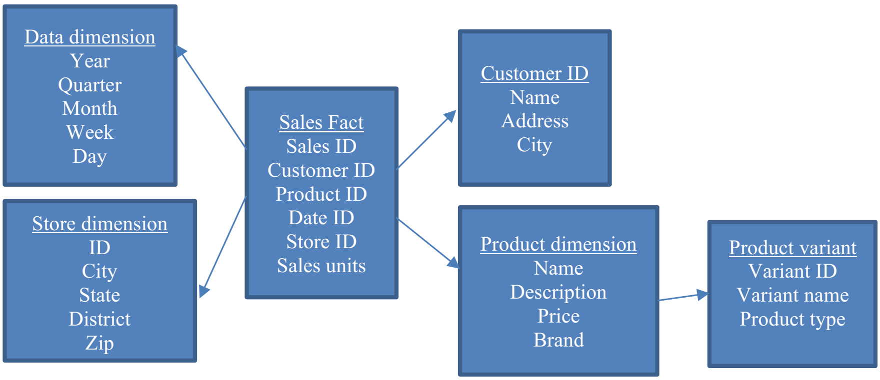



**Fig. 4** Snowflake model

star model is shown in Fig. 3. The sales data are generated in different time dimension including customer details, store details, and product details.

A snowflake model is an extension of the star model. It includes additional information about a particular dimension (Fig. 4). It uses similar disk space, is easy to install, and reduces query performance for multiple tables.

**Extract, Transform, Load (ETL) Model**

In this model, data are extracted from multiple source systems and then converted to final data after necessary calculations. The converted data are loaded into the data warehouse system for managerial decision. Source points include relevant stakeholders such as analysts, developers, testers, and top brass executives. Since ETL activities occur regularly, the data warehouse required to be updated, agile, and properly documented. ETL helps to make critical business decisions, and compare

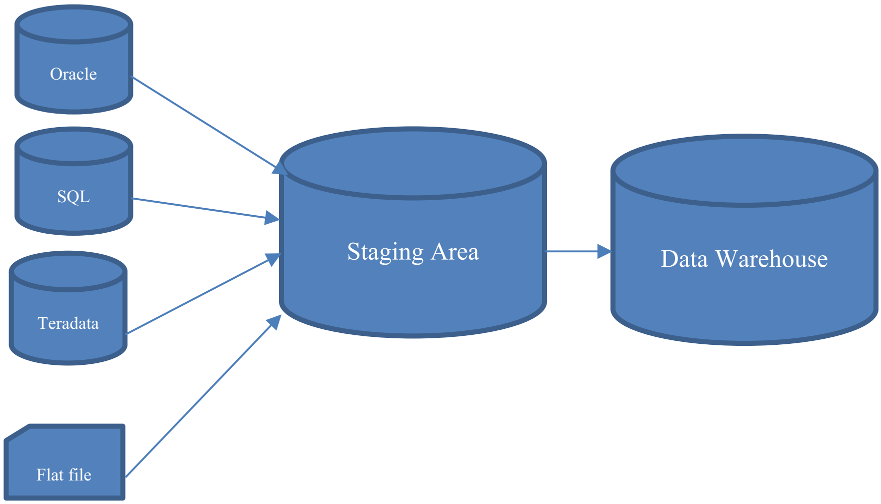



**Fig. 5** ETL model

the data of the source and target system through data migration and manipulation. Where the transactional database fails to answer complex business questions, ETL can easily and quickly address them (Hajek & Abedin, 2020; Sabtu et al., 2017). Figure 5 shows the ETL process in three steps.

In the ETL model, data are fed into the staging area by extracting them from the source points after due validations. Data are extracted from the source points in raw format, and at the transformation stage, data are cleaned, mapped, and converted. In this stage, the ETL assigns values and modifies the data so that business intelligence-based reports can be generated. Warehousing data is the last step of the ETL model. Here, a huge volume of data can be loaded in significantly less time. If the loading process fails, the recovery mechanism is activated without failure of any sort of data integrity. The entire ETL process is controlled by the warehouse administrator (Abedin et al., 2018).

**Cube Structure**

The data cube is a three-dimensional way of presenting data. In this model, the data are judged from various perspectives. When data cannot be presented in traditional column and row format due to more variables and context, data cube can make it so simple by utilizing different angles (Augenstein et al., 2018). Data cubes have the following categories.

(a) Multidimensional data cube: Most of the online analytical processing (OLAP) products are designed using a multidimensional array. These OLAPs perform better than other approaches, as they can be indexed straight to collect subsets of data. The larger the dimension, the sparser the cubes.
(b) Rational OLAP (ROLAP): This model uses a relational database to store and manage warehouse data. ROLAP servers are highly scalable and analyze

massive volumes of data across multiple dimensions. It also stores and analyzes highly volatile and changeable data.

To understand the presentation of the data in cube structure, the following information can be considered (Table 1).

The above information is shown in a three-dimensional cube (Fig. 6).

The essence of the cube structure lies in the capacity to show different data in a single image.

**Data Mining (DM) Process**

DM is an essential part of the management accounting information system (Kara et al., 2020). It combines database, statistics, machine learning, and other relevant technologies. It generates required information for managers amalgamating different data to enjoy competitive advantages (Abedin et al., 2019). Figure 7 depicts the data mining process.

## 4 Proposed Model

In light of the above analysis, this study recommends an Intelligent Management Accounting Information System (IMAIS) for the decision-making process where the following aspects are integrated. This model is the extension of Zhang (2021) where the environmental management information system was not included. In this model, the impact of the management decision on the environment has been considered. The integrated systems are as follows:

(a) Accounting analysis management system
(b) Performance management information system
(c) Accounting decision support system
(d) Risk management information system, and
(e) Environmental management information system

This recommended model can provide customized information to take decisions in time and also helps to run its business in a way better ensuring a sound internal control system. Figure 8 shows an IMAIS formation structure.

The recommended IAMAIS model covers reporting systems, risk management, performance management, decision support issues, and environmental issues. Each sub-system works autonomously and combinedly to fulfill segment and enterprise requirements.

**Test of Model Efficiency**

To verify the degree of accuracy of the proposed model, this study has used real management accounting data. Out of 380 observations, a total of 125 observations have been used classifying into 13 categories to train the model. The predicted results and actual results are shown in Fig. 9.

**Table 1** Location-wise quarterly data

| Location = "Chicago" | | | | | Location = "New York" | | | | Location = "Toronto" | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | | | | | Item | | | | Item | | | |
| Home | | | | | Home | | | | Home | | | |
| Time | Ent. | Comp. | Phone | Sec. | Ent. | Comp. | Phone | Sec. | Ent. | Comp. | Phone | Sec. |
| Q1 | 854 | 882 | 89 | 623 | 1087 | 968 | 38 | 872 | 818 | 746 | 43 | 591 |
| Q2 | 943 | 890 | 64 | 698 | 1130 | 1024 | 41 | 925 | 894 | 769 | 52 | 682 |
| Q3 | 1032 | 924 | 59 | 789 | 1034 | 1048 | 45 | 1002 | 940 | 795 | 58 | 728 |
| Q4 | 1129 | 992 | 63 | 870 | 1142 | 1091 | 54 | 984 | 978 | 864 | 59 | 784 |

**Fig. 6** Cube structure

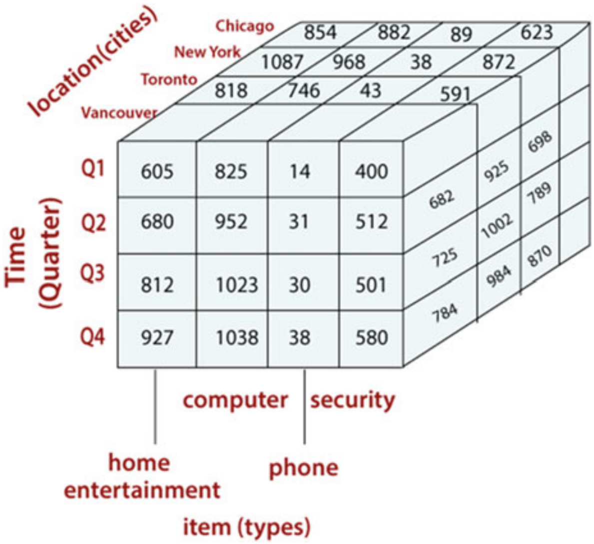



**Fig. 7** Data mining process

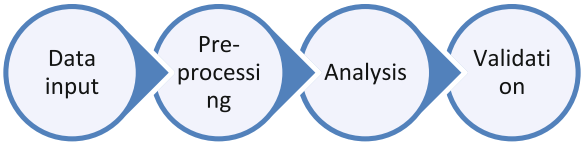
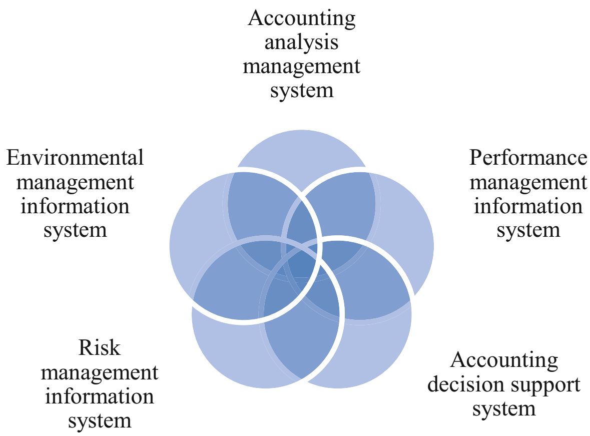



**Fig. 8** Intelligent management accounting information system




It is observed that the prediction is very close to the actual results for most of the observations. To get a further clear scenario, the residuals of the actual and predicted data are shown in Fig. 10.

It is also observed that most residuals hover within 0.05 to −0.05 and a very insignificant number of observations are above 0.1 to −0.01. This clearly indicates that the model is capable of predicting management information with an accuracy rate of 98.83%. As the rate is very close to 100%, it may be applied in the real world.

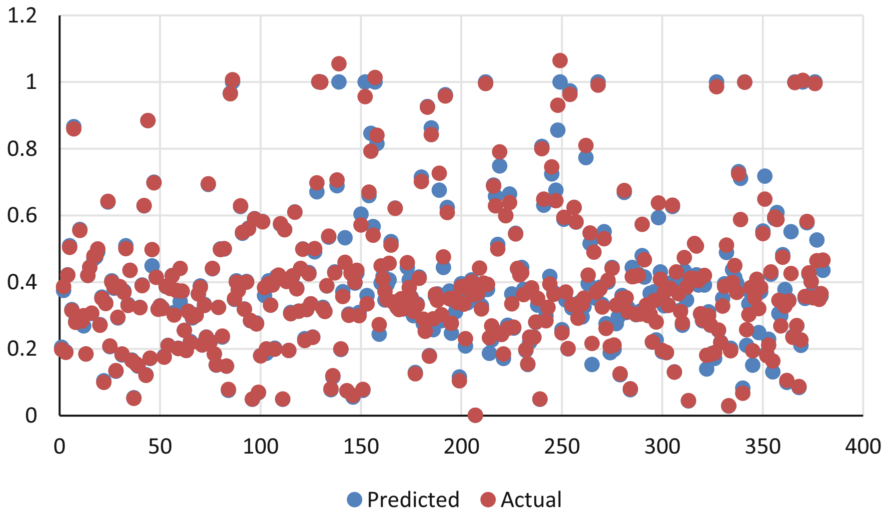



**Fig. 9** Actual vs. predicted data

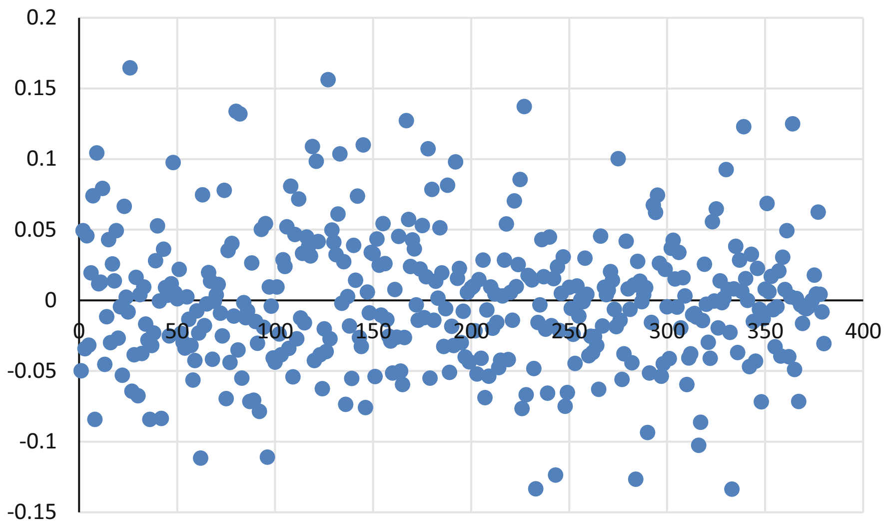



**Fig. 10** The residuals of actual vs. predicted results

## 5 Conclusion

This study aimed to formulate a management accounting information system using machine learning and an artificial neural network model. Being a vital sub-information system of management information system, the management accounting information system plays a very important role in the accounting development, therefore it should incorporate the accounting analysis management system, performance management information system, accounting decision support system, risk management information system, and environmental management information system. The recommended model can predict the accounting data with an accuracy rate of 98.83%. As the business world is complex and affected by many factors, the use of artificial intelligence technology to make management accounting decisions knows no bounds. It is assumed that the synergy of five dimensions helps in taking appropriate business decisions. Future researchers may include legal and ethical issues in the model to make this model more reliable and applicable as these issues vary from country to country.

## References


Abedin, M. Z., Chi, G., Colombage, S., & Moula, F. E. (2018). Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk*. Accessed from https://ssrn.com/abstract=3175776

Abedin, M. Z., Guotai, C., Moula, F. E., Azad, A. S., & Khan, M. S. U. (2019). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*(1), 474–507.

Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., & Hajek, P. (2020). Tax default prediction using feature transformation-based machine learning. *IEEE Access, 9*, 19864–19881.

Abedin, M. Z., Hassan, M. K., Khan, I., & Julio, I. F. (2021). Feature transformation for corporate tax default prediction: Application of machine learning approaches. *Asia-Pacific Journal of Operational Research, 2140017*.

Al-Ali, A. R., Zualkernan, I. A., Rashid, M., Gupta, R., & AliKarar, M. (2017). A smart home energy management system using IoT and big data analytics approach. *IEEE Transactions on Consumer Electronics, 63*(4), 426–434.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105–120.

Augenstein, D., Fleig, C., & Maedche, A. (2018, June). Development of a data-driven business model transformation tool. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 205–217). Springer.

Cai, J., Huang, W., Yang, S., Wang, S., & Luo, J. (2019, August). A selection method for Denoising auto encoder features using cross entropy. In *International Conference on Intelligent Computing* (pp. 479–490). Springer.

Chakraborty, T., Chattopadhyay, S., & Chakraborty, A. K. (2018). A novel hybridization of classification trees and artificial neural networks for selection of students in a business school. *Opsearch, 55*(2), 434–446.

Chi, G., Uddin, M. S., Abedin, M. Z., & Yuan, K. (2019). Hybrid model for credit risk prediction: An application of neural network approaches. *International Journal on Artificial Intelligence Tools, 28*(05), 1950017.

Chowdhury, E. K. (2019). Transformation of business model through blockchain technology. *Accounting and Finance, 47*(5), 4–9.

Cooper, D. J., Ezzamel, M., & Qu, S. Q. (2017). Popularizing a management accounting idea: The case of the balanced scorecard. *Contemporary Accounting Research, 34*(2), 991–1025.

Garrison, R. H., Noreen, E. W., Brewer, P. C., & Mardini, R. U. (2003). *Managerial accounting*. McGraw-Hill/Irwin.

Ghasemi, R., Habibi, H. R., Ghasemlo, M., & Karami, M. (2019). The effectiveness of management accounting systems: Evidence from financial organizations in Iran. *Journal of Accounting in Emerging Economies, 9*(2), 182–207.

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers and Geosciences, 81*, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Hajek, P., & Abedin, M. Z. (2020). A profit function-maximizing inventory backorder prediction system using big data analytics. *IEEE Access, 8*, 58982–58994.

Hidayat, A. T., Dewantara, A. M. D., & Saifullah, S. (2020). The development of website on management information system for e-commerce and services. *Jurnal Sisfokom (Sistem Informasi dan Komputer), 9*(3), 380–386.

Horkoff, J., Borgida, A., Mylopoulos, J., Barone, D., Jiang, L., Yu, E., & Amyot, D. (2012, September). Making data meaningful: The business intelligence model and its formal semantics in description logics. In *OTM Confederated International Conferences. On the move to meaningful Internet systems* (pp. 700–717). Springer.

Hutahayan, B. (2020). The mediating role of human capital and management accounting information system in the relationship between innovation strategy and internal process performance and the impact on corporate financial performance. *Benchmarking: An International Journal, 27*(4), 1289–1318.

Kara, M. E., Fırat, S. Ü. O., & Ghadge, A. (2020). A data mining-based framework for supply chain risk management. *Computers & Industrial Engineering, 139*, 105570.

Lau, H. C., Ip, A., Lee, C. K. M., & Ho, G. T. (2018). Development of a three-tier assessment model: A case study. *Benchmarking: An International Journal, 25*(7), 2216–2229.

Munim, Z. H., Dushenko, M., Jimenez, V. J., Shakil, M. H., & Imset, M. (2020). Big data and artificial intelligence in the maritime industry: A bibliometric review and future research directions. *Maritime Policy & Management, 47*(5), 577–597.

Nguyen, T. T., Mia, L., Winata, L., & Chong, V. K. (2017). Effect of transformational-leadership style and management control system on managerial performance. *Journal of Business Research, 70*, 202–213.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804–818.

Sabtu, A., Azmi, N. F. M., Sjarif, N. N. A., Ismail, S. A., Yusop, O. M., Sarkan, H., & Chuprat, S. (2017, July). The challenges of extract, transform and loading (ETL) system implementation for near real-time environment. In *2017 International Conference on Research and Innovation in Information Systems* (ICRIIS) (pp. 1–5). IEEE.

Shajalal, M., Hajek, P., & Abedin, M. Z. (2021). Product backorder prediction using deep neural network on imbalanced data. *International Journal of Production Research, 2021*, 1–18.

Ward, L., Agrawal, A., & Choudhary, A. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Computational Materials, 2*, 16028. https://doi.org/10.1038/npjcompumats.2016.28

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *IEEE Access, 6*, 35365–35381.

Zhang, X. (2021). Application of data mining and machine learning in management accounting information system. *Journal of Applied Science and Engineering, 24*(5), 813–820.

# The Impact of Big Data on Accounting Practices: Empirical Evidence from Africa

**Mandella Osei-Assibey Bonsu, Naheed Roni, and Yongsheng Guo**

**Abstract** Big data is much more than accounting and financial data. Big data including financial and non-accounting data have become accessible in immense volumes in distinct forms and in real time. The use of big data for accounting is immobile in initial periods. However, academics have predicted that having high-quality accessible and accelerated in real time might lead to more comprehensive financial reporting. Literature on big data is inconclusive, theoretical, and dearth empirical studies and models. This prompted us to explore the impacts of big data on accounting using accountants in an African emerging country, Nigeria. We use multiple regression for 151 responses. The samples were collected using a random sampling method. The results of the evidence show that big data has a positive and significant impact on financial reporting, performance management, corporate budgeting, audit evidence, risk management, and fraud management. Moreover, evidence indicates that while big data significantly impact accounting and auditing of accountants, utilizing the diversity of data volume, data variety, and data velocity significantly enhances it. The study can help accountants, prospective accountants, and accounting graduates hone their competencies in studying and producing big data analytics, which will benefit the industry. Moreover, business institutions of higher learning should create business curriculums that use big data in their offerings. Finally, policymakers can help by establishing governance models for big data to organize its usage and prevent its exploitation.

**Keywords** Big data · Accounting · Auditing · Financial reporting · Nigeria · Africa

---

M. O.-A. Bonsu · N. Roni (✉) · Y. Guo
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.osie-assibeybonsu@tees.ac.uk; n.roni@tees.ac.uk; y.guo@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_4

## 1 Introduction

The dynamic business environment is calling business entities to invest time, money, and efforts to adapt to envisage ways of doing things. In fact, the change of the entire business model affects the way of accounting. Technology makes accounting free from manual intervention and identifies patterns and generates the exception reports, leaving accountants with grey areas. As an emerged technical term, data is regarded as the vehicle of the accounting profession (ICAEW, 2014). The growth of accounting and auditing has been empowering the development of big data to advance technologies breakthroughs in multiple areas such as data analytics and Artificial intelligence (Gepp et al., 2018; Bullock et al., 2020).

Big data is abundant more than accounting and financial data (Petr & Abedin, 2020). Big data such as financial and non-financial data, accounting, and non-accounting data, all of which become available in abundance volumes in distinct forms and in actual time (Blazquez & Domenech, 2018; Bag et al., 2020; Basukie et al., 2020). In fact, big data could enhance financial accounting, reporting, and auditing practices (Warren & Marz, 2015; Iqbal et al., 2020). This indicates that having high-quality data available and processed in real time could lead to more comprehensive fiscal information (Elmagrhi et al., 2019), improved management, and more dependable budgeting. Furthermore, big data is thought to increase quality by increasing accuracy and making information available in actual time (Cockcroft & Russell, 2018).

In Nigeria, firms from every industry are at the frontline, experiencing first-hand the disruptive changes that affect their accountants. The rapid escalation in the volume of data demands accountants to be equipped with the available technological tools to analyze a much higher volume of data in their practice than has previously been the case (Arner et al., 2015). For example, the use of data analytics hopes to turn the accounting profession from a reactive and backward-looking exercise to a constructive, continuous source of upward-looking insights that can be used all the time, with the accountants as the custodian and translator of the underlying data framework.

Insight on the impact of big data on accounting practices from accountants in Nigeria are obtained for three reasons. First, fintech in Nigeria has grown significantly for some years and is one of the ways for fintech in Africa. Second, anecdotal evidence that Nigeria is now home to over 200 fintech firms, plus several fintech solutions offered by firms as part of the product portfolio. Nigeria's sectors are thriving and continue to amaze, exhibiting unwavering development and cutting-edge data analytics. Finally, Nigeria has advanced as one of the African top fintech hubs, attracting 25 percent ($122 million) in investment raised by African tech startups in 2019 (Disrupt Africa, 2021). In this paper, we examine the role of big data in the practice of accounting and auditing in Nigeria.

Although some recent studies have linked growing technologies to the accounting profession, there have been no scholarly empirical studies on the relationship between big data and accounting (Chen et al., 2016; Shajalal et al., 2021). Although some related literature studies have been conducted, there has been no empirical research on the topic of accounting (Schmitz & Leoni, 2019; Lamboglia et al., 2020). Furthermore, the application of big data for accounting is immobile at the early stages (Scott & Orlikowski, 2012). Big data, however, is inconclusive, theoretical, and dearth of empirical models. Therefore, more empirical studies are needed to examine the impacts of big data on the works of accountants. To the best of our knowledge, this is the first study to examine the empirical impact of big data on accounting and auditing practices evidenced from an Africa emerging economy, Nigeria.

The research provides contributions to the management accounting literature in fourfolds. First, it is the first empirical evidence to examine whether big data impacts accounting practices in Nigeria. Second, the research contributes to the scant literature on big data and accounting practices in producing higher-quality audits to serve existing purposes. Thus, this paper provides evidence of the significance of big data to auditing practices. Third, this research offers useful insights that may assist accounting regulators in recognizing the importance of big data and accounting relationships in developing accounting standards, as big data is seen as having the ability to create and refine accounting and auditing standards (Warren & Marz, 2015). Furthermore, the research could assist institutions of higher learning in updating accounting curricula to handle big data. Finally, the study provides outcomes that are more general with wider applicability by using an Africa emerging country sample, which to the best of our knowledge, no research has studied.

The next section reviews the literature, followed by hypothesis development. Section 3 presents data and methods. Section 4 reports the findings, followed by discussions. The final sections conclude with policy implications.

## 2 Literature Review

In recent times, big data has become the buzzword. Big data is described as high-volume, high-velocity, and high-variety information assets that necessitate cost-effective, novel data management to enable improved intelligence, decision-making, and process automation (Gärtner & Hiebl, 2017). The three characteristics: volume, velocity, and variety advocate that large volumes of transactions are created swiftly from a diversity of sources.

Data are considered a vehicle for the accounting profession (ICAEW, 2014). On the other hand, big data is abundant more than accounting and financial data. Financial and non-financial data, accounting and non-accounting data, and numerical and quantifiable data are all examples of big data, which is easily obtainable in various formats, and in real time (Bag et al., 2020; Basukie et al., 2020). Big data has the potential to enhance management accounting, financial reporting, and financial accounting and auditing procedures (Brown-Liburd et al., 2015; Warren & Marz, 2015; Yoon et al., 2015; Iqbal et al., 2020). The study aims to investigate the impact of big data on accounting and auditing, including big data on financial reporting, management performance, audit evidence, risk and fraud management, and corporate budgeting. Warren and Marz (2015) and Moffitt and Vasarhelyi (2013) suggested that big can enhance financial reporting, improve transparency, accounting information quality, and enrich financial reporting evidence.

However, the empirical evidence on the effectiveness of big data in accounting is dearth in the literature. Apart from Al-Htaybat and von Alberti-Alhtaybat (2017), Chen et al. (2015a, b), and Sardi et al. (2020), there is no empirical research on how big data impact accounting, and auditing in Nigeria. Using interviews with 25 participants, Al-Htaybat and von Alberti-Alhtaybat (2017), discovered that data analysts and accountants should work in conjunction to advance financial reporting utilizing data management. Sardi et al. (2020), on the other hand, found that integrated performance grounded on big data can aid attain competitive advantage for firms.

However, these studies were unable to determine whether there are empirical positive relationships between big data, and accounting and auditing practices. The approach informing in this study stresses the imperative of big data on accounting, and auditing practices within accountants. Moreover, we have considered the approach or credit risking (Abedin et al., 2018, 2022) in relation to firms. Hence, we expect empirical impacts of big data on each of accounting and auditing practices. Researching the extant literature indicates that preceding studies about big data and accounting are mainly theoretical, and there is a dearth of empirical evidence on the use of big data in accounting. Moreover, no study has studied Africa.

## 3 Research Hypothesis

In this section, we develop hypotheses based on extant literature including big data on accounting and auditing (financial reporting, performance management, corporate budgeting, audit evidence, risk, and fraud management) as a results test if these variables have positive relationships with big data.

### *3.1 Big Data and Financial Reporting Relationships*

Transparency is the primary purpose of the governance system and corporate reporting. Warren and Marz (2015) found that big data can increase transparency, improve financial reporting, and lead to improvements in accounting information quality. Moreover, big data can enrich financial reporting (Moffitt & Vasarhelyi, 2013). The results of financial accounting are financial reporting that primarily affects managers and stakeholders. However, corporate reporting does not address the customers' changing needs.

Furthermore, in the era of big data, financial reports are still made quarterly, biannually, and annually. Financial reports are often publicly disclosed after the audit at the end of the financial year, which means that certain information may be no longer relevant. Investors and stakeholders are increasingly awaiting fast financial data, perhaps daily. In this respect, one of the characteristics of big data is the speed at which the data are processed and formed; big data schemes can now analyse and produce data in actual period. This can facilitate companies' timely publication of financial reports. For example, Walmart, Amazon, and Royal Bank of Scotland have used platforms for big data that process and provide data in real time (Marr, 2016). As a result, the implementation of a big-scale data system may have a significant impact on the ability of a company to provide timely financial reports to the public.

To date, there have been few empirical studies on big data and financial reporting relations. Aside from Al-Htaybat and von Alberti-Alhtaybat (2017), who found that data analysts and accountants should collaborate to enhance financial reporting through advanced analytics (Yang et al., 2022). Moreover, Arnaboldi et al. (2017) reviewed the literature and discovered from the literature that big data can help with financial reporting. Therefore, more empirical studies are needed to close this significant gap. Overall, big data can guarantee by escalating the quality of financial reporting, and henceforth hypothesized that:

*H1: Big data is positively related to quality financial reporting.*

## 3.2 The Impact of Big Data on Performance Management

Through the collection, compilation, filtering, analysis, interpretations, and dissemination of appropriate data, performance is a set of measuring tools and dashboards aimed at assessing management decisions and to quantifying the efficacy and efficiency of the actions conducted (Tambe, 2014). Many academics believe that as competitiveness has increased, performance management has become increasingly difficult (Manyika et al., 2014). More organized and unstructured data are becoming available and a diverse set of inputs is becoming increasingly vital for long-term economic success. Information technology will provide different dimensions to performance measurement processes. Typically, accounting managers use structured data such as retention of employees, customer satisfaction surveys, and return level to collect data on the four-point balanced scorecard (Richins et al., 2017).

Accountants and financial experts need to use large data to evaluate organizational performance (ACCA and IMA, 2013). First, Vera-Baquero et al. (2015) present a big data resolution that can give firm analysts instantaneous acumens into corporate performance and make measurements and significant performance indicators accessible. Second, an efficient balance scorecard system requires extensive and varied financial and non-financial data from internal and external sources. Big data technologies can provide numerous and diverse customer data and allow managers to effectively design BSC's customers' perspectives, measures, objectives, and strategies.

Studies on big data and performance management are mainly theoretical. Elkmash et al. (2021) did a tentative investigation and discovered that big data analytics lowers the cost of unstructured data analytics for customers and improves the capacity to respond to consumer concerns quickly. Moreover, Sardi et al. (2020) observed the relationships between big data and performance management and found that big data might enhance competitive advantages. As a result, big data can help managers establish the greatest vision and strategy for future occurrences.

The literature further determined that big data could help lengthen performance measurement by creating novel performance indicators (Arnaboldi et al., 2017). However, studies remain a theoretical argument in the absence of empirical research. Therefore, we suggest that big data can positively enhance performance management and accordingly propose the following hypothesis:

*H2: Big data positively enhance the performance management of accountants.*

### 3.3 Big Data and Corporate Budgeting Relationships

Budget is described as a quantitatively articulated realistic strategy for the future (Gleim & Flesher, 2015). CIMA (2008) stated that a budget is a quantitative description of a plan for a specific time. Budgets include anticipated returns and sales, costs, reserve quantities, and expenditures, as well as liabilities, assets, and financial inflow (CIMA, 2008). However, budgeting is a management function based on forecasts. According to Collier and Berry (2002), the budgeting process often considers risk and uncertainty, as well as data on internal and external occurrences. According to the Institute of Chartered Accountants of England and Wales (ICAEW), accountants may use big data analytical models to enhance budgeting and forecasting. Big data analytics is an organizational information system that reduces uncertainty and better predicts future resource needs (Chen et al., 2015a). However, Cokins (2014) claims that the use of advanced analytics and big data in corporate operations has changed conventional costing planning and budget variation control methods. Foremost, a large data volume provides managers with many data inputs for budgeting, allowing them to create more accurate budgeting valuations and predictions and hence lessen variances. Utilizing hundreds of inputs instead of fewer can yield improved and further accurate projections in forecasting (Duan & Xiong, 2015). Secondly, "Velocity," will give data that are analyzed simultaneously, allowing managers to track the budget implementation process in real time, potentially reducing implementation errors. Data streaming, conferring to Kudyba and Kudyba (2014) is one of the most important elements of big data analytics. Real-time data streams from their source are analyzed and made accessible to decision-makers. The third dimension, "Variety," might offer a variety of data formats for managers to choose from depending on the situation. Empirically, analyzing the large quantity of data accessible on consumers' tastes, rivals' products, and economic conditions with advanced analytics should produce more accurate request and sales forecasts in actual time. This indicates that big data predictive analytics could more properly estimate the future grounded on past events (Duan & Xiong, 2015).

Studies on the impact of big data on corporate budgeting are still based on theory and dearth empirical evidence (Fisher et al., 2002; De Baerdemaeker & Bruggeman, 2015; Chen et al., 2016). Adding big data to the budgetary process can help manage performance, resource allocation, and strategic target implementation with the least amount of fluctuation. Thus,

*H3: Big data is positively improving corporate budgeting.*

### *3.4 Big Data and Audit Evidence Relationships*

The use of big data and analytics can help improve the efficiency and quality of auditing (ICAEW, 2014). Audit evidence and big data relationships indicate considerable convergence. Since it combines traditional evidence with reliable, sufficient, and relevant information (Yoon et al., 2015), increased transparency of audit standards to audit evidence sources outside common financial data. Hence, it is the key facilitator for using big data by auditors. In fact, auditing conventional permit auditors to gather evidence from any source and format if it benefits in the formulation of an opinion.

The International Standard on Auditing (500) coined audit evidence as any information utilized by the auditor, whether presented in the accounting records or vice versa. Moreover, AICPA (2004) reckoned that audit evidence is any information utilized by the auditor to arrive at an audit conclusion, whether included in accounting records or otherwise. This suggests that the flexibility of auditing standards is in line with the distinctive features of big data. However, big data characteristics can allow auditors to obtain evidence from a variety of sources, forms, and in real time for the same audited items.

However, the motive is not only to have many diverse pieces of evidence, but also for the evidence to be sufficient, relevant, and reliable following auditing standards (Alles, 2015; Brown-Liburd et al., 2015). The unique qualities of big data can provide enough accurate audit evidence (Yoon et al., 2015). The accessibility of large amounts of data in numerous formats and in real time, as well as the improved competences of big data analytics, enhances the chances of collecting the most adequate and relevant audit evidence. In summary, big data and related analysis help auditors collect more appropriate relevant audit data and conclude an opinion with a better level of assurance. However, to the best of our knowledge, no empirical evidence is provided on whether big data positively improves the audit profession via the big data audit evidence relations. Hence, the study hypothesized that:

*H4: Big data is positively related to audit evidence.*

### 3.5 Big Data and Risk and Fraud Management Relationships

Companies face a variety of risks that, if not properly assessed and handled, could jeopardize their long-term viability. Among the main managerial concerns, and a key governance necessity rule, is risk management. The board of directors of the firms must maintain sound internal control and risk management systems (Council, 2011). Bigdata can enhance risk surveillance, risk cover, and risk decision-making models (Ibrahim et al., 2021). Big data and analytics offer accountants a variety of opportunities to improve risk management (ICAEW, 2014). Incorporating risk indicator measurements will enhance the precision, and these indicators provide a predictive value while providing the KRI in real time. However, because most risks are based on the future, the more data available, the more precise the assessment and forecast of risks. Big data predictive analytics enhances the stability and predictive performance of risk assessment models, which allows managers to anticipate risk forecasts more precisely (Duan & Xiong, 2015). Furthermore, big data can assist auditors to measure the risks of their current or potential clients more precisely than ever, including the risks of management fraud, falsification of financial statements, bankruptcy, and risks related to the design and execution of internal controls (Cao et al., 2015). Aboud and Robinson (2020) discovered that data analytics may be used to detect or prevent fraud.

Equally, managers and investors can use advanced risk assessment and estimate analytics to safeguard their companies and assets from financial and market risks such as liquidity, foreign currency, and share price volatility. Aside from fraud detection, big data's exceptional characteristics could aid enhance risk assessment, measurement, and prediction. For instance, data volume and diversity will provide a vast amount of internal, external, financial, and non-financial data in a range of categories, resolving the data scarcity Chen et al. (2015b) studied the Alibaba Group and found that big data can monitor and assess fraud threats in real time and send out alerts to prevent fraud. Empowered with this, more studies are needed on how big data may help with fraud detection and prevention (Cockcroft & Russell, 2018; Aboud & Robinson, 2020). In fact, firms have begun to utilize big data resolutions to develop their risk management schemes empirically.

However, there is a dearth of academic empirical studies on the use of big data in enhancing risk management systems. Chen et al. (2015b) is the only empirical research that we have found to bring the best out of our knowledge; hence, more empirical evidence is needed to study the connections between big data, risk, and fraud management. Hence, the study proposes that:

*H5: Big data positively improves risk and fraud management.*

### 3.6 Research Framework

The research model that derives the analysis in this study is based on the empirical review above. Therefore, the explanatory variable is big data, and the five hypotheses discussed above form the basis of the empirical analysis of the research. The research model for this study is presented in Fig. 1.

## 4 Research Methods and Data

### 4.1 Population and Sample

The paper examined the impact of big data on accounting practices among accountants in Nigeria. The sample consists of chartered accountants in Nigeria with an international designation granted including ACCA and CIMA. We used the random sampling technique which allowed us to obtain a sample of 152 representing chartered accountants in Nigeria. The evidence-chartered accountants used as sample is empowered that it is vital for CAs to be sure of and have working knowledge of big data. Furthermore, Nigeria, considered as the biggest economy in the African continent, has grown in data science. Most firms have started to implement the tools and techniques used in data science and fintech. Hence, Nigeria presents a rich setting to explore the empirical impacts of big data on accounting and auditing practice.

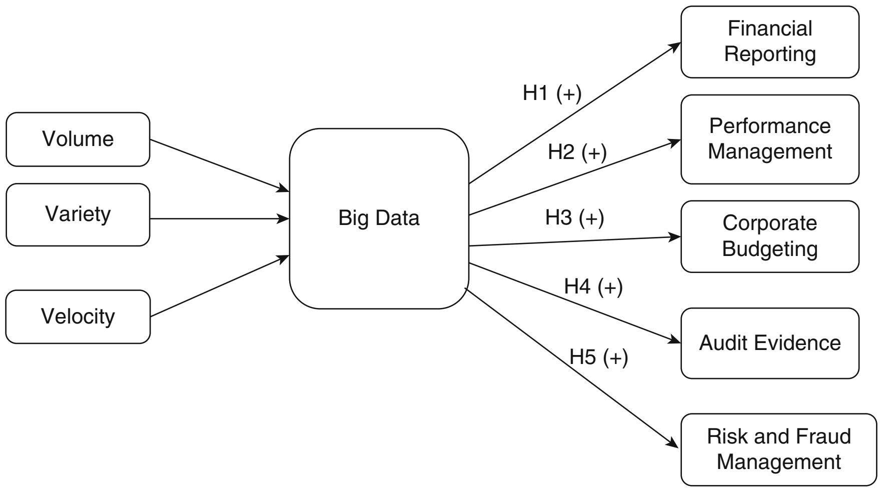



**Fig. 1** Study model

## 4.2 Questionnaires

Data were gathered among CAs through self-made questionnaires and were administered online from the period of October 2021 to January 2022. We prepared, pre-tested, and revised the draft of the three-page, and two-section questionnaires. First, pilot and pretesting were conducted by sending to 3 chartered accountants, and 2 University senior lecturers in accounting at UK recognized to the authors in big data. They were requested to review, correct, and suggest improvements of the original draft for relevance, content, and wordings. Second, we sent the refined, revised, and pre-tested questionnaires to respondents. The sections of the survey asked CAs to comment on the impact of big data on financial reporting, performance management, risk and fraud management, corporate budgeting, and audit evidence, and their respective profiles. To improve the response rate, cover letter was included stating the survey objectives, defining big data, and confidentiality were guaranteed. Finally, the survey link was generated online and sent in the email of selected respondents, which assured that their responses would be completely anonymous.

## 4.3 Measurement of Big Data

For the measuring scales for the construct of big data, we relied on the existing literature. Our study argues that the three big data characteristics (data volume, data variety, and data velocity) are essential, since combined together contribute to the big data constructions in accounting and auditing (Ghasemaghaei & Calic, 2019). Hence, we asked 9 questions on big data regarding volume, variety, and velocity on 7 Likert scale from (1, strongly disagree to, 7, strongly agree).

## 4.4 Measurement of Accounting and Auditing Practices

We used financial reporting, performance management, Risk & Fraud Management, Corporate budgeting, and audit evidence as constructs to measure accounting and auditing practices. Our self-administered questionnaires on accounting use twenty-two (22) items on 7 Likert scale from (1, strongly disagree to, 7, strongly agree).

## 4.5 Methods

To examine the proposed hypotheses, we assessed the equations for the data. We used regression as the current estimator for the impacts of big data on accounting. The model is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_t, \tag{1}$$

where $Y$ represents accounting and auditing practice practices, and $X_{1-3}$ represents big data, respectively. In the first place, we tested the effect of big data on Financial Reporting (FR) by introducing volume, variety, and velocity. Hence,

$$\mathrm{FR} = \beta_0 + \beta_1 \mathrm{VLM}_1 + \beta_2 \mathrm{VRT}_2 + \beta_3 \mathrm{VCT}_2 + \varepsilon_t. \tag{2}$$

Second, we tested the effect of big data on performance management, accordingly, we regressed the model as:

$$\mathrm{PM} = \beta_0 + \beta_1 \mathrm{VLM}_1 + \beta_2 \mathrm{VRT}_2 + \beta_3 \mathrm{VCT}_2 + \varepsilon_t. \tag{3}$$

We further tested the impact of big data volume, variety, and velocity on Corporate Budgeting, we thus estimate as:

$$\mathrm{CB} = \beta_0 + \beta_1 \mathrm{VLM}_1 + \beta_2 \mathrm{VRT}_2 + \beta_3 \mathrm{VCT}_2 + \varepsilon_t. \tag{4}$$

In addition, the single effect of big data on Audit Evidence was tested as follows:

$$\mathrm{AE} = \beta_0 + \beta_1 \mathrm{VLM}_1 + \beta_2 \mathrm{VRT}_2 + \beta_3 \mathrm{VCT}_2 + \varepsilon_t. \tag{5}$$

Finally, we tested the effect of big data on Risk and Fraud management and Eq. (5) shows as:

$$\mathrm{RFM} = \beta_0 + \beta_1 \mathrm{VLM}_1 + \beta_2 \mathrm{VRT}_2 + \beta_3 \mathrm{VCT}_2 + \varepsilon_t. \tag{6}$$

For details description of variables, see Table 1. Following the distribution of the questionnaires, we received completed one hundred and fifty-two (165) out of three hundred (300) distributed to a sample of accountants in Nigeria. After removing the missing and incomplete data, we were left with 151 responses that were detailed and adequate for analysis, accounting for 50.3 percent of the total. Table 2 reports the profile of the respondents. We discovered 95 accountants, 62.91% of whom were males and 37.09% of whom were females. Most of the respondents (54.30%) were between the ages of 26–45, with 82.12% are qualified from the Institute of Chartered Accountants of Nigeria (ICAN), followed by 10.59% with ACCA, and the majority (41.04%) had worked between 6–10 years.

Besides, we found that 61.59% works for the banking, finance, and insurance, 23.18% for the service industry, and 15.23% for the manufacturing industry. Finally, most of the respondents works in the private sector representing 75.50% leaving 24.50% for the public sector.

**Common Method Bias**

The study questionnaires are subjected to Common Method Bias (CMB) testing. Because the study used a survey to acquire data from a single provider, there is still a

**Table 1** Description of variables

| Constructs | Variable | Source |
|---|---|---|
| | (7-point Likert scale from "strongly disagree" to "strongly agree") | |
| Volume | Larger amounts of data are analyzed. | Ghasemaghaei and Calic (2019) |
| | The amount of data we examine is excessive. | |
| | We use a great deal of data, "in my opinion" | |
| Velocity | We are fast in exploring data | Ghasemaghaei and Calic (2019) |
| | We analyze data quickly | |
| | We analyze different sources of data to gain insights | |
| Variety | We examine data from multitude of sources. | Ghasemaghaei and Calic (2019) |
| | We use data to improve accounting information quality, and ensures transparency | |
| | We use data to enrich reporting information, and performance management. | |
| | (7-point Likert scale from "strongly disagree" to "strongly agree") | |
| Financial reporting | We use data to improve accounting information quality. | Developed |
| | We use data to enrich reporting information | |
| | We use data to ensure transparency, and improve accounting information quality | |
| | We use data to improve performance management | |
| Performance management | We use structured data to assess organizational performance | Developed |
| | Big data may supply enormous and diverse customer data | |
| | BDA allows to effectively design customer perspective objectives, measures, targets, and strategies | |
| | BDA gives real-time insights and makes measurements and key performance indicators | |
| | BDA provides business analytics real-time insights | |
| Corporate budgeting | Data analytics predicts models to improve budgeting and forecasting | Developed |
| | Data analytics provides managers with several inputs for budgeting, and allows budget estimations | |
| | Managers could track budget implementation budget in time | Developed |
| Audit evidence | | |
| | Extend the scope of initiatives and compare them to wider populations<br>Data may be analyzed in larger volumes and faster to provide auditors with relevant insights. | |
| | BDA helps financial auditors streamline the reporting process | |

(continued)

**Table 1** (continued)

| Constructs | Variable | Source |
|---|---|---|
| | Data analytics helps to detect fraud | |
| | (7-point Likert scale from "strongly disagree" to "strongly agree") | |
| Risk and fraud monitoring | Data can increase risk monitoring | Developed |
| | Data can enhance risk coverage, and creation of risk decisions making models | |
| | Analytics presents accountants with several chances to improve risk management | |
| | Data analytics may be used to detect or prevent fraud | |
| | Big data analytics aid to improve risk assessment, prediction, and measurement | |

**Table 2** Summary of data from the respondents

| Profile | Dimension | Frequency $n = 151$ | Percentage (%) |
|---|---|---|---|
| Sex | | | |
| | Male | 95 | 62.91 |
| | Female | 56 | 37.09 |
| Age | | | |
| | 20–25 | 30 | 19.86 |
| | 26–35 | 82 | 54.30 |
| | 36 above | 39 | 25.84 |
| Education | | | |
| | Bachelors | 54 | 35.76 |
| | Postgraduate | 97 | 64.24 |
| Certification | | | |
| | ICAN | 124 | 82.12 |
| | ACCA-UK | 16 | 10.59 |
| | CIMA-UK | 11 | 7.29 |
| Experience | | | |
| | 1–5 years | 30 | 19.86 |
| | 6–10 years | 62 | 41.04 |
| | 11-above years | 59 | 39.1 |
| Industry | | | |
| | Manufacturing | 23 | 15.23 |
| | Banking, finance, insurance | 93 | 61.59 |
| | Service | 35 | 23.18 |
| Sector | | | |
| | Public | 37 | 24.50 |
| | Private | 114 | 75.50 |

potential for CMB. As a result, the Harman single factor technique was applied, which found 35 percent less than the 50 percent requirement. This suggests that the constructs utilized in the study have no common method bias. According to the findings, the data used in the study had no CMB concerns.

**Measurement Models**

To ensure model fit and generate standardized loadings across constructs and items, as well as between each of set of variables, we built a measurement model. Hence, it is important to run a convergent and discriminant validity test prior to estimating values using multiple regression to ensure the appropriateness of the measurement model. From the results (Table 3), construct factor loading is higher than 0.7, Cronbach alpha, and composite reliability (greater than the threshold 0.7) imply strong reliability (Lance et al., 2006). Furthermore, the first-order reflective items composite reliability was robust and far above 0.8 (CR = 0.944), Table 3), showing high-scale dependability.

However, the values of the average variance estimates (AVE) were between 0.55 and 0.65, which were higher than the acceptability limit of 0.5. This indicates that the variations recorded by the questionnaire items were substantially greater than the changes caused by measurement error (Raykov, 2012). The convergent validity of all three constructs was likewise supported, as seen in Table 3. As a result, the underlying concept can account for more than half of the variance in the observed variable (Hulland, 1999).

The correlations among each set of variables remained in the range between 0.27 and 0.45 (Table 4). Any highly correlated constructs higher than 0.90 could indicate a common method bias (Bagozzi et al., 1991). All the relationships in our study are less than 0.90. Therefore, we believe that multiple regression is adequate for the study model.

We further employed the Fornell and Larker AVE metric to examine the discriminant validity. The square root of the average variance estimates (AVE) of the latent variable should be greater than the correlations across dimensions in the model to meet the discriminant validity criteria. The square root of AVE for all constructs (Table 5) is higher than their correlations (Table 4). Hence, discriminant validity was found between the two conceptions. However, all AVE square roots were larger than the correlations among all variables (evidence in Table 3). Hence, the study accepts discriminant validity.

## 5 Empirical Results and Findings

Our study explored the impact of big data on accounting and auditing of accountants in Nigeria. We used multiple regression estimates to test the hypotheses due to the limited number of data sets (Eckstein et al., 2015). First, we examined the influence

**Table 3** Results of convergent and discriminant validity

| Main variables | Mean | Std. Dev. | Factor loadings | Cronbach Alpha | CR | AVE |
|---|---|---|---|---|---|---|
| **Big data *AVE (0.652)*** | **6.295** | **0.305** | | **0.826** | **0.944** | |
| Larger amounts of data, in my opinion, are analysed | | | 0.847 | | | |
| The amount of data we examine is excessive. | | | 0.799 | | | |
| We use a great deal of data, in my opinion | | | 0.802 | | | |
| We are fast in exploring data | | | 0.802 | | | |
| We analyse data quickly | | | 0.802 | | | |
| We analyse different sources of data to gain insights | | | 0.806 | | | |
| We examine data from multitude of sources. | | | 0.806 | | | |
| We use data to improve accounting information quality and ensure transparency<br>We use data to enrich reporting information, and performance management | | | 0.806<br>0.799 | | | |
| **Accounting and auditing practice AVE *(0.659)*** | **6.295** | **0.514** | | **0.886** | **0.875** | |
| **Financial reporting** | **6.217** | 0.433 | | **0.743** | **0.865** | **0.616** |
| We use data to improve accounting information quality. | | | 0.798 | | | |
| We use data to enrich reporting information | | | 0.732 | | | |
| We use data to ensure transparency, and improve accounting information quality | | | 0.789 | | | |
| We use data to improve performance management | | | 0.818 | | | |
| **Performance management** | **6.236** | 0.459 | | **0.751** | **0.863** | **0.558** |
| We use structured data to assess organizational performance | | | 0.723 | | | |
| Big data may supply enormous and diverse customer data | | | 0.790 | | | |
| BDA allows to effectively design customers perspective objectives, measures, targets, and strategies | | | 0.794 | | | |
| BDA gives real-time insights and makes measurements and key performance indicators | | | 0.714 | | | |
| BDA provides business analytics real-time insights | | | 0.711 | | | |
| **Corporate budgeting** | **6.221** | 0.437 | | **0.756** | **0.783** | **0.546** |
| Data analytics predicts models to improve budgeting and forecasting | | | 0.756 | | | |

(continued)

**Table 3** (continued)

| Main variables | Mean | Std. Dev. | Factor loadings | Cronbach Alpha | CR | AVE |
|---|---|---|---|---|---|---|
| Data analytics provides managers with several inputs for budgeting, and allows budget estimations | | | 0.749 | | | |
| Managers could track budget implementation budget in time | | | 0.712 | | | |
| **Audit evidence** | **6.247** | 0.500 | | **0.790** | **0.866** | **0.564** |
| Extend the scope of initiatives and compare them to wider populations. | | | 0.732 | | | |
| Data may be analyzed in larger volumes and faster to provide auditors with relevant insights. | | | 0.730 | | | |
| BDA helps financial auditors streamline the reporting process | | | 0.752 | | | |
| Data analytics helps detect fraud | | | 0.745 | | | |
| Overall, data analytics can aid to collect more suitable and relevant evidence | | | 0.793 | | | |
| **Risk and fraud management** | 6.277 | 0.477 | | **0.749** | **0.884** | **0.559** |
| Data can increase risk monitoring | | | 0.747 | | | |
| Data can enhance risk coverage, and creation of risk decisions making models | | | 0.744 | | | |
| Analytics presents accountants with several chances to improve risk management<br>Data analytics may be used to detect or prevent fraud<br>Big data analytics aid to improve risk assessment, prediction, and measurement | | | 0.749<br>0.748<br>0.750<br>0.751 | | | |

**Table 4** Correlation results

| | CA | AVE | Big | FRep | PMgt. | CBugt. | AEvid. | RFMgt. |
|---|---|---|---|---|---|---|---|---|
| Big data | 0.826 | 0.652 | | | | | | |
| FRep. | 0.743 | 0.616 | 0.338 | | | | | |
| PMgt. | 0.751 | 0.558 | 0.347 | 0.450 | | | | |
| CBugt. | 0.756 | 0.554 | 0.351 | 0.342 | 0.285 | | | |
| AEvid. | 0.790 | 0.564 | 0.356 | 0.352 | 0.352 | 0.325 | | |
| RFMgt. | 0.749 | 0.559 | 0.360 | 0.387 | 0.384 | 0.471 | 0.271 | |

**Table 5** Variables, Cronbach Alpha, and AVE square root

| Variable | Cronbach Alpha | Average variance estimate | Square root AVE |
|---|---|---|---|
| Big data | 0.826 | 0.652 | 0.81 |
| FRep. | 0.743 | 0.616 | 0.78 |
| PMgt. | 0.751 | 0.558 | 0.751 |
| CBugt. | 0.756 | 0.554 | 0.744 |
| AEvid. | 0.790 | 0.564 | 0.751 |
| RFMgt. | 0.749 | 0.559 | 0.75 |

**Table 6** Results of big data, accounting, and auditing relationships

| | FRep | PMgt. | CBugt. | AEvid. | RFMgt. |
|---|---|---|---|---|---|
| Variable | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| *Big data* | 0.345 (0.00)*** | 0.432 (0.00)*** | 0.333 (0.00)*** | 0.378 (0.00)*** | 0.360 (0.00)*** |

Notes: The table presents the results of big data, accounting, and auditing relationships. Big data, FRep, PMgt, CBudgt, AEvid, RFMgt represent big data, Financial Reporting, Performance and Management, Control budgeting, Audit Evidence, Risk and Financial Management, ***, **, * indicate significance at 1%, 5%, and 10% level, the *p*-value is provided in the parathesis

of each data volume, data variety, and data velocity on accounting and auditing practice and explored their effects together.

## 5.1 Results of Big Data, Accounting, and Auditing Relationships

Table 6 provides estimates highlights and empirical findings on the impact of big data on accounting and auditing using the multiple regression model employed. The results indicate that big data is statistically positive and significant in financial reporting ($\beta = 0.345$, $p$-value $= 0.000$). Hence, H1 is approved. Likewise, the use of big data is positive and significant in performance management ($\beta = 0.432$, $p$-value $= 0.000$), confirming H2. Moreover, big data is positive on corporate budgeting ($\beta = 0.333$, $p$-value $= 0.00$), supporting H3, big data is positive and significant on audit evidence ($\beta = 0.378$, $p$-value $= 0.000$), risk and fraud management ($\beta = 0.360$, $p$-value $= 0.000$). These findings support Hypothesis H4-H5 are further discussed in the Conclusions and Discussions sections.

The $R^2$ of endogenous constructions value for financial reporting, performance management, corporate budgeting, audit evidence and risk and fraud management is 0.54, 0.55, 0.51, 0.52, and 0.52, respectively, which indicates that the model is responsible for roughly 54%, 55%, 51%, 52%, and 52% of the volatility in accounting and auditing practices. The summary showing which of the hypotheses were approved is also presented in Table 8.

**Table 7** Results of data variety, variety, velocity, accounting, and auditing relationships

| Variable | FRep<br>Model 1 | PMgt.<br>Model 2 | CBugt.<br>Model 3 | AEvid.<br>Model 4 | RFMgt.<br>Model 5 |
|---|---|---|---|---|---|
| *Volume* | 0.116 (0.00)*** | 0.149 (0.00)*** | 0.103 (0.00)*** | 0.126 (0.00)*** | 0.120 (0.00)*** |
| *Variety* | 0.146 (0.00)*** | 0.142 (0.00)*** | 0.110 (0.00)*** | 0.126 (0.00)*** | 0.120 (0.00)*** |
| *Velocity* | 0.126 (0.00)*** | 0.141 (0.00)*** | 0.120 (0.00)*** | 0.126 (0.00)*** | 0.120 (0.00)*** |

Notes: The table presents the results of data characteristics. Big data, FRep, PMgt, CBudgt, AEvid, RFMgt represent big data, Financial Reporting, Performance and Management, Control budgeting, Audit Evidence, Risk and Financial Management, ***, **, * indicate significance at 1%, 5%, and 10% level, the *p*-value is provided in the parathesis

**Table 8** Hypothesis testing

| Hypothesis | Relationships | Total estimates | Percentage | Prove |
|---|---|---|---|---|
| *H1* | Big→FRep | 0.345 | 0.000*** | Approved |
| *H2* | Big→PMgt. | 0.432 | 0.000*** | Approved |
| *H3* | Big→CBugt. | 0.333 | 0.000*** | Approved |
| *H4* | Big→AEvid. | 0.378 | 0.000*** | Approved |
| *H5* | Big→RFMgt. | 0.360 | 0.000*** | Approved |

Notes: The table presents the hypothesis for the study. Big, FRep, PMgt, CBudgt, AEvid, RFMgt, represent big data, Financial Reporting, Performance and Management, Control budgeting, Audit Evidence, Risk and Financial Management, ***, **, * indicate significance at 1%, 5%, and 10% level, the *p*-value is provided in the parathesis

Table 7 reports highlights of the estimations and empirical evidence from the models employed. From the results, the volume is positive on financial reporting at a significance level of 1%. Similarly, variety is positive and significant at the 1% level in financial reporting, and velocity is positively related and significant in financial reporting. The positive impact suggests that the use of big data will significantly improve the financial reporting of accountants by about 0.345%. The results confirm with Marr (2016) suggestion that implementing big data system may strongly affect firm capacity to timely disclose financial reporting. With regard to performance management, there is evidence of positive and significant impact of volume, variety, and velocity on performance management at significance level of 1%. This suggests that accountants use high levels of data volume, data velocity, and data variety have the best means to assess firm performance. The results, however, confirm with Sardi et al. (2020) who found that big data might help organization attain competitive advantage. Moreover, Elkmash et al. (2021) found that big data analytics lowers the cost of unstructured data analysis for customers and improves the capacity to respond to consumer concerns swiftly. As results, big data can help managers establish the greatest vision and strategy for future occurrences.

From model 3, volume, variety and velocity are positive and significant on corporate budgeting. The positive impact on corporate budgeting indicates high-level use of high levels of data velocity, data volume, and data variety leads 0.333 percent in corporate budgeting of accountants in Nigeria. The result affirms that accountants use the predictive model of large data to improve budget and forecasting (ICAEW, 2020). For example, a large data volume provides accountants and managers with many inputs for budgeting, allowing them to create more accurate budgeting estimations and predictions and hence reduce variances. However, the result is novel and contributes to the extant literature as studies on big data impact on corporate budgeting are still based on theory and dearth empirical evidence (Fisher et al., 2002; De Baerdemaeker & Bruggeman, 2015; Chen et al., 2016).

With respect to model 4, the results discovered a positive and significant effect of data volume, data velocity and data variety on audit evidence at 1% significance level value suggesting that, accountants of Nigeria firms high level use of big data enhance audit evidence at a coefficient of 0.378%. The results confirm with ICAEW (2014) that the use of big data and analytics could help improve the quality and efficiency of auditing. Between big data and audit evidence, there is a consideration convergence, and therefore big data will play an essential role in auditing. Therefore, unique qualities of big data can provide sufficient and accurate audit evidence (Yoon et al., 2015). However, no empirical evidence is provided, and there this finding contributes to the knowledge base.

Finally, the results (model 5) evidence positive and significant impact data volume, velocity, and variety on risk and fraud management at 1% significance level value. The coefficient magnitude indicates that high-level use of data volume, data variety and data velocity will lead increased risk and fraud monitoring at 0.36%. The result confirms with Chen et al. (2015b), who found that Alibaba Group's big data system can monitor and assess fraud threats in real time and send out alerts to prevent fraud. Moreover, the volume, variety, and velocity of high levels of data could help improve risk assessment, prediction, and measurement. For example, volume and variety will supply a large amount of internal, external, financial, and non-financial data in various formats, overcoming the data shortage issue (Table 8).

## 6 Discussion on Results

The results of this study present that while big data significantly impact accounting and auditing of accountants, utilizing the diversity of data (i.e., data volume, data variety, and data velocity) significantly improves it. This indicates that analyzing data from both multiple sources yields economically valuable insights, focusing on swiftly processing data or analyzing large volumes, variety, and velocity does necessarily provide financial benefits for accountants and auditors. The results indicate that big data has significant positive impact on financial reporting. The results confirm Marr (2016), who suggested that the implementation of a big data system has a major effect on firm capacity to provide timely financial reporting to the public. However, the finding supports Warren and Marz (2015) who found that big data can enhance financial reporting and enrich. Moreover, Moffitt and Vasarhelyi (2013) established that big data enrich financial reporting information. The finding suggests that accountants could improve the quality and accuracy of financial reports, especially when big data and continuous analytics is used.

The findings further find a significant positive impact of big data on performance management. The finding is consistent with Sardi et al. (2020), who indicated that big data improve competitive advantage. Besides, ACCA and IMA (2013) asserted that big data used by accountants and finance experts is paramount to examine firm performance. Hence, big data can help managers establish the greatest vision and strategy for future occurrences. Moreover, the results indicate a significant positive impact of big data on corporate budgeting. The results affirm that the more data obtainable and more reliable an organization revenue and expenses, the more effective a static budget is at delivering useful information for decision-making and predict future budgets. However, the result is novel and contributes to the literature, as studies of the relationship between big data and corporate budgeting is still theoretical (De Baerdemaeker & Bruggeman, 2015; Chen et al., 2016). However, the results found a significant and positive impact on the audit evidence. The results support the notion that the accessibility of large amounts of data in various formats and simultaneously, as well as the improved competences of big data analytics, enhances the chances of collecting the most adequate and relevant audit evidence. Finally, the results show that big data has a significant and positive effect on risk and fraud management. The finding is in line with Chen et al. (2015b) who found that Alibaba Group in China big data system can monitor and assess fraud threats in real time and send out alerts to prevent fraud. This suggests that big data can increase risk coverage, risk monitoring, and creation of risk decision-making models, permeating managers to anticipate risk forecasts more precisely (Duan & Xiong, 2015).

To further explore the effect of each big data dimensions on accounting and auditing practices, our study assessed data volume, variety, and velocity when accountants utilized diverse levels of big data dimensions. The findings show that although accountants use high levels of data volume, velocity, and variety regarding their accounting and auditing practices, data variety has the highest means regarding accounting and auditing practices. The result is scholarship (theoretically and practically) significant, with the assumption that one needs to have a farther comprehension of effect of big data on accountants.

## 6.1 Theoretical Contribution

Academics and the literature view big data as a vehicle for the accounting profession (ICAEW, 2014) and have the potential to add value to companies and enhance their performance. However, studies argue that big data is far more than accounting data. Moreover, big data have potential to advance management accounting, financial reporting, financial accounting procedures, and auditing (Iqbal et al., 2020).

Researching the extant literature indicates that preceding studies on big data and accounting are mainly theoretical. Therefore, the empirical study on the effect of big data in accounting is dearth in the literature. However, to the best of our knowledge, there are no empirical studies that investigated the impact of big data on accounting and auditing practice in emerging markets. Moreover, no study has studied Africa. As such, the role of big data utilizing in enhancing accounting and auditing works is not well understood. The gap is what our study examined. To address the study objectives, we surveyed chattered accountants from the African emerging economy, Nigeria to examine the impact of big data on accounting and auditing practice. We make numerous theoretical contributions.

1. We underline the need for accountants and managers adopting big data to publish high-quality information to lessen agency costs and vagueness from an agency theory approach. We illustrated the need to theoretically distinguish between big data dimensions when assessing their effects on accounting and auditing methods, but it might be treated holistically.
2. The results extend to the understanding of the big data literature of the impact of data volume, variety, and velocity on accounting and auditing. Generally, the findings show that each big data dimension might have a different impact on accounting and auditing procedures. Even though data volume, variety, and velocity all have an impact on accounting and auditing, data variety has the most impact. Our findings contribute to the big data literature by examining how each of the big data's three primary characteristics affects accountants and auditors' work.
3. Big data has a large and positive impact on budgeting, as per the results. However, because studies of the interaction between big data and corporate budgeting are still theoretical (De Baerdemaeker & Bruggeman, 2015; Chen et al., 2016), the result is novel and contributes to the literature.
4. A novel contribution of our work to the big data literature is the difference in the influence of data volume, variety, and velocity on accounting and auditing. Our findings represent the first step in determining the effects of big data characteristics on accounting and auditing in Africa's emerging economy.

Furthermore, our study provides a significant theoretical contribution by developing a measurement scale in the context of accounting and auditing. To summarize, this is the first empirical research to examine the effect of big data on accounting and auditing in the African emerging economy. Moreover, this is also the first to empirically examine the relationships in Africa context.

## *6.2 Policy Implications*

The study preceding discussions supplies the following implications. First, big data can help develop accounting and overcome the constraints of numerous accounting procedures in relation to the data. As a result, accountants, prospective accountants, and accounting graduates should hone their competencies in studying and producing big data analytics, which will benefit the industry. Second, the study is important to managers, since it shows how big data represents a hopeful future. Furthermore, accounting teaching bodies have a strong demand for data analysis and data science employment, and there is a lack of such jobs on the job market (Ibrahim et al., 2021). As a result, business institutions of higher learning should create business curriculums that use big data in their offerings. As an outcome of our results, prospective accountants should have a thorough understanding of numerous business matters, as well as a solid understanding of various big data features and how to apply them in accounting and auditing operations. Finally, policymakers can help by establishing governance frameworks for big data to organize its usage and prevent its exploitation.

## 7 Conclusion, Limitations, and Further Studies

The main objective of our study was to close an indispensable gap in the literature concerning the effectiveness of big data on accounting and auditing practice. The study sampled respondents from Nigeria, which is an African emerging economy. Results indicate that big data significantly and positively improves financial reporting, performance management, audit evidence, corporate budgeting, risk, and fraud management of accountants. Moreover, the study found that big data positively and significantly impact risk and fraud management. Interestingly, the effect of data volume, data variety, and data velocity enhances accounting and auditing practices. One of the unique contributions of this study is creating fascinating insights about the empirical impact of big data on accounting when accountants use different characteristics of big data.

Albeit data volume, variety, and velocity could be significant and positively impact accounting and auditing, data variety has the strongest impact. Our results add to the big data literature by investigating how each of the three main dimensions of big data impacts the work of accountants and auditors. These findings assist accountants in using big data analytics to help businesses obtain deeper insight, anticipate future outcomes, and streamline non-routine processes. Furthermore, big data presents prospects for the accounting profession to add value and assist businesses in transforming decision-making in a variety of ways.

There are some potential caveats to this study. First, this study employed a cross-sectional survey to test statistical relations in the proposed study framework. We are calling further studies to employ longitudinal approach as cross-sectional data are inadequate to test the causal relations amid constructs in the study model. Second, we selected participants through the random sample technique. Despite it was considered necessary due to the nature of data received from the Nigerian market, it has caveats in terms of generalizability of the conclusions. Finally, we call for further studies to further validate the results of this study, as our study recruited respondents from Nigeria. Empirical studies from advanced countries would be helpful.

## References


Abedin, M. Z., Guotai, C., Colombage, S., & Moula, F. E. (2018). Credit default prediction by using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27. https://doi.org/10.21314/JCR.2017.233

Abedin, M. Z., Guotai, C., Hajek, P., Tong, Z., & Moula, F. E. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems.* https://doi.org/10.1007/s40747-021-00614-4

Aboud, A., & Robinson, B. (2020). Fraudulent financial reporting and data analytics: An explanatory study from Ireland. *Accounting Research Journal, 35*(1), 21–36.

Disrupt Africa. (2021). *Disrupt Africa*. Accessed from https://disrupt-africa.com/funding-report/

Al-Htaybat, K., & von Alberti-Alhtaybat, L. (2017). Big data and corporate reporting: Impacts and paradoxes. *Accounting, Auditing & Accountability Journal, 30*(4), 850–873.

Alles, M. G. (2015). Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession. *Accounting Horizons, 29*(2), 439–449.

American Institute of Certified Public Accountants (AICPA). (2004). *Audit evidence. Statement on auditing standards no. 106*. AICPA.

Arnaboldi, M., Busco, C., & Cuganesan, S. (2017). Accounting, accountability, social media and big data: Revolution or hype? *Accounting, Auditing & Accountability Journal, 30*(4), 762–776.

Arner, D. W., Barberis, J., & Buckley, R. P. (2015). The evolution of Fintech: A new post-crisis paradigm. *Georgetown Journal of International Law, 47*, 1271.

Association of Chartered Certified Accountants (ACCA) and Institute of Management Accountants (IMA). (2013). *Big data: Its powers and perils*. ACCA. Accessed from http://www.accaglobal.com/content/dam/acca/global/PDF-technical/futures/pol-afa-bdpap.pdf

Bag, S., Wood, L. C., Xu, L., Dhamija, P., & Kayikci, Y. (2020). Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. *Resources, Conservation and Recycling, 153*, 104559.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421–458.

Basukie, J., Wang, Y., & Li, S. (2020). Big data governance and algorithmic management in sharing economy platforms: A case of ridesharing in emerging markets. *Technological Forecasting and Social Change, 161*, 120310.

Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*, 99–113.

Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data's impact on audit judgment and decision making and future research directions. *Accounting Horizons, 29*(2), 451–468.

Bullock, J., Young, M. M., & Wang, Y. F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity, 25*(4), 491–506.

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. *Accounting Horizons, 29*(2), 423–429.

Chen, D. Q., Preston, D. S., & Swink, M. (2015a). How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems, 32*(4), 4–39.

Chen, J., Tao, Y., Wang, H., & Chen, T. (2015b). Big data based fraud risk management at Alibaba. *The Journal of Finance and Data Science, 1*(1), 1–10.

Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big data analytics and big data science: A survey. *Journal of Management Analytics, 3*(1), 1–42.

CIMA (Chartered Institute of Management Accounting). (2008). Budgeting: Topic gateway series no. 27. CIMA. Available at: http://www.cimaglobal.com/Documents/ImportedDocuments/cig_tg_budgeting_mar08.pdf

Cockcroft, S., & Russell, M. (2018). Big data opportunities for accounting and finance practice and research. *Australian Accounting Review, 28*(3), 323–333.

Cokins, G. (2014). Mining the past to see the future: CFOs and their teams can use business analytics to make better decisions. *Strategic Finance, 96*(5), 23–31.

Collier, P. M., & Berry, A. J. (2002). Risk in the process of budgeting. *Management Accounting Research, 13*(3), 273–297.

Council, F. R. (2011). *The UK corporate governance code*. Financial Reporting Council.

De Baerdemaeker, J., & Bruggeman, W. (2015). The impact of participation in strategic planning on managers' creation of budgetary slack: The mediating role of autonomous motivation and affective organisational commitment. *Management Accounting Research, 29*, 1–12.

Duan, L., & Xiong, Y. (2015). Big data analytics and business analytics. *Journal of Management Analytics, 2*(1), 1–21.

Eckstein, D., Goellner, M., Blome, C., & Henke, M. (2015). The performance impact of supply chain agility and supply chain adaptability: The moderating effect of product complexity. *International Journal of Production Research, 53*(10), 3028–3046.

Elmagrhi, M. H., Ntim, C. G., Elamer, A. A., & Zhang, Q. (2019). A study of environmental policies and regulations, governance structures, and environmental performance: The role of female directors. *Business Strategy and the Environment, 28*(1), 206–220.

Elkmash, M. R. M., Abdel-Kader, M. G., & El Din, B. B. (2021). An experimental investigation of the impact of using big data analytics on customers' performance measurement. *Accounting Research Journal.*

Fisher, J. G., et al. (2002). Using budgets for performance evaluation: Effects of resource allocation and horizontal information asymmetry on budget proposals, budget slack, and performance. *The Accounting Review, 77*(4), 847–865.

Gärtner, B., & Hiebl, M. R. (2017). *Issues with big data* (pp. 161–172). Routledge.

Gepp, A., Linnenluecke, M. K., O'Neill, T. J., & Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature, 40*, 102–115.

Ghasemaghaei, M., & Calic, G. (2019). Does big data enhance firm innovation competency? The mediating role of data-driven insights. *Journal of Business Research, 104*, 69–84.

Gleim, & Flesher, D. L. (2015). *Gleim CMA Review, Part. 1: Financial reporting, planning, performance, and control*. GLEIM Publications, Inc.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic Management Journal, 20*(2), 195–204.

Ibrahim, A. E. A., Elamer, A. A., & Ezat, A. N. (2021). The convergence of big data and accounting: Innovative research opportunities. *Technological Forecasting and Social Change, 173*, 121171.

ICAEW (Institute of Chartered Accountants in England and Wales). (2020). *Big data and analytics – what's new?* Chartered Accountants' Hall.

Institute of Chartered Accountants in England and Wales (ICAEW). (2014). *Big data and analytics - What's new?* Chartered Accountants' Hall.

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2020). Big data analytics: Computational intelligence techniques and application areas. *Technological Forecasting and Social Change, 153*, 119253.

Kudyba, S., & Kudyba, S. (2014). *Big data, mining, and analytics*. Auerbach Publications.

Lamboglia, R., Lavorato, D., Scornavacca, E., & Za, S. (2020). Exploring the relationship between audit and technology. A bibliometric analysis. *Meditari Accountancy Research, 29*(5), 1233–1260.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202–220.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2014, May). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Marr, B. (2016). *Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results*. Wiley.

Moffitt, K. C., & Vasarhelyi, M. A. (2013). AIS in an age of big data. *Journal of Information Systems, 27*(2), 1–19.

Petr, H., & Abedin, M. Z. (2020). A profit function maximizing inventory backorder prediction system using big data analytics. *IEEE Access, 8*, 58982–58994.

Raykov, T. (2012). Scale construction and development using structural equation modeling.

Richins, G., Stapleton, A., Stratopoulos, T. C., & Wong, C. (2017). Big data analytics: Opportunity or threat for the accounting profession? *Journal of Information Systems, 31*(3), 63–79.

Sardi, A., Sorano, E., Cantino, V., & Garengo, P. (2020). Big data and performance measurement research: Trends, evolution and future opportunities. *Measuring Business Excellence*. https://doi.org/10.1108/MBE-06-2019-0053

Schmitz, J., & Leoni, G. (2019). Accounting and auditing at the time of blockchain technology: A research agenda. *Australian Accounting Review, 29*(2), 331–342.

Scott, S. V., & Orlikowski, W. J. (2012). Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society, 37*(1), 26–40.

Shajalal, M., Petr, H., & Abedin, M. Z. (2021). Product backorder prediction with deep neural network on imbalance data. *International Journal of Production Research, 2021*, 1.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science, 60*(6), 1452–1469.

Vera-Baquero, A., Palacios, R. C., Stantchev, V., & Molloy, O. (2015). Leveraging big-data for business process analytics. *The Learning Organization, 22*(4), 215–228.

Warren, J., & Marz, N. (2015). *Big data: Principles and best practices of scalable realtime data systems*. Simon & Schuster.

Yang, F., Qiao, Y., Abedin, M. Z., & Huang, C. (2022). Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/TII.2022.3151917

Yoon, K., Hoogduin, L., & Zhang, L. (2015). Big data as complementary audit evidence. *Accounting Horizons, 29*(2), 431–438.

# Part II
# Financial Risk Prediction Using Machine Learning

# Using Outlier Modification Rule for Improvement of the Performance of Classification Algorithms in the Case of Financial Data

**Md. Rabiul Auwul, Md. Ajijul Hakim, Fahmida Tasnim Dhonno, Nusrat Afrin Shilpa, Ashrafuzzaman Sohag, and Mohammad Zoynul Abedin**

**Abstract** This study aims to improve the performance of Data Analytics (DA) algorithms by mining outliers from credit card fraud detection datasets. In doing so, we analyze the performance of data analytics algorithms, such as Linear Discriminant Analysis (LDA), *k*-Nearest Neighbor (*k*-NN), Naïve Bayes (NB) and Support Vector Machine (SVM), by comparing the original and modified datasets in the absence and presence of outliers. To generate modified dataset, this chapter proposes an outlier mining method based on Median (MED) and Median Absolute Deviation (MAD). Performance measures such as accuracy, sensitivity, specificity, detection rate, misclassification error rate, AUC, and pAUC evaluate the performance of the DA algorithms. Empirical findings show that the performance of the DA algorithms on modified dataset shows better results than the original data for both simulated dataset and real-life credit card datasets. This study offers new insights into financial decision makers and stakeholders in the credit card industry.

**Keywords** Financial data · Classification · Outlier detection · Modification

---

M. R. Auwul
Department of Mathematics, Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh

M. A. Hakim
Foreign Exchange and Remittance Department, Travelex Qatar, Golbex Business Center, Doha, Qatar

F. T. Dhonno · N. A. Shilpa · A. Sohag
Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

M. Z. Abedin (✉)
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_5

## 1 Introduction

The competition condition of credit markets has altered since financial technology companies (FinTechs) and giant technology companies (BigTechs) have initiated providing alternative credit (Kowalewski & Pisany, 2022). Since the last two decades, financial institutions have undergone extensive financial technologies which have brought changes in credit provision, information, savings, communication, transactions, and cyber security (Murinde et al., 2022). Machine learning, predictive analytics, data science, and artificial intelligence are all widely used by FinTech organizations to facilitate financial decision-making, as well as eliminate credit default risks (Abedin et al., 2022).

Data science (DS) plays a vital role in managing credit default risk and detecting fraud in financial decisions. DS refers to the process of categorizing a set of tested elements, such as bonds, enterprises, stocks, countries, and so on, into predetermined similar groups (Kulczycki & Franus, 2020). DS algorithms predict credit and fraud risks quickly, helping to increase the efficiency of financial decision-making processes (Hajek & Abedin, 2020). To improve financial decision-making performance, there is a wide range of DS algorithms such as Naïve Bayes (NB), *k*-Nearest Neighbor (*k*-NN), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM) (Li et al., 2018; Abedin et al., 2018a; Chen et al., 2018). Despite DS being one of the key research topics in financial decision support systems, datasets with outliers have a significant impact on the performance of DS classifiers. Note that throughout this book chapter, modeling credit scoring data and credit default risk prediction procedures to support decision-making activities refer to financial decision support systems (FDSSs) (Abedin et al., 2018b).

The existence of abnormal data, for instance, an outlier seriously affects the accuracy of DS performance and other operations (Souiden et al., 2022). The outlier represents the data points in which there are anomalies or errors that affect the data analysis and modeling (Zhang et al., 2021). Outliers occur in numerous ways including omitted variables, data errors, sampling errors, variable construction, and nonnormality (Adams et al., 2019). These problems seriously affect DS models in the field of credit risk forecasting, customer churn prediction, facial recognition, medical diagnosis, speech recognition, and web text classification (Kim, 2017; Ma et al., 2020; Moula et al., 2017; Shen et al., 2018; Xu et al., 2017; Kamishima et al., 2018; Xiao et al., 2019). Therefore, outlier detection, that is, the action of detecting patterns that are significantly different from the data sample, is a vital challenge in machine learning (ML) (Fernández et al., 2022). The outlier in credit cards deals with the fraudulent of clients. Outliers in computer systems indicate fundamental malicious activities. Industrial outlier represents system faults, and medical outliers indicate underlying diseases (Wang & Mao, 2020).

As the outlier affects the performance of the DS models, it generates poor financial decisions in many organizations, including banks and other financial institutions. The financial sector by nature is an intensively data-driven industry, as it manages large quantities of client data. That is why FDSSs such as bankruptcy data, credit data, etc., have the potential to contain outliers (Nyitrai & Miklos, 2019;

Zhang et al., 2021). Outliers in financial decision-making may lead to invalid inferences, model parameter biases, and poor volatility prediction (Granea & Veiga, 2010). Therefore, the detection of outliers is an important concern for the detection of rare objects in real application domains, such as in finance, materials science, health, and industry (Ma et al., 2020). Outlier detection is a technique that improves the efficiency of FDSSs and exhibits a huge difference from other financial decisions (Cai et al., 2020). Outlier detection intends to detect fraud and money laundering of financial decision systems by finding unusual customer behavior patterns (Jun, 2006). Moreover, outlier detection in credit scoring domain helps to reduce subjective elements in detecting outliers, eliminate the required time and effort, and enhance the effectiveness of FDSSs (Okada et al., 2013; Yang et al., 2022). Besides detection, outlier modification should also be taken seriously, because modification helps preserve useful information at the time of modeling FDSSs (Granea & Veiga, 2010).

The presence and absence of outliers are the causes that differ the performance of DS algorithms. More specifically, it is estimated that the performance of DS algorithms may vary when there are 0%, 10%, and 20% outliers. The presence of outliers can lead to destructive effects on the performance of DS algorithms if these are not detected and modified precisely (Liu et al., 2021). Therefore, detection and modification of these outliers are the primary steps to generate more stability of DS algorithms. This study applies simulated data to see the performance of DS algorithms in the presence and absence of outliers before and after modification. But simulated data are often unable to reflect the present situation of corresponding domains, which leads to unreliable and unrealistic reaction of people to the simulation. That is why, to reduce prediction bias and enhance stability as well as effectiveness of DS algorithms, the adaptation of real-life datasets (FDSS) is important. For these reasons, this study uses credit scoring data as FDSS data.

In modeling FDSSs, this study analyzes the performance of four DS algorithms such as LDA, k-NN, NB, and SVM by comparing the original and modified datasets. Following the study by Nyitrai and Miklos (2019), this study trains multiple DS classifiers to enhance the stability and minimize the forecast bias of the decision support system. The modified dataset originates from the original data by applying an outlier detection and modification model based on Median (MED) and Median Absolute Deviation (MAD). The current study applies both simulated and real-life datasets to train the model. Real-life data refers to FDSSs data which come from Credit Scoring Default Datasets. For measuring the performance of the DS algorithms in absence (0) and presence of 10 and 20 outliers on original and modified datasets, this study uses Accuracy, Sensitivity, Specificity, Detection rate (DR), AUC, and pAUC. Our study makes notable contributions to DS performance and FDSSs. This study extends the existing literature by comparing the performance of DS algorithms on original and modified datasets. This paper informs stakeholders that the detection and modification of outliers is important to improve the performance of DS algorithms and financial decision-making. The current study suggests that policymakers to motivate stakeholders to detect and modify outliers precisely because an outlier-free dataset can result in a precise financial decision. This study also motivates financial decision makers to improve the performance of applied DS algorithms while making financial and managerial decisions.

The paper proceeds as follows. In Sect. 2, we present a review of the related literature. Section 3 briefly describes the proposed methodology along with applied data science methods. The results and discussions are presented in Sect. 4. Finally, Sect. 5 concludes the paper with further road maps.

## 2 Related Literature

DS plays an important role in improving the performance of FDSS. Regarding existing studies, Wang and Mao (2020) develop a dynamic ensemble outlier detection model to generate a base classifier, determine the validation set, and estimate the competence by using k-NN. Abedin et al. (2018a, b) utilize the topological applications of support vector machines (SVMs) and multilayer perceptrons (MLPs) to confirm the competitive performance of statistical intelligence mechanisms. Their study deals with bankruptcy prediction and credit scoring in eight different databases to assess FDSS. Li et al. (2022) offer a Fisher LDA classification method attached with Naïve Bayes (B-FLDA) for the event-related potential-based brain-computer interface (ERP-BCI) to concurrently recognize the works, intentions, and idle states of subject intentions.

Decision-making in banking and finance is now comparatively more complex than in previous decades. One of the factors influencing financial and banking decisions is the existence of outliers. Leontitsis and Vorlow (2006) use the surrogate data analysis (SDA) technique to deal with outliers which have an impact on stock return. Their approach is based on the scale parameters of mean-stationary time series and robust estimation of location. The study of Shen et al. (2018) determines the effect of outliers on the relationship between financial development and economic growth. To conduct the study, they used a dynamic panel model by collecting data from 48 countries between 1988 and 2014. To determine the effect of different levels of outliers on the positive-valued insurance dataset, Okhli and Nooghabi (2021) develop the contaminated exponential distribution as an alternative platform.

Detecting outliers is a vital phase in evaluating the impact of outliers in empirical finance research. Adams et al. (2019) employ a multivariate identification strategy to identify and treat outliers appropriately in financial data. To successfully detect the financial crisis, Domino (2020) introduces fourth-order multivariate cumulate method as an outlier detection algorithm. Granea and Veiga (2010) applied a wavelet-based general detection and correction method to detect isolated outliers and outlier patches when modeling financial time series data. Okada et al. (2013) propose a case model to detect financial outliers of the hospital industry, which helps to reduce the required time and effort and enhance the quality of analysis. Based on distance, Jun (2006) develops a cross-outlier detection model to detect outliers of financial transaction data. To minimize the negative impacts of outliers in the noise-filled credit datasets, Zhang et al. (2021) propose a novel multistage ensemble model with enhancing outlier adaption.

Based on the literature cited above, the present study determines the following research gaps. First, there is a range of studies dealing with the performance of DS algorithms in the presence of outliers (Wang & Mao, 2020; Ling et al., 2020), but none of them analyze the performance of DS algorithms by considering the absence and presence of outliers, especially in FDSSs. Second, most studies apply MED or MAD to detect outliers from datasets (Leys et al., 2013; Park & Moon, 2015; Abbas, 2019). That means previous studies are unable to demonstrate the performance of outlier detection and modification by combining both MED and MAD.

To fill in the above research gaps, this study provides significant theoretical contributions to the existing literature on DS and FDSSs. First, this study extends to previous studies by investigating the performance of DS algorithms by comparing the original and modified datasets in the absence (0) and presence of 10% and 20% outliers of FDSSs. Second, this paper combines MED and MAD as an outlier detection and modification algorithm in financial decision-making.

## 3 Materials and Methods

To evaluate the performance of different DS algorithms for binary classification, this chapter applies the Receiving Operating Characteristics (ROC) curve, the area under the ROC curve (AUC), and other classification measures as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}), \tag{1}$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \tag{2}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}), \tag{3}$$

$$\text{Detection Rate} = \text{TP}/(\text{TP} + \text{FP} + \text{TN} + \text{FN}), \tag{4}$$

$$\text{MER} = 1\text{–Accuracy}, \tag{5}$$

where TP, FP, TN, and FN are the numbers of True Positive, False Positive, True Negative, and False Negative, respectively. MER, AUC, and pAUC are the misclassification error rate, area under the ROC curve, and partial area under the ROC curve, respectively.

On the basis of these parameters, this chapter declares a method as a good performer if it produces larger values of Accuracy, Specificity, Sensitivity, Detection Rate, AUC, and pAUC and low values of MER.

## *3.1 Statistical Methods to Be Compared*

In this study, four popular classification algorithms are evaluated, namely LDA, *k*-NN, Naïve Bayes and SVM.

**Linear Discriminant Analysis**

LDA is a dimensionality reduction approach that is used as a pre-processing step to classify patterns. LDA aims to design the features of higher dimensions of a space on a lower dimension space to reduce resources and dimensional cost (Treder et al., 2016). LDA represents a general discriminant function that uses a linear decision boundary. For example, the target data instance $z$ is classified by solving the discriminant function $d_j$ for each $K_j$ class with the classification rule $S_j$. Let, the prior probabilities is $p(K_j)$, the mean of each class is $c_j$, and the common covariance matrix is $\text{cov}_w$. Then, the discriminant function is defined as follows:

$$d_j(z) = \log\left(p\left(K_j\right)\right) - \frac{1}{2}c_j^T \text{cov}_w^{-1} c_j + z^T \text{cov}_w^{-1} c. \tag{6}$$

The classification rule for the target data instance is defined as:

$$S_k(Z) = j** :\Leftrightarrow j** = \underbrace{\arg\max}_{j} d_j(Z). \tag{7}$$

**k-Nearest Neighbors (k-NN)**

*k*-NN is a widely used machine learning algorithm that is utilized in numerous applications. *k*-NN is based on the assumption that the prediction value of the example is probably the same as those of neighbors Jang et al. (2020). The *k*-NN algorithm explains a metric in the predictor vector space, plots all applicants to a position in this space, and evaluates posterior probability through the relative amount of good risks between the *k*-nearest points in the training set.

Suppose $Z_j$ are the feature values, and $K_j$ denotes the labels of $Z_j$ for each $j$. Let the number of classes be $n$ and $z$ be the points for which the label is not known. To find the classes for unknown labels using *k*-NN, $d(z, Z_j)$, $j = 1, 2, \ldots, n$ first must be determined for all values of $k$ ($d$ is a distance metric). Second, the distances are determined for all $n$, the values are arranged in increasing order, and the distances are taken from the sorted list ($D \geq 0$). Third, $D$ points are found that correspond to the $D$ distances. In the fourth step, let $D_j$ represent the number of data points belonging to the $j$th class. In the fifth step, put $x$ in class $i$ if $D_j > D_i$, $\forall j \neq i$.

**Naïve Bayes (NB)**

The NB classifier is a probabilistic algorithm that is used for solving classification tasks based on the Bayes Theorem, where the independence of features is assumed. The NB classifier is widely applied in the data mining and product review sentiment classifications domains (Xu et al., 2020).

Let $z$ be a class variable that needs to be predicted and $x_1, x_2, \ldots, x_n$ are features, then according to the Bayes Theorem, the probability of obtaining classes for $z$ based on $x$'s is:

$$P(z|x_1, x_2, \ldots, x_n) = \frac{P(x_1|z)P(x_2|z)\ldots P(x_n|z)P(z)}{P(x_1)P(x_2)\ldots P(x_n)}. \tag{8}$$

As the denominator is unchangeable and the features are independent, the denominator can be removed, and the result is proportionally given as:

$$P(z|x_1, x_2, \ldots, x_n) \infty P(z)\prod\nolimits_{i=1}^{n} P(x_i|z). \tag{9}$$

So, the class is obtained by finding the maximum probability as follows:

$$z = \underbrace{\arg\max}_{z} P(z)\prod_{i=1}^{n} P(x_i|z). \tag{10}$$

**Support Vector Machine (SVM)**

SVM refers to a machine learning model that is used to fix pattern recognition problems such as outlier detection, classification, and regression. It utilizes the idea of decision planes that apply decision boundaries to optimally distinct data into numerous categories (Huang et al., 2021). The main objective of SVM is to find the hyper plane that classifies the classes accurately with the maximum margin. The linear SVM formula is given below. Suppose $X$ are the features and $z$ are the target values that need to be predicted. Then predict $z$ as a function of the weighted values of $X$. The Hinges loss function with a regularization term is defined as:

$$\text{Total cost} = \|\omega\|^2/_2 + K^*. \tag{11}$$

That is, the total cost is the sum of all losses for each observation. Here, $\omega$ denotes the weight value, and $K$ is the hyperparameter that controls the amount of regularization. If $K$ is sufficiently small, this indicates a hard-margin classifier, while for large $K$ we obtain a soft-margin classifier.

## 3.2 *Proposed Method*

The current chapter proposes a novel methodology by combining MED and MAD as an outlier mining (detection and modification) method to evaluate the performance of data analytics algorithms. This paper considers 0%, 10%, and 20% outliers to assess how machine learning algorithms perform on original and modified datasets at

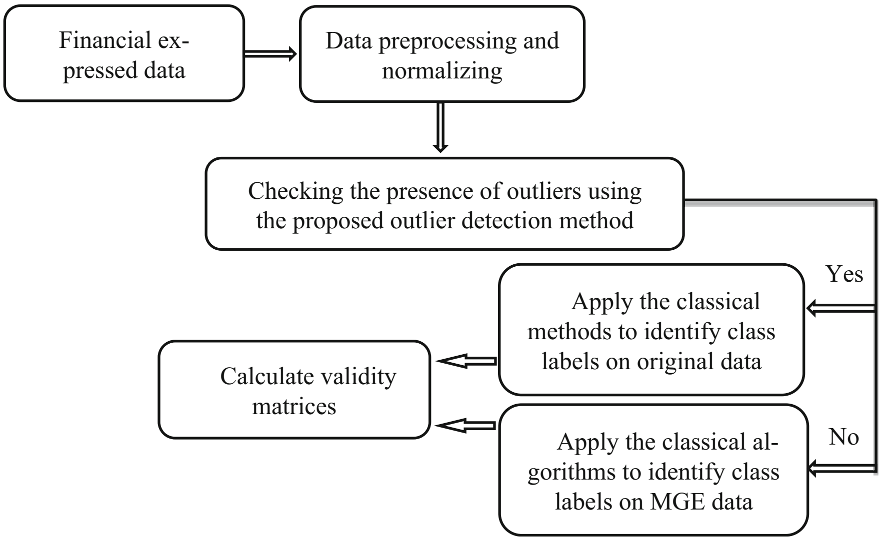



**Fig. 1** Flow chart of the proposed outlier modification algorithm

different levels of outlier existence. MED and MAD are the robust estimator of location and robust measure of dispersion, respectively.

For this study, suppose that $z_{ijk}$ is the $i$th data instance for the $j$th that replicates in the $k$th class and $E(z_{ijk}) = \mu_{ik}$, $\mathrm{var}(z_{ijk}) = \sigma^2{}_{ij}$ represent the mean and variance value for the $i$th data instance and the $k$th class ($i = 1, 2, \ldots, p; j = 1, 2, \ldots, n_k; k = 1, 2, \ldots, m$), respectively. Based on this concept, we propose the following outlier modification rule:

(i) This chapter classifies an expression of a sample as an outlier, if it does not fall in the interval $[\mathrm{MED}_{i,(k)} - L^*\mathrm{NMAD}_{i,(k)}, \mathrm{MED}_{i,(k)} + L^*\mathrm{NMAD}_{i,(k)}]$. We declare the corresponding sample as an outlying sample. Here, $L = 3$ (for our study), $\mathrm{MED}_{i,(k)} = \mathrm{median}(z_{ij,(k)})$; $i = 1, 2, \ldots, p; j = 1, 2, \ldots, n_k; k = 1, 2, \ldots, m$) are the median expressions of the $i$th data instance in the $k$th class, $\mathrm{MAD}_{i,(k)} = \mathrm{median}_{j = 1,2,\ldots,nk}(|z_{ij,(k)}) - \mathrm{MED}_{i,(k)}|)$ is the median absolute deviation and $\mathrm{NMAD}_{i,(k)} = \mathrm{MAD}_{i,(k)}/0.6745$ is the normalized $\mathrm{MAD}_{i,(k)}$ of the $i$th instance in the $k$th group.

(ii) For each sample from each group ($k = 1, 2, \ldots, m$), check separately the presence of outliers using Step 1. If an outlier is present, then replace it by the median of the respective group $[\mathrm{MED}_{i,(k)}]$, and get our desired modified financial expression (MFE) data.

(iii) Finally, apply the classical methods (DS Algorithms) in the MFE data to identify the class label and finding different indices measurement such as accuracy, sensitivity, specificity, detection rate, misclassification error rate, AUC and pAUC.

The flow chart of the proposed outlier modification algorithm is depicted in Fig. 1.

# 4 Results

This section illustrates the results of credit card fraud detection by using four DA algorithms such as LDA, *k*-NN, NB, and SVM. All experiments were carried out on a simulated dataset and three real-life credit card fraud datasets. Performance was evaluated by comparing the original and modified datasets. This study utilized R packages for these algorithms: class, caret, ROC, kkNN, e1071, and rpart. To judge the performance of these algorithms, we used the MASS R package. The comprehensive R archive network (cran) or Bio-conductor are the main sources of these packages. In this chapter, the terms "proposed" and "classical" refer to the application of four traditional methods in the proposed and original MFED datasets, respectively.

## 4.1 Simulated Data Analysis

Simulated data were generated for two groups ($k = 2$) with known characteristics both in the presence of 0%, 10%, and 20% outliers that mimic the nature of real-life credit card data modeling scenarios. This study uses a data generation model that is described in Table 1. In Table 1, the row represents the feature, and the column represents the sample groups. For randomization, this study adds Gaussian noise to the datasets. The generated data contains $p = 1{,}000$ features consisting of two groups ($P1 = P2 = 500$) with sample size $n = 10$. We set the value of the parameter $d$ as 0.2 and the noise parameter, $\sigma^2 = 0.05$ to generate datasets for each of the data types.

This study generates 100 datasets from the data generating model as presented in Table 1. The performance of four DA algorithms (LDA, *k*-NN, NB, and SVM) was evaluated by comparing the original and modified datasets with two groups ($k = 2$). This study also evaluates the performance of these methods in the presence and absence of outliers. To generate outlier datasets, this study randomly selects a dataset containing 0%, 10%, and 20% outlier and replaces it with Gaussian noise with mean 60 and variance 3, respectively. This study measures different percentage of outlier features (10% and 20%) with randomly choosing one or two samples. This study computes different performance measures such as accuracy, sensitivity, specificity,

**Table 1** Matrix used to generate simulation study

| | Sample | | | |
|---|---|---|---|---|
| | $S1$ | $S2$ | $S3$ | Gaussian noise |
| *Group-1*($p_1$) | $-a-d$ | $-a+d$ | + d | $+N(0,\ \sigma^2)$ |
| *Group-2*($p_2$) | $a-$ d | a + d | $-d$ | |

**Table 2** Performance evaluation of four classifiers based on original and modified training dataset for simulated data

| Data structure | Validity matrices | Classical algorithms | | | | Proposed algorithms | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LDA | *k*-NN | NB | SVM | LDA | *k*-NN | NB | SVM |
| In absence of outliers | Accuracy | 0.977 | 0.941 | 0.977 | 0.968 | 0.977 | 0.941 | 0.977 | 0.968 |
| | Sensitivity | 0.977 | 0.943 | 0.976 | 0.968 | 0.977 | 0.943 | 0.976 | 0.968 |
| | Specificity | 0.977 | 0.938 | 0.979 | 0.968 | 0.977 | 0.938 | 0.979 | 0.968 |
| | Detection rate | 0.977 | 0.943 | 0.976 | 0.968 | 0.977 | 0.943 | 0.976 | 0.968 |
| | AUC | 0.997 | 0.984 | 0.997 | 0.995 | 0.997 | 0.984 | 0.997 | 0.995 |
| | pAUC | 0.198 | 0.186 | 0.198 | 0.195 | 0.198 | 0.186 | 0.198 | 0.195 |
| In the presence of 10% outliers | Accuracy | 0.495 | 0.940 | 0.500 | 0.500 | 0.976 | 0.957 | 0.977 | 0.966 |
| | Sensitivity | 0.573 | 0.938 | 0.550 | 0.550 | 0.976 | 0.961 | 0.975 | 0.968 |
| | Specificity | 0.417 | 0.942 | 0.450 | 0.450 | 0.976 | 0.953 | 0.979 | 0.964 |
| | Detection rate | 0.573 | 0.938 | 0.550 | 0.550 | 0.976 | 0.961 | 0.975 | 0.968 |
| | AUC | 0.516 | 0.982 | 0.724 | 0.598 | 0.997 | 0.966 | 0.997 | 0.994 |
| | pAUC | 0.046 | 0.184 | 0.088 | 0.055 | 0.197 | 0.178 | 0.197 | 0.195 |
| In the presence of 20% outliers | Accuracy | 0.500 | 0.928 | 0.500 | 0.500 | 0.976 | 0.946 | 0.976 | 0.962 |
| | Sensitivity | 0.750 | 0.940 | 0.650 | 0.550 | 0.977 | 0.957 | 0.978 | 0.965 |
| | Specificity | 0.250 | 0.915 | 0.350 | 0.450 | 0.974 | 0.936 | 0.973 | 0.959 |
| | Detection rate | 0.750 | 0.940 | 0.650 | 0.550 | 0.977 | 0.957 | 0.978 | 0.965 |
| | AUC | 0.559 | 0.977 | 0.673 | 0.570 | 0.996 | 0.969 | 0.997 | 0.993 |
| | pAUC | 0.070 | 0.197 | 0.048 | 0.175 | 0.194 | 0.062 | 0.179 | 0.197 |

detection rate, AUC, and pAUC for each of the 100 datasets using the seven DA algorithms. Then, this paper determines the average of these performance measures for each of the data types.

For creating 100 Modified Financial Expressed Datasets (MFED), this chapter first applies the proposed outlier modification technique for 100 training datasets. The value of validity matrices such as accuracy, sensitivity, specificity, detection rate, AUC, and pAUC are averaged over 100 datasets that are obtained from MFED datasets. These average performance values are summarized in Table 2. We perceived that in absence of outlier all four classifiers (LDA, *k*-NN, NB, and SVM) produce same results using original data and proposed modified training dataset. Nevertheless, in the presence of 10% and 20% outliers, the four classifiers performed much better using modified training data than original training data. For instance, the average accuracies 0.976, 0.957, 0.977, and 0.966 are produced by LDA, k-NN, NB, and SVM, respectively, in the presence of outliers in each of 10% outliers that are larger than 0.495, 0.940, 0.500, and 0.500, those were produced by the classical classifiers in the same condition. The average accuracies 0.976, 0.946, 0.976, and 0.962 are produced by LDA, k-NN, NB, and SVM, respectively, in the presence of outliers in each of 20% outliers that are larger than 0.500, 0.928, 0.500, and 0.500, those were produced by the classical classifiers in the same condition. Hence, we

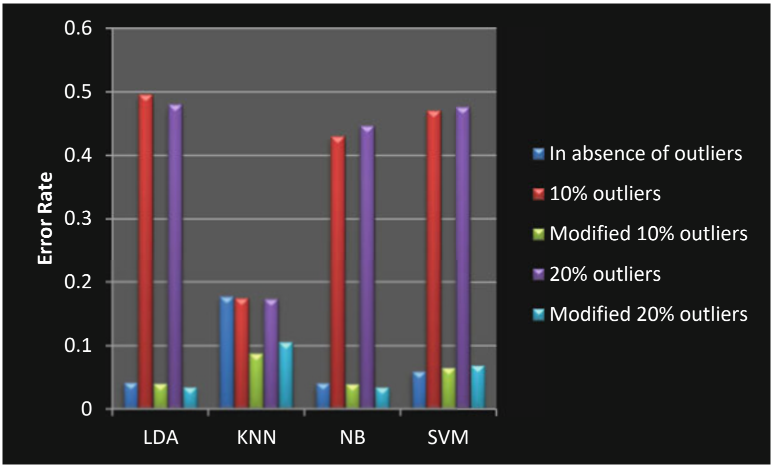



**Fig. 2** Performance evaluation using the average value of the error rate

perceived that the performance of the classifiers improves by using MFED datasets instead of the original datasets.

The bar plot of the average value of error rate is presented in Fig. 2. From this plot, this chapter determines that the error rate is approximately similar for both classical and proposed algorithms in the absence of outliers (0%). But in case of 10% and 20% outliers, error rate is raised for classical algorithms and the values are getting lower for using MFED datasets.

In Fig. 3a, b, this study represented the box plot of the accuracies for 100 datasets for 10% and 20% outlying datasets including original datasets for both classical and proposed algorithms. Figure 3 shows that for this simulation study, the performance of the popular DS algorithms improves when the training datasets are modified by the proposed method in the presence of outliers. Otherwise, these DS algorithms produce the same results on original datasets.

### *4.2 Credit Card Default Data (CCDD)*

To examine the performance of the four well-known DS algorithms (LDA, *k*-NN, NB, and SVM), this study generated training and test datasets by randomly partitioning (70% training and 30% test) the whole CCDD dataset into two independent datasets. The log-transformed dataset was considered to remove unusual or extreme values in this study. First, the training CCDD dataset was used in the proposed outlier modification procedure to obtain a modified training dataset as described above. Thereafter, the performance of DS algorithms was determined based on performance measures such as accuracy, sensitivity, specificity, detection rate, and misclassification error rate (MER) on CCDD datasets. Table 3 shows the

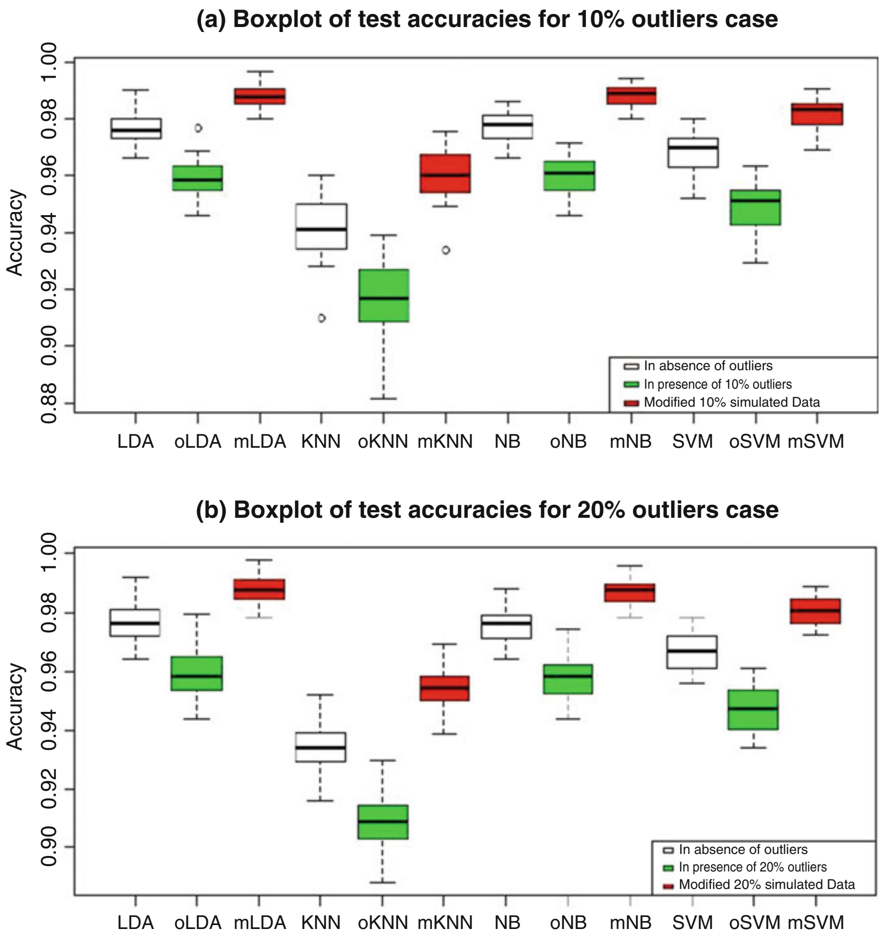



**Fig. 3** Performance evaluation of four classifiers using boxplot (**a**) in presence of 10% outliers (**b**) in presence of 20% outliers

average accuracy value using 100 simulations. The results indicate that all four DS algorithms (LDA, *k*-NN, NB, and SVM) produce similar performance to those for the original CCDD training dataset. On the contrary, these DS algorithms performed far better on the modified CCDD datasets. For example, LDA produces accuracy = 0.791 for the modified CCDD dataset, which is better than accuracy = 0.768 using the original CCDD dataset. Figure 4a represents the boxplot of test values.

**Table 3** Performance evaluation of four classifiers based on the original and modified training dataset for real credit default datasets

| Data | Measure | Original data | | | | Modified data | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LDA | *k*-NN | NB | SVM | LDA | *k*-NN | NB | SVM |
| Default data | Accuracy | 0.768 | 0.769 | 0.767 | 0.768 | 0.791 | 0.770 | 0.769 | 0.776 |
| | Sensitivity | 0.999 | 0.998 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| | Specificity | 0.001 | 0.002 | 0.001 | 0.053 | 0.062 | 0.413 | 0.018 | 0.001 |
| | Detection rate | 0.999 | 0.998 | 0.982 | 0.999 | 0.904 | 0.997 | 0.999 | 0.990 |
| | MER | 0.232 | 0.231 | 0.233 | 0.232 | 0.209 | 0.230 | 0.231 | 0.224 |
| Taiwan credit default data | Accuracy | 0.735 | 0.773 | 0.478 | 0.782 | 0.775 | 0.779 | 0.624 | 0.817 |
| | Sensitivity | 0.919 | 0.990 | 0.459 | 0.990 | 0.990 | 0.999 | 0.579 | 0.959 |
| | Specificity | 0.087 | 0.010 | 0.544 | 0.078 | 0.013 | 0.002 | 0.579 | 0.320 |
| | Detection rate | 0.918 | 0.990 | 0.489 | 0.990 | 0.990 | 0.999 | 0.637 | 0.959 |
| | MER | 0.265 | 0.227 | 0.522 | 0.212 | 0.225 | 0.221 | 0.376 | 0.183 |
| PAK credit default data | Accuracy | 0.739 | 0.738 | 0.715 | 0.738 | 0.739 | 0.739 | 0.738 | 0.739 |
| | Sensitivity | 0.999 | 0.999 | 0.937 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| | Specificity | 0.001 | 0.001 | 0.087 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Detection rate | 0.999 | 0.999 | 0.937 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| | MER | 0.261 | 0.262 | 0.285 | 0.262 | 0.261 | 0.261 | 0.262 | 0.261 |

### 4.3 *Taiwan Credit Default Data*

As in the same procedure as in the previous subsection, the entire Taiwan credit dataset was divided into two independent datasets. To remove the unusual or extreme values in this dataset, the log-transformed Taiwan dataset was considered in this study. Firstly, the training Taiwan dataset was used in the proposed outlier modification procedure to obtain the modified training dataset as described above. Thereafter, accuracy, sensitivity, specificity, detection rate, and MER were measured using test Taiwan datasets. Table 3 summarizes the average values of accuracy over 50 simulations. Table 3 shows that all four classifiers (LDA, *k*-NN, NB, and SVM) produce slightly better results using the modified Taiwan dataset than the original one. For example, LDA produces an accuracy of 0.775 using the modified training Taiwan dataset, which is greater than the accuracy of 0.735 using the original training Taiwan credit dataset. Figure 4b represents the test accuracy values, supporting the results in Table 3.

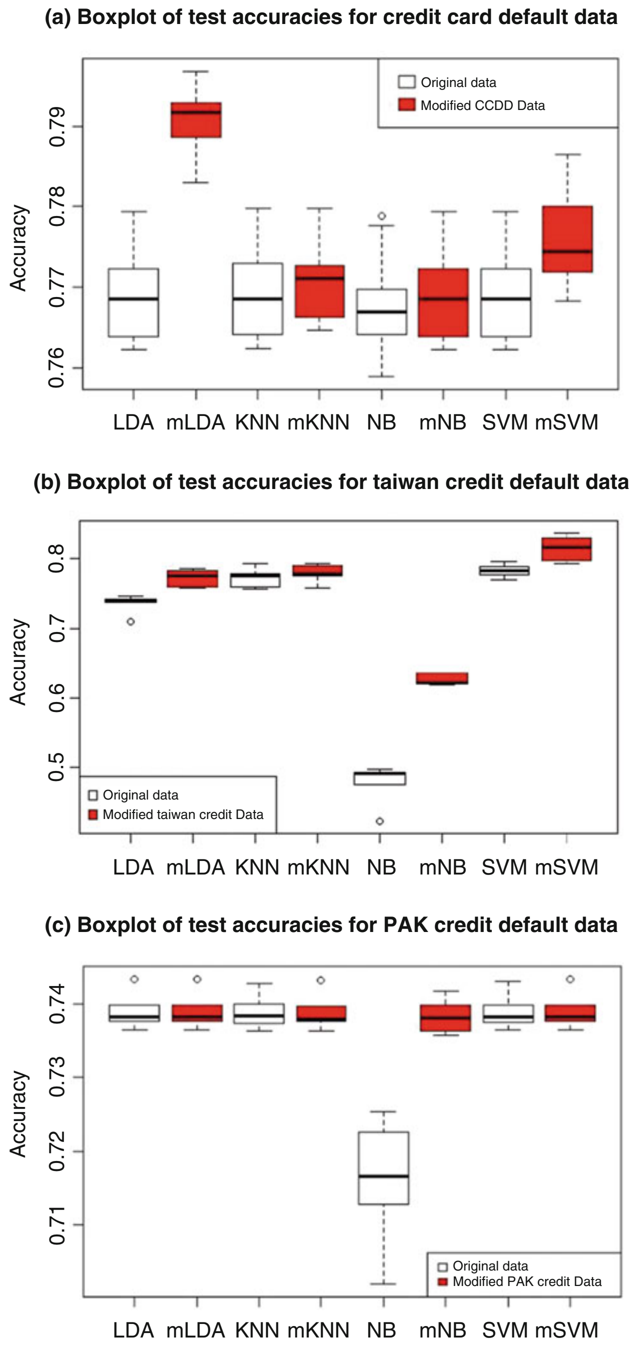



**Fig. 4** Performance evaluation of four classifiers using (**a**) CCDD dataset (**b**) Taiwan credit dataset, and (**c**) PAK credit default dataset

### *4.4 PAK Credit Default Data*

Again, the whole PAK credit dataset was divided into two independent datasets, and the log-transformed PAK credit dataset was used. Firstly, the PAK training credit dataset was used in the proposed procedure to obtain the modified dataset. Thereafter, accuracy, sensitivity, specificity, detection rate, and MER were measured using test PAK credit datasets. Table 3 summarizes the average accuracies for 50 simulations. From Table 3, notice that all four classifiers (LDA, *k*-NN, NB and SVM) produce almost equal results using both the original PAK credit training dataset and the modified PAK credit datasets except NB classifiers that gave better result for the modified data than the original data. For example, NB produces accuracy $= 0.738$ using the modified PAK credit dataset, which is greater than accuracy $= 0.715$ using the original training PAK credit dataset. The box plot of test accuracies is presented in Fig. 4c.

Table 3 summarizes the average values of the performance criteria estimated for three well-known financial datasets by different algorithms, respectively. We reconnoiter similar interpretations like boxplots based on this table. We also perceived that the proposed method produces almost parallel values of performance measures. Therefore, we may conclude that the performance of the proposed algorithms improved substantially over the performance of the classical algorithms.

## 5 Discussion

This is the first study, as far as we know, that applies outlier mining-based data analytics approaches in predicting credit card fraud. This chapter compares the results and findings with some recently published papers. For example, Carcilloa et al. (2021) apply hybrid unsupervised and supervised learning to detect credit card fraud. Their results illustrate that the combined approach is more workable than the baseline methods. Carneiro et al. (2017) develop a data mining-based methodology to assess credit card default for an electronic merchant. They also state that a combination of automatic and manual intelligent methodology offers feasible insights. Vlasselaer et al. (2015) apply the data mining methodology and explain that intrinsic and network-based features produce the most optimum results in predicting credit card fraud customers. Bhattacharyya et al. (2011) also applied data mining-based approaches to detect credit card fraud. They conclude that traditional SVM, RF, and LR generate optimum prediction results than others. By comparing and contrasting the results of other studies with ours, we can assert that none of the existing studies covers outlier mining-based data analytic approaches in predicting financial status of credit card users as does this study.

## 6 Conclusion

One of the major objectives of DS algorithms is to extract knowledge from large amount of data. In the literature, there exist many algorithms to perform this task. However, it should be noted that most of them provide vague results in the presence of outliers. Therefore, in this chapter, an outlier detection method and a modification rule were proposed to improve the classification performance of several classification algorithms (LDA, *k*-NN, Naïve Bayes, and SVM). The performance of the proposed methods was evaluated using both simulated and real financial datasets. The results indicate that all classification algorithms produce misleading results in the presence of outliers. However, their performance improved substantially when using the proposed MFE data both for small and large datasets. From the data analysis of the CCDD, Taiwan credit default, and PAK credit default tasks, we confirmed the effectiveness of the proposed method under real conditions.

## References


Abbas, N. (2019). A robust S2 control chart with Tukey's and MAD outlier detectors. *Quality and Reliability Engineering International, 36*(1), 403–413. https://doi.org/10.1002/qre.2588

Abedin, M. Z., Guotai, C., Moula, F. E., Azad, A. S. M., & Khan, M. S. U. (2018a). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*(1), 474–507. https://doi.org/10.1002/ijfe.1675

Abedin, M. Z., Guotai, C., Colombage, C., & Moula, F. E. (2018b). Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27. https://doi.org/10.21314/JCR.2017.233

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems, 1–21*. https://doi.org/10.1007/s40747-021-00614-4

Adams, J., Hayunga, D., Mansi, S., Reeb, D., & Verardi, V. (2019). Identifying and treating outliers in finance. *Financial Management, 48*(2), 345. https://doi.org/10.1111/ma.12269

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Cai, S., Li, L., Li, S., Sun, R., & Yuan, G. (2020). An efficient approach for outlier detection from uncertain data streams based on maximal frequent patterns. *Expert Systems with Applications, 160*, 113646. https://doi.org/10.1016/j.eswa.2020.113646

Carcilloa, F., Borgnea, Y. L., Caelen, O., Kessaci, Y., Obleb, F., & Bontempia, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences, 557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems, 95*, 91–101. https://doi.org/10.1016/j.dss.2017.01.002

Chen, W., Zhang, S., Li, R., & Shahabi, H. (2018). Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the Total Environment, 644*, 1006–1018. https://doi.org/10.1016/j.scitotenv.2018.06.389

Domino, K. (2020). Multivariate cumulants in outlier detection for financial data analysis. *Physica A, 558*, 124995. https://doi.org/10.1016/j.physa.2020.124995

Fernández, Á., Bella, J., & Dorronsoro, J. R. (2022). Supervised outlier detection for classification and regression. *Neurocomputing, 486*, 77–92.

Granea, A., & Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Computational Statistics and Data Analysis, 54*, 2580–2593. https://doi.org/10.1016/j.csda.2009.12.010

Hajek, P., & Abedin, M. Z. (2020). A profit function-maximizing inventory backorder prediction system using big data analytics. *IEEE Access, 8*, 58982–58994. https://doi.org/10.1109/ACCESS.2020.2983118

Huang, J., Jin, T., Liang, M., & Chen, H. (2021). Prediction of heat exchanger performance in cryogenic oscillating ow conditions by support vector machine. *Applied Thermal Engineering, 182*, 116053. https://doi.org/10.1016/j.applthermaleng.2020.116053

Jang, S., Jang, Y. E., Kim, Y. J., & Yu, H. (2020). Input initialization for inversion of neural networks using k-nearest neighbor approach. *Information Science, 519*, 229–242. https://doi.org/10.1016/j.ins.2020.01.041

Jun T. (2006, April 9). A cross datasets referring outlier detection model applied to suspicious financial transaction discrimination. In *Intelligence and Security Informatics, International Workshop, WISI 2006, Proceedings*, Vol. 3917, pp. 58–65. https://doi.org/10.1007/11734628_7

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2018). Model-based and actual independence for fairness-aware classification. *Data Mining and Knowledge Discovery, 32*, 258–286. https://doi.org/10.1007/s10618-017-0534-x

Kim, K. (2017). Liquidity basis between credit default swaps and corporate bonds markets. *International Review of Economics & Finance, 48*, 98–115.

Kowalewski, O., & Pisany, P. (2022). Banks' consumer lending reaction to fintech and bigtech credit emergence in the context of soft versus hard credit information processing. *International Review of Financial Analysis, 81*, 102116.

Kulczycki, P., & Franus, K. (2020). Outlier detection, clustering, and classification – Methodologically united procedures for conditional approach. In N. T. Nguyen, B. H. Hoang, C. P. Huynh, D. Hwang, B. Trawinski, & G. Vossen (Eds.), *Computational collective intelligence. ICCCI 2020. Lecture notes in computer science* (Vol. 12496). Springer. https://doi.org/10.1007/978-3-030-63007-2_26

Leontitsis, A., & Vorlow, C. E. (2006). Accounting for outliers and calendar effects in surrogate simulations of stock return sequences. *Physica A, 368*, 522–530. https://doi.org/10.1016/j.physa.2005.12.037

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*, 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Li, X., Yang, S., Fan, R., Yu, X., & Chen, D. (2018). Discrimination of soft tissues using laser-induced breakdown spectroscopy in combination with k nearest neighbors (kNN) and support vector machine (SVM) classifiers. *Optics and Laser Technology, 102*, 233–239. https://doi.org/10.1016/j.optlastec.2018.01.028

Li, M., Zhang, P., Yang, G., Xu, G., Guo, M., & Liao, W. (2022). A fisher linear discriminant analysis classifier fused with naïve Bayes for simultaneous detection in an asynchronous brain-computer interface. *Journal of Neuroscience Methods, 371*, 109496.

Ling, X., Yanlin, J., Jin, X., Xin, G., & Jing, H. (2020). GMDH-based outlier detection model in classification problems. *Journal of Systems Science and Complexity, 33*, 1516–1532. https://doi.org/10.1007/s11424-020-9002-6

Liu, W., Fan, H., & Xia, M. (2021). Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Engineering Applications of Artificial Intelligence, 97*, 104036. https://doi.org/10.1016/j.engappai.2020.104036

Ma, Y., Liu, S., & Li, Q. (2020). An advanced multiple outlier detection algorithm for 3D similarity datum transformation. *Measurement, 163*, 107945.

Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: an application of support vector machine. *Risk Manag, 19*, 158–187. https://doi.org/10.1057/s41283-017-0016-x

Murinde, V., Rizopoulos, E., & Zachariadis, M. (2022). The impact of the FinTech revolution on the future of banking: Opportunities and risks. *International Review of Financial Analysis, 81*, 102103.

Nyitrai, T., & Miklos, V. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences, 67*, 34–42. https://doi.org/10.1016/j.seps.2018.08.004

Okada, S., Nagase, K., Ito, A., Ando, F., Nakagawa, Y., Okamoto, K., Kume, N., Takemura, T., Kuroda, T., & Yoshihara, H. (2013). Development of a methodology for the detection of hospital financial outliers using information systems. *The International Journal of Health Planning and Management, 29*, e207–e232. https://doi.org/10.1002/hpm.2194

Okhli, K., & Nooghabi, M. J. (2021). On the contaminated exponential distribution: A theoretical Bayesian approach for modeling positive-valued insurance claim data with outliers. *Applied Mathematics and Computation, 392*, 125712. https://doi.org/10.1016/j.amc.2020.125712

Park, J. W., & Moon, Y. S. (2015). Robust estimation of target scale by removing outlier motion vectors using MAD. *Electronics Letters, 58*(9), 691–693. https://doi.org/10.1049/el.2014.3821

Shen, C. H., Fan, X., Huang, D., Zhu, H., & Wu, M. W. (2018). Financial development and economic growth: Do outliers matter? *Emerging Markets Finance and Trade, 54*(13), 2925–2947.

Souiden, I., Omri, M. N., & Brahmi, Z. (2022). A survey of outlier detection in high dimensional data streams. *Computer Science Review, 44*, 100463.

Treder, M. S., Porbadnigk, A. K., Avarvand, F. S., Muller, K. R., & Blankertz, B. (2016). The LDA beamformer: Optimal estimation of ERP source time series using linear discriminant analysis. *NeuroImage, 279–291*, 279. https://doi.org/10.1016/j.neuroimage.2016.01.019

Vlasselaer, V. V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems, 75*, 38–48. https://doi.org/10.1016/j.dss.2015.04.013

Wang, B., & Mao, Z. (2020). A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule. *Information Fusion, 63*, 30–40. https://doi.org/10.1016/j.inus.2020.05.00

Xiao, J., Tian, Y., Xie, L., Jiang, X., & Huang, J. (2019). A hybrid classification framework based on clustering. *IEEE Transactions on Industrial Informatics, 16*(4), 2177–2188. https://doi.org/10.1109/TII.2019.2933675

Xu, P., Ding, Z., & Pan, M. (2017). An improved credit card users default prediction model based on RIPPER. In *In 2017 13th International conference on natural computation, fuzzy systems and knowledge discovery (ICNCFSKD)* (pp. 1785–1789). IEEE.

Xu, F., Pan, Z., & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing and Management, 57*, 102221. https://doi.org/10.1016/j.ipm.2020.102221

Yang, F., Qiao, Y., Abedin, M. Z., & Huang, C. (2022). Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics, 2022*, 1. https://doi.org/10.1109/TII.2022.3151917

Zhang, L., Su, H., Zio, E., Zhang, Z., Chi, L., Fan, L., Zhou, J., & Zhang, J. (2021). A data-driven approach to anomaly detection and vulnerability dynamic analysis for large-scale integrated energy systems. *Energy Conversion and Management, 234*, 113926.

# Default Risk Prediction Based on Support Vector Machine and Logit Support Vector Machine

**Fahmida-E-Moula, Nusrat Afrin Shilpa, Preity Shaha, Petr Hajek, and Mohammad Zoynul Abedin**

**Abstract** This chapter aims to predict the credit customer default risk. We propose a machine learning algorithm such as Support Vector Machine and a hybrid default risk prediction model such as Logistic Regression and Support Vector Machine being known as LogitSVM (LSVM) to access the credit default risk. We apply three real-world credit databases to validate the probability and value of the proposed risk appraisal hybrid approaches. This chapter uses Type-I Error, Type-II Error, and Root Mean Squared Error (RMSE) to evaluate the performance of the algorithms. Empirical findings show that hybrid model experimentation (LogitSVM) maximizes overall accuracy and minimizes RMSE, Type-I error, and Type-II error. This study is useful for stakeholders to develop a wide variety of approaches to predict risk of default of the credit customer.

**Keywords** Credit default prediction · Support vector machine · Logistic regression · Hybrid methodology

Fahmida-E-Moula
School of Economics and Management, Dalian University of Technology, Dalian, China

N. A. Shilpa · P. Shaha
Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

P. Hajek
Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic
e-mail: petr.hajek@upce.cz

M. Z. Abedin (✉)
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_6

## 1 Introduction

Risk assessment is the most significant and crucial concern in banking companies and financial institutions (Zhao et al., 2022; Hajek et al., 2022; Efat et al., 2022; Abedin et al., 2021; Shajalal et al., 2021). It is the process of managing the uncertainty related to risks. A sound risk assessment allows banks to plan a strong decision-making procedure that reduces financial losses. Three different types of risk are causes of financial uncertainties, such as market risk, credit risk, and operational risk (Abedin et al., 2019a). Credit risk or default risk means the risk that a lender takes when borrowers are unable to make the required payments on their debt obligations. According to Abedin et al. (2022), credit risk involves 60% of the total risk for the banking industries. Therefore, credit risk is considered as a wide-ranging multifaceted crisis that aims to know credit the performance of the credit customers and forecast their uncertainties. This financial crisis drastically reduces the profit margin. As a result, many banks and other financial institutions face complexities, and some experience economic failure. However, one of the drastic end results of the collapse is that the bank's client or creditor bankruptcy is known as the credit default. Credit Default Prediction (CDP) is essential for financial institutions that aim to decline future losses by eliminating the new credit proposal and estimating the probable default risk. The accuracy of credit forecasting is fundamental to the profitability and efficiency of financial institutions. Even a few positive adjustments in the precision of credit risk prediction of potential applicants with credit default will lessen a massive future loss for the financial industry (Abedin et al., 2019b).

According to Vapnik (1995), the Support Vector Machine (SVM) is an extensive applied algorithm for credit approval data classification. SVM-based non-parametric intelligent methods are more appropriate for default risk data classification since financial data involve specific character such as non-linearity and neutrality of covariance matrices between two groups of credit customers' class. However, the standalone predictive algorithm cannot create the best credit risk accuracy for all prediction problems. Therefore, there is a growing concentration that existing applications of standalone learners may be further enhanced by utilizing blending or hybrid methods. The hybrid forecasting model means the blend of traditional and current artificial intelligence (AI) techniques, which signifies improved forecasting capacity than the application of a single classifier (Chi et al., 2019). Additionally, the hybrid learning system outperforms a standalone algorithm that provides better accuracy and fewer prediction errors when employed in modeling credit approval datasets (Moula et al., 2017). The purpose of this chapter is to predict the default risk of the credit customer to minimize the burden of the applied credit risk prediction classifiers. Therefore, this chapter proposes one machine learning algorithm such as a Support Vector Machine and a hybrid default risk prediction model such as Logistic Regression and Support Vector Machine known as LogitSVM (LSVM) to access the credit default risk. Empirical findings show that experimentation with the hybrid model (LogitSVM) minimizes the RMSE, Type-I error, and Type-II error and maximizes overall accuracies. This study is useful for policymakers who have the opportunity to inspect customer financial practices that are able to increase their future capability.

## 2 Literature Review

Researchers use many statistical classifiers to predict the default risk of credit customers. For instance, multivariate adaptive regression splines (MARS) (Lee et al., 2006), survival models (Luo et al., 2016), linear discriminant analysis (LDA) (Lu et al., 2022), and fuzzy logistic regression analysis (Yang et al., 2022). Jiashen You and Tomohiro Ando (2013) use a statistical model for the concurrent estimation of hazard rate, risk-free interest rate, and loss given default, as well as the credit risk dependency structure. However, there are difficulties with using these statistical classifiers to predict credit approval data analysis. For instance, some hypothesizes such as the multivariate normality hypothesizes for independent variables are usually violated in reality which makes these models hypothetically unacceptable for an example set.

Researchers also used many machine learning classifiers to predict credit customer default risk analysis. Boyacioglu et al. (2009) employed SVMs, three multivariate statistical methods, and four different neural network models to the problem of forecasting bank credit failures. Huang et al. (2007) investigated that SVM-based credit prediction approach can properly classify applications as either accepted or rejected, reducing creditors' risk and interpreting future savings. Lee (2007), Kim and Ahn (2012) and Shin et al. (2005) used SVMs to Korean credit risk approval dataset and bankruptcy prediction. Ding et al. (2008), Hui and Sun (2006), and Xie et al. (2011) utilized SVMs for the credit modeling of Chinese listed companies. Experimenting with a Peruvian microfinance credit database, Blanco et al. (2013) employed several intelligence credit risk assessment models based on the MLP approach. However, the standalone analytical algorithm cannot create the best credit risk accuracy for all prediction problems.

Therefore, nowadays corporate analysts and academic modelers have paid special attention to hybridization along with the non-parametric approaches (Son et al., 2016). In order to deal with the restrictions of statistical models and standalone predictive algorithm and to generate the best credit risk accuracy for all forecasting problems, SVM and LogitSVM-based default risk prediction models (hybrid models) are proposed in the literature. SVM is a flexible and intelligent method that creates additive data connections with fewer predictors. LogitSVM (hybrid model) increases credit risk discrimination ability by ensuring variety of prediction assignments, model augmentation, and multifunctionality. Lin (2009) explores a two-stage blending method of LR with BPN to Taiwanese banks' distress database in the bankruptcy prediction domain. The hybrid model not only improves the prediction power but also minimizes the misclassification error. Besides, the hybrid technique applied in this chapter solves the over fitting concerns of other studies. Consequently, it improves the ability to discriminate default risk.

# 3 Methodology

## *3.1 Datasets*

We focus on three credit datasets including "Credit Approval," "German Credit" and "Japanese Credit" to verify the probability and effectiveness of the proposed credit risk assessment model. The "Credit Approval" data comes from Alyuda NeuroIntelligence (http://www.alyuda.com). This database consists of 238 samples of non-risky customers and 262 samples of risky customers. Each case includes twelve financial and non-financial characteristics and one class attribute. The German and Japanese credit databases come from the UCI (University of California, Irvine) machine learning database repository. The "German credit" dataset consists of 700 non-risky and 300 risky customers. Each credit customer seizes seven numerical, thirteen categorical attributes, and one target variable. The "Japanese credit" dataset includes a total of 690 instances having 307 non-default creditors and 383 default creditors. It has fifteen attributes that include nine nominal variables, six continuous variables, and one class attribute. This chapter applies three different types of training scheme, 30%:70%, 50%:50%, and 70%:30%, respectively, to determine the most optimal one.

## *3.2 Forecast Algorithms*

**Support Vector Machine**

The SVM is suitable for a small sample, nonlinear, and high-dimensional data. Two types of SVM are now accessible (i) Linear SVM and (ii) Kernel SVM. Linear SVM acts as an extremely fast machine learning algorithm and performs an original proprietary algorithm with a view to solve multiclass problems in large datasets. Kernel-based SVM is used for nonlinear data classification. In a nonlinear situation, SVM mainly uses a kernel function to chart the preliminary data in the high-dimensional factor to attain linear separability. Through this, it assists to solve the issue of linear inseparability in the initial factor.

For a linear separable data set ($x_i$, $y_i$; $i = 1, 2, \ldots, n$), $x \in R^n$ and $y \in R^n$, the separation hyperplane is gained by maximizing the interval or solving the corresponding convex quadratic programming problem:

$$\omega^T x + b = 0, \tag{1}$$

where $\omega$ is a parameter vector, $x$ and $b$ are sample data and offset, respectively. The corresponding classification decision function is:

$$f(x) = sgn\left(\omega^T x + b\right). \tag{2}$$

For a linearly non-separable data set, each sample point presents a relaxation variable to symbolize a non-negative measure of the misclassification error. The following optimization problem represents the linear-non-separable SVM:

$$\min \frac{1}{2}\|\omega\|^2 + C\sum\nolimits_{i=1}^{n}\xi_i, \tag{3}$$

$$\text{s.t. } y_i(\omega x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\ldots,n, \tag{4}$$

where $C$ is the penalty factor that controls the association between accuracy and generalization in the credit prediction training set.

Combining the kernel function with soft interval maximization principle, the decision function of nonlinear SVM can be obtained by using the dual function and Lagrange optimization algorithm, as follows:

$$f(x) = sgn\left(\sum\nolimits_{i=1}^{n} a_{i,} y_i K(x,\ x) + b\right), \tag{5}$$

where $a_i \geq 0$ symbolizes the Lagrange multiplier and $K(x, x_i)$ represents the kernel function, in agreement with the Mercer theorem.

To reduce computationally expensive calculations, the inner product is replaced with kernel function $K(x_i, x_j)$. It converts the credit forecasting input data into a high-dimensional feature space where the credit forecasting problems are separable and hence increases the ability of the learning machine. Common forms of such kernel functions include:

(a) The linear kernel, $K(x_i, x_j) = x_i^T x_j$
(b) The sigmoid kernel, $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$
(c) The polynomial kernel, $K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d$; and
(d) The radial basis kernel, $K(x_i, x_j) = \exp.\ (-\gamma \parallel x_i - x_j \parallel^2)$

As a final SVM classifier, this chapter obtains the decision function as follows:

$$Y(x) = \text{sgn}\ \left(\sum\nolimits_{i=0}^{n} y_i \alpha_i\ K(x,\ x_i\ ) + b\right), \tag{6}$$

where $Y(x)$ represents the SVM decision function, sgn is the sign of the decision parameter, $K(x, x_i)$ represents the kernel function, $\alpha_i$ is the Lagrange multipliers, and $b$ is the bias of the model.

**Logistic Regression**

Logistic regression (LR) is a widely applied credit default prediction data modeling method. The response variable of (LR), i.e., the outcome is binary (0, 1). Therefore, researchers can employ it to clarify the relationship between the occurrence of an incident of interest and a set of probable descriptive variables. In the circumstance of credit approval data modeling, the outcome links up to the borrowing loan performance of a borrower during a specified period, usually twelve months.

However, LR represents a valuable classifier on the basis of two foundations in the context of credit risk appraisal modeling. Firstly, in an LR, subsequent probabilities are determined directly, which makes it more comprehensible than more versatile "black box" techniques. Secondly, LR-based data classification has exposed it to make robust and better predictions in benchmarking studies for credit risk assessment (Guo et al., 2016; Caigny et al., 2018). Therefore, LR can affirm more difficult data classifiers in credit approval data modeling.

## 3.3 Performance Measures

Previous studies propose a number of appraisal performance measures to assess forecasting methods in the field of credit approval data analysis (Abedin et al., 2018). The evaluation of the forecasting capacity of a classifier is built from a confusion matrix. This matrix is a special tabulation of correctly and incorrectly predicted examples for each class. A confusion matrix for binary classification is as stated in Table 1, where *tp* refers to true positive, *tn* is true negative, *fp* means false positive and *fn* represents false negative.

The Type-I error represents that a creditor with good status is misclassified as a creditor with bad status in Eq. (7) and the Type-II error states that a creditor with a bad status is misclassified as a creditor with a good status in Eq. (8):

$$\text{Type I error} = fn/(tp + fn), \tag{7}$$

$$\text{Type II error} = fp/(fp + tn). \tag{8}$$

The root mean square error (RMSE) is the average root square difference between the estimated and actual values, that is:

$$\text{RMSE} = \sqrt{(1/N)\sum_{i=1}^{n}(\theta_i - P_i)^2}, \tag{9}$$

where $N$ refers to the total number of credit approval samples, $\theta_i$ represents a binary display for the actual execution of the status variable (0 if non-default, 1 if default), and $P_i$ is the estimated probability of default. Credit scoring with large errors is

**Table 1** Confusion matrix for a classification problem

| | | Predicted observations | |
|---|---|---|---|
| | | Predicted positive | Predicted negative |
| Actual observations | Actual positive | *tp* | *fp* |
| | Actual negative | *fn* | *tn* |

weighted more deeply in Eq. (9) since the errors are squared before determining the mean. Therefore, this performance indicator is efficient in estimating huge surplus deviations.

# 4 Results

## 4.1 Description of the Data

A summary of the three datasets is available in Table 2. The total number of customers ranged from 500 to 1000, while two datasets were almost balanced (Credit approval and Japanese credit datasets), and the German dataset was imbalanced in favor of risky customers. The dimensionality of the databases ranged from 13 to 20.

## 4.2 Prediction of Credit Risk

From the experimental results shown in Tables 3 and 4, for the "Credit Approval" dataset, we can find that the overall total accuracy of LSVM is 92.8%, while SVM is 92.7%. Moreover, it reveals that the overall LSVM generates the smallest RMSE and Type–II error than the SVM. The total RMSE and Type–II error of LSVM are 5.35 and 1.98%, while SVM are 5.58 and 2.17%, respectively.

For the German credit dataset shown in Tables 5 and 6, the total RMSE is the same for both LSVM and SVM. The average Type–I error is 29.3% in LSVM, while it is 30.0% for SVM. Regarding the kernel functions used, LSVM with linear and polynomial kernel functions performed best, with high accuracy and low RMSE and Type–I and Type–II errors.

The results for the Japanese credit approval database are presented in Tables 7 and 8. The results report that the overall total accuracy of the LSVM is 92.1% while it is 90.2% for the SVM. Furthermore, the results expose that the total RMSE is 0.557, the Type–I error is 34.2%, and the Type–II error is 21.2% of the LSVM. The total RMSE, Type–I error, and Type–II error of SVM are 0.572, 36.2%, and 22.4%, respectively. It is clear that the errors of LSVM are smaller than those of SVM for the Japanese credit database.

**Table 2** Description of databases used in the experiments

| | Total cases | Non-risky/risky customers | No. of attributes |
|---|---|---|---|
| Credit approval | 500 | 238/262 | 13 |
| German credit | 1000 | 700/300 | 20 |
| Japanese credit | 690 | 307/383 | 15 |

**Table 3** Blending LogitSVM performance for the "Credit Approval" database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset[a] | Te dataset[a] | Overall[a] | | Type-I | Type-II |
| 30:70 | LSVM – 1 (LinK) | 86.67 | 85.43 | 85.80 | 0.1867 | 19.85 | 7.17 |
| | LSVM – 2 (RbfK) | 86.67 | 87.14 | 87.00 | 0.4315 | 19.43 | 4.61 |
| | LSVM – 3 (PolK) | 86.67 | 87.14 | 87.00 | 0.3737 | 19.43 | 4.61 |
| | LSVM – 4 (SigK) | 51.33 | 48.86 | 49.60 | 0.7025 | 52.59 | 47.83 |
| 50:50 | LSVM – 1 (LinK) | 84.40 | 87.60 | 86.00 | 0.3735 | 19.78 | 6.76 |
| | LSVM – 2 (RbfK) | 85.60 | 88.40 | 87.00 | 0.3601 | 19.43 | 4.61 |
| | LSVM – 3 (PolK) | 85.60 | 88.40 | 87.00 | 0.3601 | 19.43 | 4.61 |
| | LSVM – 4 (SigK) | 50.80 | 50.80 | 50.80 | 0.7085 | 52.00 | 47.33 |
| 70:30 | LSVM – 1 (LinK) | 84.86 | 80.00 | 83.40 | 0.4181 | 19.84 | 13.17 |
| | LSVM – 2 (RbfK) | 86.86 | 87.33 | 87.00 | 0.3732 | 19.43 | 4.61 |
| | LSVM – 3 (PolK) | 86.86 | 87.33 | 87.00 | 0.3592 | 19.43 | 4.61 |
| | LSVM – 4 (SigK) | 50.00 | 50.67 | 50.20 | 0.7071 | 52.56 | 47.72 |

Note: [a]Tr refers to in-sample instances, while Te refers to out-sample instances. The overall results are the average outcomes of the Tr and Te instances

**Table 4** SVM performance for the "Credit Approval" database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset | Te dataset | Overall | | Type-I | Type-II |
| 30:70 | SVM – 1 (LinK) | 71.33 | 92.22 | 85.92 | 0.4072 | 16.87 | 11.29 |
| | SVM – 2 (RbfK) | 63.33 | 93.37 | 84.31 | 0.4315 | 11.33 | 18.71 |
| | SVM – 3 (PolK) | 76.00 | 93.37 | 88.13 | 0.3737 | 15.89 | 7.53 |
| | SVM – 4 (SigK) | 50.00 | 51.30 | 50.91 | 0.7025 | 52.31 | 47.02 |
| 50:50 | SVM – 1 (LinK) | 83.60 | 87.20 | 85.40 | 0.3814 | 20.00 | 8.00 |
| | SVM – 2 (RbfK) | 85.60 | 88.40 | 87.00 | 0.3601 | 19.44 | 4.61 |
| | SVM – 3 (PolK) | 85.60 | 88.40 | 87.00 | 0.3601 | 19.44 | 4.61 |
| | SVM – 4 (SigK) | 48.80 | 50.80 | 49.80 | 0.7085 | 52.36 | 47.56 |
| 70:30 | SVM – 1 (LinK) | 86.86 | 81.33 | 85.20 | 0.3973 | 20.29 | 8.04 |
| | SVM – 2 (RbfK) | 88.00 | 84.00 | 86.80 | 0.3732 | 18.15 | 7.39 |
| | SVM – 3 (PolK) | 88.00 | 84.67 | 87.00 | 0.3771 | 19.44 | 4.61 |
| | SVM – 4 (SigK) | 50.00 | 50.00 | 50.00 | 0.7071 | 52.61 | 47.78 |

### 4.3 Comparative Analysis of Prediction Models

To observe more reliability of the findings of current experimental setups, this chapter applies a non-parametric Wilcoxon signed-ranks (WSR) test, which sets the significance level at $p = 0.01/0.05$ to attach the statistically significant performance differences among the LogitSVM-based credit risk assessment classifiers. Moreover, the objective of the study is to establish that the proposed hybrid

**Table 5** Blended LogitSVM performance for the German credit database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset | Te dataset | Overall | | Type-I | Type-II |
| 30:70 | LSVM – 1 (LinK) | 73.00 | 77.14 | 75.90 | 0.4988 | 19.97 | 37.61 |
| | LSVM – 2 (RbfK) | 75.00 | 72.26 | 73.10 | 0.5132 | 26.05 | 34.69 |
| | LSVM – 3 (PolK) | 74.33 | 77.71 | 76.70 | 0.4893 | 19.92 | 35.13 |
| | LSVM – 4 (SigK) | 42.00 | 45.86 | 44.70 | 0.7487 | 30.00 | 70.16 |
| 50:50 | LSVM – 1 (LinK) | 75.00 | 75.60 | 75.30 | 0.4970 | 20.75 | 38.39 |
| | LSVM – 2 (RbfK) | 70.20 | 72.00 | 71.10 | 0.5375 | 27.75 | 43.24 |
| | LSVM – 3 (PolK) | 74.80 | 76.60 | 75.70 | 0.4928 | 20.64 | 37.27 |
| | LSVM – 4 (SigK) | 34.00 | 50.20 | 42.10 | 0.7591 | 29.67 | 70.00 |
| 70:30 | LSVM – 1 (LinK) | 76.43 | 75.33 | 76.10 | 0.4911 | 20.00 | 36.96 |
| | LSVM – 2 (RbfK) | 70.86 | 73.33 | 71.60 | 0.5281 | 27.51 | 39.73 |
| | LSVM – 3 (PolK) | 75.43 | 74.67 | 75.20 | 0.4995 | 20.92 | 38.46 |
| | LSVM – 4 (SigK) | 46.00 | 41.67 | 44.70 | 0.7493 | 30.00 | 70.16 |

**Table 6** SVM performance for the German credit database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset | Te dataset | Overall | | Type-I | Type-II |
| 30:70 | SVM – 1 (LinK) | 70.33 | 77.14 | 75.10 | 0.5114 | 21.42 | 37.92 |
| | SVM – 2 (RbfK) | 72.00 | 70.86 | 71.20 | 0.5345 | 28.27 | 38.46 |
| | SVM – 3 (PolK) | 71.67 | 77.29 | 75.60 | 0.5045 | 21.21 | 36.54 |
| | SVM – 4 (SigK) | 42.00 | 70.00 | 61.60 | 0.6547 | 30.00 | 70.00 |
| 50:50 | SVM – 1 (LinK) | 74.20 | 78.00 | 76.10 | 0.4885 | 20.41 | 36.20 |
| | SVM – 2 (RbfK) | 71.80 | 70.60 | 71.20 | 0.5366 | 28.45 | 36.36 |
| | SVM – 3 (PolK) | 74.00 | 77.80 | 75.90 | 0.4906 | 20.99 | 35.88 |
| | SVM – 4 (SigK) | 34.00 | 50.00 | 42.00 | 0.7648 | 30.00 | 70.00 |
| 70:30 | SVM – 1 (LinK) | 74.43 | 74.67 | 74.50 | 0.5045 | 21.22 | 40.09 |
| | SVM – 2 (RbfK) | 72.57 | 70.00 | 71.80 | 0.5357 | 27.95 | 32.69 |
| | SVM – 3 (PolK) | 75.00 | 77.33 | 75.70 | 0.4764 | 20.44 | 37.45 |
| | SVM – 4 (SigK) | 46.00 | 42.00 | 44.80 | 0.7482 | 30.00 | 70.00 |

algorithms are reliable learners to distinguish the non-risky customers from their risky counterparts. However, in the database, all credit assessment classifiers (Model Z) are verified for significant dissimilarity from the perfect classifier (Model A). The null hypothesis represents the overall accuracy of Model A/type-I error/type-II error = the overall accuracy of Model Z/type-I error/type-II error, while the inverse is the alternative hypothesis. The column "improvement" states the relative progress of the average CRA accuracy (type-I error/type-II error) that model A achieves over model Z. The results are summarized in Tables 9, 10, and 11.

Table 3 shows that for 30%:70%, 50%:%50% and 70%:30% TSs, LSVM-3 has the highest averages in overall credit risk assessment (CRA) accuracies. For the

**Table 7** Blended LogitSVM performance for the Japanese credit database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset | Te dataset | Overall | | Type-I | Type-II |
| 30:70 | LSVM – 1 (LinK) | 76.81 | 84.68 | 82.32 | 0.4364 | 14.28 | 19.72 |
| | LSVM – 2 (RbfK) | 82.13 | 86.96 | 85.51 | 0.3920 | 21.33 | 6.99 |
| | LSVM – 3 (PolK) | 82.13 | 86.96 | 85.51 | 0.3920 | 21.33 | 6.99 |
| | LSVM – 4 (SigK) | 48.79 | 50.10 | 49.71 | 0.7110 | 55.49 | 44.48 |
| 50:50 | LSVM – 1 (LinK) | 82.90 | 87.25 | 85.07 | 0.3853 | 19.82 | 10.23 |
| | LSVM – 2 (RbfK) | 83.48 | 87.54 | 85.51 | 0.3797 | 21.33 | 6.99 |
| | LSVM – 3 (PolK) | 83.48 | 87.25 | 85.36 | 0.3818 | 21.55 | 7.01 |
| | LSVM – 4 (SigK) | 53.04 | 52.75 | 52.90 | 0.6863 | 52.68 | 41.81 |
| 70:30 | LSVM – 1 (LinK) | 83.02 | 89.37 | 84.93 | 0.3690 | 18.96 | 11.57 |
| | LSVM – 2 (RbfK) | 83.64 | 89.86 | 85.51 | 0.3614 | 21.33 | 6.99 |
| | LSVM – 3 (PolK) | 83.64 | 90.34 | 85.65 | 0.3576 | 21.11 | 6.97 |
| | LSVM – 4 (SigK) | 51.76 | 54.59 | 52.61 | 0.7175 | 53.01 | 42.18 |

**Table 8** SVM performance for the Japanese credit database

| TS ratio (%) | LSVM model | Risk assessment accuracy (%) | | | RMSE | Error (%) | |
|---|---|---|---|---|---|---|---|
| | | Tr-dataset | Te dataset | Overall | | Type-I | Type-II |
| 30:70 | SVM – 1 (LinK) | 77.78 | 78.47 | 78.26 | 0.4677 | 28.84 | 13.48 |
| | SVM – 2 (RbfK) | 78.26 | 86.96 | 84.35 | 0.4341 | 20.48 | 11.05 |
| | SVM – 3 (PolK) | 82.13 | 86.96 | 85.51 | 0.3920 | 21.33 | 6.99 |
| | SVM – 4 (SigK) | 47.83 | 52.17 | 50.87 | 0.7070 | 55.93 | 44.76 |
| 50:50 | SVM – 1 (LinK) | 75.07 | 86.67 | 80.87 | 0.4322 | 16.88 | 21.08 |
| | SVM – 2 (RbfK) | 83.48 | 87.54 | 85.51 | 0.3798 | 21.33 | 6.99 |
| | SVM – 3 (PolK) | 83.48 | 87.25 | 85.36 | 0.3818 | 21.39 | 7.27 |
| | SVM – 4 (SigK) | 51.01 | 47.83 | 49.42 | 0.7111 | 55.53 | 44.52 |
| 70:30 | SVM – 1 (LinK) | 82.82 | 85.99 | 83.77 | 0.3944 | 22.22 | 10.03 |
| | SVM – 2 (RbfK) | 83.85 | 89.37 | 85.51 | 0.3639 | 21.33 | 6.99 |
| | SVM – 3 (PolK) | 83.85 | 89.37 | 85.51 | 0.3639 | 21.33 | 6.99 |
| | SVM – 4 (SigK) | 45.76 | 51.21 | 47.39 | 0.6951 | 55.41 | 44.19 |

German credit dataset, Table 5 shows that LSVM-3 has the highest averages in overall credit risk assessment (CRA) accuracies in 30%:70% and 50%:%50% TSs, but LSVM-1 has the highest accuracies for 70%:30% TSs. For the Japanese credit dataset, Table 7 represents that LSVM-3 has the highest accuracies in 30%:70% and 70%:30% TSs, but LSVM-2 has the highest accuracies in 50%:%50% TSs.

Evidence from Tables 9–11 shows that in 30%:70% and 50%:50% TSs, LSVM-3 on the German credit database obtain a remarkable improvement compared to other classifiers considering the overall CRA accuracy criterion. For type-I error, LSVM-3 yields more than 30% improvement for the same dataset in 50%:50%, while for type-II error, LSVM-3 on a similar database attains more than 46% improvement. It

**Table 9** Results of Wilcoxon signed-ranks test for the "Credit Approval" database

| TS Ratio (%) | Model A | Model Z | Overall accuracy | | Type-I Error | | Type-II Error | |
|---|---|---|---|---|---|---|---|---|
| | | | Impr. (%) | $p$ | Impr. (%) | $p$ | Impr. (%) | $p$ |
| 30:70 | LSVM-3 | LSVM-1 | 1.3980 | 0.800 | 2.1159 | 1.02E−18[a] | 35.7043 | 5.97E−13[a] |
| | | LSVM-2 | 0.0000 | 0.502 | 0.0000 | 1.62E−25[a] | 0.0000 | 1.31E−5[a] |
| | | LSVM-4 | 75.4032 | 5.78E−7[a] | 63.0538 | 4.53E−55[a] | 90.3617 | 8.66E−28[a] |
| 50:50 | LSVM-3 | LSVM-1 | 1.1628 | 0.525 | 1.8013 | 1.44E−16[a] | 31.8147 | 8.83E−38[a] |
| | | LSVM-2 | 0.0000 | 0.001[a] | 0.0000 | 1.29E−77[a] | 0.0000 | 9.49E−20[a] |
| | | LSVM-4 | 71.2598 | 4.03E−7[a] | 62.6346 | 8.05E−32[a] | 90.2599 | 2.11E−37[a] |
| 70:30 | LSVM-3 | LSVM-1 | 4.3165 | 2.69E−6[a] | 2.0665 | 0.044[b] | 64.9962 | 6.09E−16[a] |
| | | LSVM-2 | 0.0000 | 0.638 | 0.0000 | 2.17E−11[a] | 0.0000 | 7.09E−30[a] |
| | | LSVM-4 | 73.3068 | 1.15E−4[a] | 63.0327 | 7.21E−83[a] | 90.3395 | 3.87E−43[a] |

[a]$\alpha = 0.01$, [b]$\alpha = 0.05$

**Table 10** Results of Wilcoxon signed-ranks test for the German credit database

| TS Ratio (%) | Model A | Model Z | Overall accuracy | | Type-I Error | | Type-II Error | |
|---|---|---|---|---|---|---|---|---|
| | | | Impr. (%) | $p$ | Impr. (%) | $p$ | Impr. (%) | $p$ |
| 30:70 | LSVM-3 | LSVM-1 | 1.0540 | 1.81E−192[a] | 2.1159 | 4.47E−7[a] | 6.5940 | 4.36E−14[a] |
| | | LSVM-2 | 4.9248 | 0.744 | 0.0000 | 5.58E−98[a] | −1.2684 | 0.098 |
| | | LSVM-4 | 71.5884 | 2.92E−22[a] | 63.0538 | 2.07E−17[a] | 49.9287 | 2.21E−33[a] |
| 50:50 | LSVM-3 | LSVM-1 | 0.5312 | 0.841 | 0.5329 | 1.34E−53[a] | 2.9174 | 1.07E−4[a] |
| | | LSVM-2 | 6.4698 | 1.91E−17[a] | 25.6216 | 2.15E−101[a] | 13.8067 | 3.52E−10[a] |
| | | LSVM-4 | 79.8100 | 4.37E−11[a] | 30.4348 | 4.29E−31[a] | 46.7571 | 8.63E−32[a] |
| 70:30 | LSVM-3 | LSVM-1 | 6.2849 | 2.04E−16[a] | 27.2992 | 0.953 | 6.9721 | 0.115 |
| | | LSVM-2 | 1.1968 | 0.453 | 4.3977 | 0.072 | 3.9002 | 3.78E−24[a] |
| | | LSVM-4 | 70.2461 | 1.40E−26[a] | 33.3333 | 9.11E−39[a] | 47.3204 | 7.64E−31[a] |

[a]$\alpha = 0.01$

**Table 11** Results of Wilcoxon signed-ranks test for the Japanese credit database

| TS Ratio (%) | Model A | Model Z | Overall accuracy | | Type-I Error | | Type-II Error | |
|---|---|---|---|---|---|---|---|---|
| | | | Impr. (%) | $p$ | Impr. (%) | $p$ | Impr. (%) | $p$ |
| 30:70 | LSVM-3 | LSVM-1 | 3.8751 | 9.78E−186 | −49.370 | 6.07E−7[a] | 64.5538 | 6.12E−5[a] |
| | | LSVM-2 | 0.0000 | 0.451 | 0.0000 | 1.64E−8[a] | 0.0000 | 2.53E−23[a] |
| | | LSVM-4 | 72.0177 | 2.09E−14[a] | 61.561 | 3.03E−31[a] | 84.2851 | 1.90E−7[a] |
| 50:50 | LSVM-3 | LSVM-1 | 0.5172 | 0.743 | −7.6186 | 2.69E−25[a] | 31.6716 | 8.59E−60[a] |
| | | LSVM-2 | 0.1757 | 0.421 | 1.0209 | 1.84E−61[a] | 0.2853 | 4.75E−38[a] |
| | | LSVM-4 | 61.6446 | 1.96E−4[a] | 59.510 | 0.007[a] | 83.2815 | 3.17E−73[a] |
| 70:30 | LSVM-3 | LSVM-1 | 0.8478 | 0.027[b] | −11.340 | 9.32E−5[a] | 39.7580 | 1.07E−36[a] |
| | | LSVM-2 | 0.1637 | 0.003[a] | 1.0314 | 1.14E−13[a] | 0.2861 | 1.44E−71[a] |
| | | LSVM-4 | 62.8017 | 0.344 | 60.177 | 4.38E−85[a] | −83.4756 | 1.30E−83[a] |

[a]$\alpha = 0.01$, [b]$\alpha = 0.05$

is clear from Tables 9–11 that all improvements in type-I error and type-II error on all databases are statistically significant with respect to the best-performing blending classifiers. On the contrary, in some cases, the improvements of the accuracy criterion regarding the best algorithms are statistically insignificant, and this is mentioned in the fact that the best algorithms have spaces for further improvements relative to their competing learners.

## 5 Discussion

Jiashen You and Tomohiro Ando (2013) show that their numerical results verify the practicality of their proposed statistical methodology. The empirical findings of Boyacioglu et al. (2009) show that, as learning algorithms, SVMs with some neural network architectures outperform the multivariate statistical methods. The findings of Blanco et al. (2013) reveal that neural models outperform statistical techniques. SVMs are the better approach to learn a small size of data patterns as opposed to common DA, LR, and MLP (Kim & Ahn, 2012; Shin et al., 2005). On the other hand, the result of Lin (2009) claims that the hybrid methodology outperforms the baseline models by generating 80.8% prediction accuracy, while the baseline LR and BPN provide 75.6% and 75.34%, respectively. Therefore, in this chapter, we discuss SVM and LogitSVM (hybrid model), which are better than the performance of other statistical methods and baseline models.

## 6 Conclusion

Credit default risk prediction is important to survive for both financial and non-financial companies. Since the recent global financial crisis has exposed, insufficient decision-making not only affects profitability but also threatens firm solvency in the credit approval procedure. As a result, the accuracy of credit forecasting is essential for the profitability and solvency of financial institutions. This study presents SVM and LogitSVM as new blended intelligent algorithms to assess credit risk. We evaluate the performance of the algorithms using Type–I error, Type–II error, and Root Mean Squared Error (RMSE). The results demonstrate that the experimentation with the hybrid model (LogitSVM) minimizes the RMSE, Type–I error, and Type–II error.

The present methodology is extensively applicable in many previous works. Therefore, as a further avenue, further study will improve the investigated technique utilizing more advanced algorithms. We would like to expand the current study as a future line of research by including credit approval databases from other regions. Moreover, the findings of this chapter relate to empirical approaches. Therefore, future work may be further verified by applying a real-life case study.

## References

Abedin, M. Z., Guotai, C., Colombage, S., & Moula, F. E. (2018). Credit default prediction by using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27.

Abedin, M. Z., Guotai, C., Moula, F. E., Zhang, T., & Hassan, M. K. (2019a). An optimized support vector machine intelligent technique using optimized feature selection methods: Evidence from Chinese credit approval data. *Journal of Risk Model Validation, 13*(2), 1–46.

Abedin, M. Z., Guotai, C., Moula, F. E., Azad, A. S. M. S., & Khan, M. S. U. (2019b). Topological applications of multilayer perceptrons and support vector machine in financial decision support systems. *International Journal of Finance & Economics, 24*, 474–507.

Abedin, M. Z., Hasan, M. M., Hassan, M. K., & Hajek, P. (2021). Deep learning-based exchange rate prediction during the COVID–19. *Annals of Operations Research.* https://doi.org/10.1007/s10479-021-04420-6

Abedin, M. Z., Guotai, C., Hajek, P., & Tong, Z. (2022). Combining weighted smote with ensemble learning for class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*.

Blanco, A., Mejías, R. P., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications, 40*, 356–364.

Boyacioglu, M. A., Kara, Y., & Baykan, O. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications, 36*, 3355–3366.

Chi, G., Uddin, M. S., Abedin, M. Z., & Yuan, K. (2019). Hybrid model for credit risk prediction: An application of neural network approaches. *International Journal on Artificial Intelligence Tools, 28*(5), 1–33. https://doi.org/10.1142/S0218213019500179

Ding, Y., Song, X., & Zen, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications, 34*, 3081–3089.

Efat, M. I. A., Hajek, P., Abedin, M. Z., Azad, R. U., Jaber, M. A., Aditya, S., & Hassan, M. K. (2022). Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales. *Annals of Operations Research.* https://doi.org/10.1007/s10479-022-04838-6

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research, 249*, 417–426.

Hajek, P., Abedin, M. Z., & Sivarajah, S. (2022). Fraud detection in mobile money transactions using an XGBoost-based framework. *Information Systems Frontiers.* https://doi.org/10.1007/s10796-022-10346-6

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, 847–856.

Hui, X., & Sun, J. (2006). An application of support vector machine to companies' financial distress prediction. *Lecture Notes in Artificial Intelligence, 3885*, 274–282.

Jiashen You & Tomohiro Ando. (2013). A statistical modeling methodology for the analysis of term structure of credit risk and its dependency. *Expert Systems with Applications, 40*(12), 4897–4905.

Kim, K. J., & Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research, 39*, 1800–1811.

Lee, Y. C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications, 33*(1), 67–74.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*(4), 1113–1130.

Lin, S. L. (2009). A new two-stage hybrid approach of credit risk in banking industry. *Expert Systems with Applications, 36*, 8333–8341.

Lu, Y., Yang, L., Shi, B., Li, J., & Abedin, M. Z. (2022). A Novel Framework of Credit Risk Feature Selection for SMEs during Industry 4.0. *Annals of Operations Research.* https://doi.org/10.1007/s10479-022-04849-3

Luo, S., Kong, X., & Nie, T. (2016). Spline based survival model for credit risk modeling. *European Journal of Operational Research, 253*(3), 869–879.

Moula, E. F., Goutai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management, 19*(2), 158–187.

Shajalal, M., Petr, H., & Abedin, M. Z. (2021). Product backorder prediction with deep neural network on imbalance data. *International Journal of Production Research.* https://doi.org/10.1080/00207543.2021.1901153

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*(1), 127–135.

Son, Y., Byun, H., & Lee, J. (2016). Nonparametric machine learning models for predicting the credit default swaps: An empirical study. *Expert Systems with Applications, 58*, 210–220.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

Xie, C., Luo, C., & Yu, X. (2011). Financial distress prediction on SVM and MDA methods: The case of Chinese listed companies. *Quality & Quantity, 45*, 671–686.

Yang, F., Qiao, Y., & Abedin, M. Z. (2022). Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics*.

Zhao, Y., Goodell, J. W., Dong, Q., Wang, Y., & Abedin, M. Z. (2022). Overcoming spatial stratification of fintech inclusion: Inferences from across Chinese provinces to guide policy makers. *International Review of Financial Analysis, 84*, 102411. https://doi.org/10.1016/j.irfa.2022.102411

# Predicting Corporate Failure Using Ensemble Extreme Learning Machine

**David Veganzones**

**Abstract** Corporate failure prediction has become a major topic in the accounting and finance literature. Effective prediction models are essential for banks and financial institutions to solve financial decision-making problems. In general, artificial intelligence and machine learning techniques have been mainly employed to develop corporate failure models due to their prediction superiority in comparison to the traditional statistical method. Extreme learning machine is a newly developed artificial intelligence technique with an extremely fast learning speed. Nonetheless, its performance instability may be a major constraint for its practical application. The literature documents that the ensemble is one of the widely used methods to improve the generalization performance of weak classifiers. Therefore, we propose in this study an ensemble of extreme learning machine for improving the prediction performance on corporate failure task. In particular, we compare four benchmark ensemble methods (multiple classifiers, bagging, boosting, and random subspace) to evaluate which is best suited for extreme learning machine. Experimental results on French firms indicated that bagged and boosted extreme learning machine showed the best-improved performance.

**Keywords** Forecasting · Corporate failure · Machine learning · Extreme learning machine · Ensemble

## 1 Introduction

The global economic developments of recent decades have put corporate failure and their consequences for economic well-being under the spotlight, to the extent that bankruptcy or business failure has become a crucial task in finance. This, in turn, has emphasized that financial institutions need effective prediction mechanisms in order to make an appropriate lending decision.

---

D. Veganzones (✉)
ESCE International Business School, OMNES Education, Paris La Défense, France

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_7

In general, the objective of corporate failure prediction is to forecast the likelihood that a firm will survive or fail with the minimum possible classification error. That is why corporate failure research aims at binary classification (Séverin & Veganzones, 2021; Ouenniche & Tone, 2017). From the binary classification point of view, the model's output is a dichotomous variable that takes the value of 1 when the firm follows a bankruptcy procedure and is set to 0 when the firm survives. The explanatory variables to design corporate failure prediction models are often financial ratios, which measure the relationship between any two items on financial statements.

Since the pioneer studies of Beaver (1966) and Altman (1968) who documented the predictive power of ratio analysis, many prediction techniques have been employed to develop corporate failure prediction models, including statistical and artificial intelligence methods (Veganzones & Severin, 2020; Kumar & Ravi, 2007; Moula et al., 2017). On the one hand, researchers still employ well-known statistical methods, notably linear discriminant analysis and logistic regression, due to their simplicity and capacity to interpret the data, even though they are clearly outperformed by machine learning techniques. On the other hand, artificial intelligence techniques (i.e., support vector machine, decision trees, neural networks, fuzzy set theory, self-organizing map) have become indispensable tools in the field of corporate failure prediction, especially in this era of advanced informatics and computing technology (Abedin et al., 2021). Their superiority relies on the fact that they learn directly from the data, which makes it possible to test complex data using nonlinear approaches, and therefore, their predictions are more reliable. Nonetheless, these mentioned methods are not free of drawbacks: low learning rate, slow computational time, converge in local minima, etc. (Yu et al., 2014; Abedin et al., 2018), which could make corporate failure prediction time consuming and arduous.

To overcome these, we consider a novel prediction method, Extreme Learning Machine (ELM) (Huang et al., 2006a) to predict corporate failure. There are several reasons behind choosing ELM as the classifier for the prediction of corporate failures. Firstly, despite many existing methodologies for predicting corporate failure, new methods of research should be continually explored by researchers and practitioners. Secondly, the main concept behind ELM is the random initialization of the Single Layer Feed-Forward Neural Network (SLFN), which replaces the computationally cost procedure of training the hidden layer performed by other artificial intelligence techniques. Unlike the AI techniques, it does not need to calibrate parameters, such as the learning rate. For this reason, ELM has good performance with an extremely fast learning speed (Akusok et al., 2015) and it is proven to be a universal approximator given enough hidden neurons (Huang et al., 2006b).

However, as other techniques, ELM possesses a main drawback: the random initialization that allows ELM to be an extremely fast algorithm, it becomes ELM a highly unstable classifier as well. In ELM, even if we train the same training sample several times, it performs differently due to the random initialization of bias and weights between the input and hidden nodes. Although the reliance on a single ELM may be misguided, the ensemble of predictions might improve the generalization performance of the ELM. Indeed, ensemble methods are usually used as an instrument for improving the accuracy of the learning algorithm by constructing and combining a set of weak classifiers (Kim & Kang, 2010; Abedin et al., 2022). This rationale motivates our specific study of the performance of the ensemble extreme learning machine to predict corporate failure.

Consequently, the aim of this current work is to fully examine which is the best ensemble procedure to improve the performance of ELM for corporate failure prediction. This is of significant importance because the diversity generation method is key in the process of creating an ensemble of classifiers. According to Rokach (2010), diversity creation can be obtained in several ways: by manipulating the training sample, by manipulating the inducer, by varying the representation of the target attribute and by changing the search space. Of all possible ensemble techniques, we selected 4 based on their popularity in the literature (Verikas et al., 2010): Multiple classifiers, Bagging, Boosting, and Random Subspace. The fact that the chosen techniques rely on different ensemble procedures might provide further insight into the general characteristics of ensemble techniques that are influenced by the base classifier. In turn, a rigorous study of such methods would provide assistance in designing a model of corporate failure based on ensemble ELM. Furthermore, optimal performance of prediction models developed based on ensemble ELM models can be employed as a baseline prediction model for future research.

The rest of the paper is organized as follows. Section 2 presents the research methodology. Sections 3 and 4 describe the experimental design and results, respectively. Finally, in Sect. 5, the conclusions are summarized.

## 2 Research Methodology

In this section, we present the method employed in this study. In particular, we describe the extreme learning machine classifier as well as the ensemble modeling techniques.

### *2.1 Extreme Learning Machine*

The Extreme Learning Machine (ELM) classifier was proposed by Huang et al. (2006a). The ELM represents a fast way of creating a Single Layer Hidden Feed-Forward Neural Network (SLFN) by the random initialization of the internal bias and weights. The hidden layer does not need to be iteratively tuned; it bypasses the time-consuming calibration setup performed by artificial intelligence algorithms. As a result, ELM is an extremely fast learning speed while being a simple method. The ELM algorithm can be described as follows:

Consider a set of $N$ observations with features $\boldsymbol{x}_i \in \mathbb{R}^N$ and the corresponding output labels $\boldsymbol{Y} \in \{-1, 1\}^{Nxc}$. A SLFN with $m$ neurons in the hidden layer is written by the following sum:

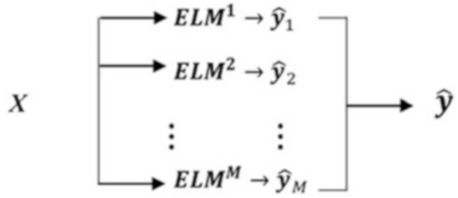



**Fig. 1** Architecture of the multiple classifier

$$\Sigma_{j=1}^{m} \beta_j\, \phi\left(w_j\, \boldsymbol{x}_i + b_j\right) = \boldsymbol{Y}_{ik}, i = 1, \ldots, N\ k = 1, \ldots, c, \tag{1}$$

where $\beta_j$ are the output weights, $\phi$ is the activation function, $w_j$ are the input weights and $b_j$ represents the biases. The Eq. (1) can be expressed in the form of a matrix as $\boldsymbol{H\beta} = \boldsymbol{Y},$ where

$$\boldsymbol{H} = \begin{pmatrix} \phi(w_1\, x_1 + b_1) & \cdots & \phi(w_m\, x_1 + b_m) \\ \vdots & \ddots & \vdots \\ \phi(w_1\, x_N + b_1) & \cdots & \phi(w_m\, x_N + b_m) \end{pmatrix}. \tag{2}$$

$$\boldsymbol{\beta} = (\beta_1 \ \cdots \ \beta_m)^c\ \boldsymbol{Y} = (\,Y_1 \quad \cdots \quad Y_N\,)^c.$$

Then, the output weights $\boldsymbol{\beta}$ can be calculated by the Ordinary Least Squares method using the Moore-Penrose pseudo inverse of $\boldsymbol{H}$ (Rao & Mitra, 1971):

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{Y}. \tag{3}$$

## 2.2 *Ensemble Techniques*

### 2.2.1 Multiple Classifiers Technique

The multiple classifier technique relies on the simple idea that the combination of multiple classifiers leads to higher classification prediction and efficiency than the single classifier. This approach is equivalent to the wisdom of crowds: the combined opinion of diverse and independent experts usually outperforms the opinion of single individuals. According to Kitter et al. (1998), the multiple classifier technique achieves higher efficiency when learners generalize in different ways, i.e., the diversity of the ensemble is generated. As ELM is based on the random initialization of internal bias and weights, each learner will be different; there is diversity in the ensemble. Therefore, the forecast of several ELMs will be combined using majority voting to produce the final decision rule. Figure 1 shows the general architecture of the multiple classifier.

The classifiers $\boldsymbol{C^1}(\boldsymbol{X}), \ldots, \boldsymbol{C^M}(\boldsymbol{X})$ are built based on the data set $\{(\boldsymbol{x}_1,\boldsymbol{y_1}), (\boldsymbol{x}_2,\boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n,\boldsymbol{y_n})\}$. Each classifier provides an output $\widehat{\boldsymbol{y}}_M$ that will be combined into the final output $\widehat{\boldsymbol{y}}$.

### 2.2.2 Bagging

Bagging (short for bootstrap aggregating) is one of the primal ensemble techniques (Breiman, 1996). Its popularity lies in the fact that it is intuitive and simple to implement, with notably good performance. Bagging generates the essential diversity to create the ensemble process that manipulates the training set. In this regard, the training set samples are randomly resampled in order to generate several different bags of samples. Thus, each bag represents a set of training samples. Finally, the base classifier is applied to each bag, and the output classification is made by a majority vote of all the base classifier results.

Bagging technique generates an improvement in generalization performance due to the reduction in variance while maintaining steady or only slightly increasing the bias, in particular, when it is applied to weak classifiers (Grandvalet, 2004). The bagging algorithm can be expressed as follows:

Given a data set $\{(\boldsymbol{x}_1,\boldsymbol{y_1}), (\boldsymbol{x}_2,\boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n,\boldsymbol{y_n})\}$ .

1. Repeat for $i = 1, 2, \ldots, I$.
   (a) Build a bootstrap sample $\left\{\left(\boldsymbol{x}_1^*,\ \boldsymbol{y}_1^*\right),\ \left(\boldsymbol{x}_2^*,\ \boldsymbol{y}_2^*\right),\ \ldots,\ \left(\boldsymbol{x}_n^*,\ \boldsymbol{y}_n^*\right)\right\}$ by randomly selecting $n$ times with replacement from the data $\{(\boldsymbol{x}_1,\boldsymbol{y_1}), (\boldsymbol{x}_2,\boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$.
   (b) Fitting the bootstrapped classifier $\boldsymbol{C}_i$ on corresponding bootstrap sample.
2. Calculate the output of the final classifier:

$$\boldsymbol{C}(\boldsymbol{x}) = I^{-1}\sum\nolimits_i^I C_i(x). \tag{4}$$

### 2.2.3 Boosting

Unlike the bagging technique, the boosting technique combines inaccurate and relatively weak rules to produce highly accurate predictions. That is, it progressively gives more weight to observations that have been misclassified by previously generated classifiers in order to generate new classifiers and then combines the classifiers of different iterations with weighted voting to make final predictions. Since numerous algorithms for boosting have been proposed, we use the Adaboost algorithm (Freund & Schapire, 1996) which is one of the most popular boosting techniques applied to pattern recognition (Verikas et al., 2010). The Adaboost algorithm can be described as follows:

Given a data set $\{(\boldsymbol{x}_1, \boldsymbol{y_1}), (\boldsymbol{x}_2, \boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$ .

1. Initialize the weight vector of the training set:

$$W_1(i) = {}^1\!/_N \text{ for } i = 1, \ldots, N. \tag{5}$$

2. For $t = 1, \ldots, T$,

   (a) Train the weak classifier $C_t$ on the weighted training samples.
   (b) Calculate the sum of weighted errors of $C_t$:

$$\varepsilon_t = \sum\nolimits_{i=1}^{N} W_i^t, Y_i \neq C_t\ (X_i). \tag{6}$$

(c) Choose

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right). \tag{7}$$

(d) Update the weights:

$$W_i^{t+1} = \frac{W_i^t \exp\left( - \alpha_t Y_i C_t\ (X_i) \right)}{Z_t}, \tag{8}$$

where $Z_t$ is a normalization factor.

3. Output:

$$f(x) = sign \left( \sum\nolimits_{t=1}^{T} \alpha_t C_t\ (x) \right). \tag{9}$$

### 2.2.4 Random Subspace

The random subspace (Ho, 1998) bases its ensemble process on the modification of the feature space. That is, it creates different bags of training samples by randomly selecting features drawn for the initial feature set that characterizes each sample. The training sample $\boldsymbol{X}_i (i = 1, \ldots, n)$ in the training set $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_\mathrm{n})$ is a $p$-dimensional vector $\boldsymbol{X}_i = (\boldsymbol{x}_{\mathrm{i}1}, \boldsymbol{x}_{\mathrm{i}2}, \ldots, \boldsymbol{x}_{\mathrm{ip}})$, where $p$ represents the feature components. Within the random subspace, the $k$-dimensional subspace is randomly selected from the original $p$-dimensional feature space, $k < p$. The new learning samples $\boldsymbol{X}^b = \left(\boldsymbol{X}_1^b, \boldsymbol{X}_2^b, \ldots, \boldsymbol{X}_n^b\right)$ in a $k$-dimensional subspace $\boldsymbol{X}_i^b = \left(\boldsymbol{x}_{i1}^b, \boldsymbol{x}_{i2}^b, \ldots, \boldsymbol{x}_{in}^b\right)$ , where $x_{ij}^{b}(j = 1, \ldots, r)$, are built and then, the classifiers in the random subspace $X^b$ are combined using majority voting to create the final decision rule. Thus, the random subspace can be organized as follows:

1. Repeat $b$ times, with $b = 1, 2, \ldots, B$
   (a) Randomly select a $k$-dimensional subspace $X^b$ among the initial $p$-dimensional feature space $X$.
   (b) Design a classifier $C^b(x)$ using the sample $X^b$.
2. Combine the forecast of $C^b(x)$ classifiers using majority voting to a final decision rule.

$$\mathrm{Prev}(x) = \underset{y \in \{-1; 1\}}{\operatorname{argmax}} \sum\nolimits_{b=1}^{B} \delta_{sgn\left(C^b(x)\right), y}. \tag{10}$$

# 3 Experimental Design

## 3.1 Data

Our empirical study uses non-listed French firms taken from the Diane database created by Bureau Van Dijk. The French companies must submit annual reports to the French Commercial Court under French law provide accounting and income statements to the Bureau Van Dijk authority. We drew firms from all sectors of activity (excluding financial companies) for the years 2016–2018, allowing us to examine the model's capacity to create good prediction rules in a real-world scenario.

The Diane database provides the information on whether firms have failed or remain healthy; in the case of failure, it also provides the date. A firm is considered to be failed if it proceeded to be liquidated or reorganized, and non-failed firms were those that continued their activity for at least a year after the period studied. We decided to be conservative in the selection of non-failed firm in order to avoid the inclusion of healthy companies that may suddenly fail and ensure a reliable sample that does not fail. Moreover, firms that presented missing values in their financial statement, as well as outliers, were excluded to ensure the prediction model stability. Consequently, the collected dataset is composed of 3000 failed and 3000 non-failed firms.[1]

[1]Corporate failure is a rare phenomenon in the real world, so failed firms are clearly outnumbered by non-failed ones. That is why the sample selection process becomes a significant paradigm. If one design a model based on the actual population, the dataset must be imbalanced. However, this procedure has a main drawback: it is likely to lead to significant degradation of the prediction performance due to low percentage of failed firm in the entire sample (López et al., 2013; Shajalal et al., 2021). Therefore, we collect a stratified sample with same observations of failed and non-failed based on matched pair technique (Ciampi, 2015), in which failed firms are matched with non-failed firms according to industry sector, size, and firm age.

To minimize the bias effect and sample variability that might influence the model prediction performance, we carried out a tenfold cross-validation method in which the dataset is split into ten distinct training and test set in order to learn and evaluate the model prediction. This procedure was repeated ten times to ensure the reliability of our results. Therefore, the final prediction performance is calculated as the average of 100 testing results.

## 3.2 Variables

Financial dimensions characterize the main explanatory factors for corporate failure. Therefore, the balance sheets and income statements of the collected firms were used to calculate 30 financial ratios to use as explanatory variables. This representation layer is important because it guarantees that the variables, we have used actually represent all aspects of the phenomenon.

The initial set of financial ratios that we compute includes at least four indicators representing six categories: liquidity, solvency, profitability, financial structure, turnover, and activity. These variables are presented in Table 1.

However, using all financial ratios may result in very high-dimensional feature space, which may reduce model predictive capability. Therefore, a variable selection process has been performed in order to choose a subset of the most relevant financial ratios. Following the study by Kainulainen et al. (2011), a feed-forward variable selection process was performed to retain the necessary information for prediction.

## 3.3 Evaluation Metrics

The evaluation criteria of our experiments are adopted from standard measures established in the field of prediction (Shahriare et al., 2021). These measures include average accuracy, type error I, and type error II. The formula of these measures provided below can be explained with respect to the confusion matrix shown in Table 2.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{11}$$

$$\text{Type} - \text{I error} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

$$\text{Type} - \text{II error} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{13}$$

In addition to these evaluation metrics, we also used the area under the receiver operating characteristic curve (AUC) to estimate the model performance. This is a

**Table 1** Initial set of variables

| Profitability | | Liquidity | |
|---|---|---|---|
| X1 | Profit before Tax/Shareholders' Funds | X16 | Cash/total assets |
| X2 | Net income/shareholders' funds | X17 | Current assets/current liabilities |
| X3 | EBITDA/Total assets | X18 | Current assets/total debts |
| X4 | EBIT/Total assets | X19 | Quick assets/Total assets |
| X5 | Net income/Total assets | X20 | (Cash +Marketable securities)/Total sales |
| Financial structure | | Turnover | |
| X6 | Shareholder's funds/Total assets | X21 | Inventory/Total sales |
| X7 | Total debt/shareholders' funds | X22 | Net operating working /Total sales |
| X8 | Total debt/Total assets | X23 | Accounts receivable/Total sales |
| X9 | Net operating working/Total assets | X24 | Accounts payable/Total sales |
| X10 | Long term debt/Total assets | X25 | Current assets/Total sales |
| Solvency | | Activity | |
| X11 | Financial expenses/Total sales | X26 | Cash flow/total sales |
| X12 | Labor expenses/Total sales | X27 | Total sales/total assets |
| X13 | Financial debts/equity | X28 | Value added/total sales |
| X14 | Financial expenses/EBITDA | X29 | Net income/value added |
| X15 | Financial expenses/net income | X30 | EBITDA/Total sales |

EBIT, earnings before interest and taxes; EBITDA, earning before interest, taxes, depreciation, and amortization

**Table 2** Confusion matrix for the prediction of corporate failure

| | | Actually | |
|---|---|---|---|
| | | Failed | Healthy |
| Prediction | Failed | *True positive (TP)* | *False positive (FP)* |
| | Healthy | *False negative (FN)* | *True negative (TN)* |

graphical plot used to represent the model performance while changing the cutoff value. In this case, the proportion of true positive and false positive are plotted on the *x*-axis and *y*-axis of the curve. AUC has become a widely used evaluation metric in corporate failure prediction because it is insensitive to the matrix of misclassification cost[2] to assess the discrimination ability of a model. In summary, two classifiers can be easily compared according to differences in the ROC curve performance. A classifier should get as close to the top left corner as possible, where its value will be close to 1.

With the data set mentioned above, a cross-validation loop (tenfold cross-validation repeated ten times) was performed to estimate the average evaluation measures. To compare the classifier performance, Demšar (2006) recommends a

[2]The misclassification of a failed firm (predict that a firm is healthy when it fails) represent a loss in capital, while the misclassification of a healthy firm (predict that a firm is failed when it survives) represents only a loss of commercial bargain. That is why, misclassified a failed firm is considered to be more costly.

Wilcoxon signed ranks non-parametric test because it only assumes limited commensurability and can be applied to prediction accuracy, misclassification errors or any other evaluation metric. It is expressed as follows:

Given $R^+$ be the sum of ranks when the second classifier outperforms the first one, $R^-$ be the sum of ranks for the opposite and the ranks of $d_i = 0$ are split evenly among the sums:

$$R^+ = \sum_{d_i > 0} \mathrm{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \mathrm{rank}(d_i), \tag{14}$$

$$R^- = \sum_{d_i < 0} \mathrm{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \mathrm{rank}(d_i). \tag{15}$$

Let $T$ be the smaller of the sums, $T = \min(R^+, R^-)$, the normal approximation can be used and the following statistic is used to calculate the $z$-statistics with a corresponding $p$-value:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}. \tag{16}$$

However, Garcia and Herrera (2008) caution that several repeated pairwise comparison tests between algorithms conducted by us may lead to loss of control over family-wise errors.

# 4 Results

Experimental analysis is designed to compare the prediction ability of different ensemble methods based on extreme learning machine classifier. Table 3 indicates the evaluation metrics achieved to assess the performance of the methods. Furthermore, this table is complemented by Table 4, which highlights whether the differences between the methods are statistically significant.[3]

We first analyze the overall performance of the methods. Boosting ELM and Bagging ELM achieve the best mean accuracy values, 82.2% and 82.6%, respectively, while Random subspace ELM attains mean accuracy value of 81.7% and that of 81.4% is achieved with Multiple ELM. All ensemble methods are more accurate than the single ELM (80.4% of the mean accuracy). Thus, it confirms that ensemble ELM methods produce greater predictive power compared to a single ELM

[3] Appendix 1 shows the results on the database using ELM and ELM-ensemble methods. Figures 2 and 3 indicates the testing results with different number of hidden nodes and the average classification error of the ELM-ensemble methods as a function of the number of ensemble members.

**Table 3** Performance of different ELM-based ensemble methods

| | Accuracy | Type-I error | Type-II error | AUC |
|---|---|---|---|---|
| ELM | 80.4% | 21.7% | 17.5% | 0.821 |
| Multiple ELM | 81.4% | 20.3% | 16.7% | 0.834 |
| Bagging ELM | 82.6% | 18.2% | 16.5% | 0.849 |
| Boosting ELM | 82.2% | 18.8% | 16.8% | 0.842 |
| Random subspace ELM | 81.7% | 20.0% | 16.6% | 0.836 |

**Table 4** Significance levels of a test of differences by method and evaluation metric

| | Multiple ELM | Bagging ELM | Boosting ELM | Random subspace ELM |
|---|---|---|---|---|
| | *Accuracy* | | | |
| ELM | 0.0866* | 0.0001*** | 0.0012*** | 0.0338** |
| Multiple ELM | | 0.0463** | 0.0971* | 0.3372 |
| Bagging ELM | | | 0.2908 | 0.985* |
| Boosting ELM | | | | 0.2883 |
| | *Type-I error* | | | |
| ELM | 0.0976* | 0.0001*** | 0.0001*** | 0.0652* |
| Multiple ELM | | 0.0179** | 0.0751* | 0.7871 |
| Bagging ELM | | | 0.5584 | 0.0386** |
| Boosting ELM | | | | 0.182 |
| | *Type-II error* | | | |
| ELM | 0.4275 | 0.0987* | 0.4752 | 0.1255 |
| Multiple ELM | | 0.7213 | 0.6531 | 0.6466 |
| Bagging ELM | | | 0.7889 | 0.6777 |
| Boosting ELM | | | | 0.5133 |
| | *AUC* | | | |
| ELM | 0.0610* | 0.0001*** | 0.0001*** | 0.0462** |
| Multiple ELM | | 0.0133** | 0.1170 | 0.8674 |
| Bagging ELM | | | 0.2891 | 0.0811* |
| Boosting ELM | | | | 0.3746 |

*Significant at 10% threshold; **Significant at 5% threshold; ***Significant at 1% threshold

classification. The fact that Bagging and Boosting ensembles lead to the best reduction in the generalization error is not entirely surprising, as it is well documented their robustness to overfitting (Xiao et al., 2013; González et al., 2020). In contrast, variation of the parameters of the classifiers, such as Multiple ensemble and Random Subspace, can generate greater diversity (Bi, 2012). Nonetheless, the information perceived by the varying diversity does not generate consistent guidance so that the ensemble classifier can obtain a good generalization. On the whole, the key of Boosting and Bagging is that they build a set of diverse classifiers, while they benefit from the balance between diversity and accuracy, which is an important determinant of the performance of ensemble classifiers.

Secondly, we find no uniform improvement among the ensemble methods. If the misclassification errors are analyzed, Boosting ELM and Bagging ELM, here as well, lead to lower misclassification error for failed firms, 18.8% and 18.2%, respectively, significant at 1% threshold in comparison with ELM. In contrast, we do not observe any significant differences in misclassification error for non-failed firms across ensemble methods; rather, the mean type-II error ranges from 16.5% with Bagging ELM and Random Subspace ELM to 18.8% with Bagging ELM.

Finally, the Bagging and Boosting ELM-based methods lead to higher AUC values than the other ensemble methods, which is in line with the previous results. In particular, Bagging ELM seems to be the most optimal ensemble method for corporate failure prediction as results are significantly better than those achieved with the other ensemble methods, but with respect to Boosting ELM.

In sum, the better overall prediction of Bagging and Boosting methods over the other ensemble methods, as previously observed, is due to their capacity to better identify failed firms. The superiority of Bagging ELM is based on the creation of a unique training set for each ensemble member because the perturbation generated in the learning set causes a significant change in the prediction constructed. As a model's prediction is order-correct for most of the replicated observation, the bagging-based ELM can be transformed into a nearly optimal predictor, in particular, for failed firms. Furthermore, one of major reasons why boosted ELM better identifies failed firms may be due to the fact that the new classifier generation gives more relevance to misclassified observation, mostly failed firms. That is, the likelihood of instances that have been misclassified by the previously generated classifier increases, and the set of classifiers grows progressively diverse. This trend explains why this method provides higher accuracy for the minority class without jeopardizing the accuracy of the majority class.

## *4.1 Further Validation*

In order to further evaluate the effectiveness of the ensemble extreme learning machine for the corporate failure prediction task, a new data set has been collected. In general, there is no universal accepted definition of corporate failure; bankruptcy, the more severe form of failure, is commonly used. The popularity of bankruptcy as the definition of failure is based on two concepts: on the one hand, it provides an objective criterion to distinguish failed and non-failed firms, and, on the other hand, the moment of failure can be dated when a firm fills in the bankruptcy procedure. Therefore, the bankruptcy notion offers a discrimination criterion for obtaining a well-defined dichotomy, or at least, a representation of corporate failure, that can be applied methodologically. Nonetheless, numerous studies (Sun et al., 2014; Brédart et al., 2021) consider that corporate failure begins when a firm experiences financial distress. That is, when a firm encounters financial difficulties or struggles to fulfill its obligations. Accordingly, we collected a data set considering financial distress as the definition of corporate failure. We consider the criterion provided by Balcaen et al. (2011), who define financial distress as a firm with negative recurring profit after

**Table 5** Performance of different prediction methods

| | Accuracy | Type-I error | Type-II error | AUC |
|---|---|---|---|---|
| ELM | 78.2% | 24.7% | 18.9% | 0.790 |
| Multiple ELM | 79.5% | 23.0% | 18.0% | 0.804 |
| Bagging ELM | 81.1% | 20.7% | 17.1% | 0.824 |
| Boosting ELM | 80.5% | 21.4% | 17.6% | 0.812 |
| Random subspace ELM | 80.0% | 22.1% | 17.9% | 0.808 |

**Table 6** Significance levels of a test of differences by method and evaluation metric

| | *Accuracy* | | | |
|---|---|---|---|---|
| | Multiple ELM | Bagging ELM | Boosting ELM | Random subspace ELM |
| ELM | 0.0753* | 0.0001*** | 0.0032** | 0.0217** |
| Multiple ELM | | 0.0265** | 0.1333 | 0.2766 |
| Bagging ELM | | | 0.1267 | 0.0836* |
| Boosting ELM | | | | 0.3045 |
| | *Type-I error* | | | |
| | Multiple ELM | Bagging ELM | Boosting ELM | Random subspace ELM |
| ELM | 0.0592* | 0.0001*** | 0.0001*** | 0.0154** |
| Multiple ELM | | 0.0144** | 0.0869* | 0.1936 |
| Bagging ELM | | | 0.1709 | 0.0935* |
| Boosting ELM | | | | 0.2423 |
| | *Type-II error* | | | |
| | Multiple ELM | Bagging ELM | Boosting ELM | Random subspace ELM |
| ELM | 0.2611 | 0.0348** | 0.0107 | 0.2414 |
| Multiple ELM | | 0.2560 | 0.3987 | 0.5612 |
| Bagging ELM | | | 0.6214 | 0.3521 |
| Boosting ELM | | | | 0.3951 |
| | *AUC* | | | |
| | Multiple ELM | Bagging ELM | Boosting ELM | Random subspace ELM |
| ELM | 0.0509* | 0.0001*** | 0.0028*** | 0.0131** |
| Multiple ELM | | 0.0106** | 0.1635 | 0.5145 |
| Bagging ELM | | | 0.0958* | 0.0439** |
| Boosting ELM | | | | 0.3153 |

*Significant at 10% threshold; **Significant at 5% threshold; ***Significant at 1% threshold

taxes over two consecutive years. Consequently, the collected dataset is composed of 2500 failed and 2500 non-failed firms.[4]

The results presented in Tables 5 and 6 are consistent with those of the previous ones. Boosting ELM and Bagging ELM achieve the highest accuracy values, in particular, due to their effectiveness in the reducing the type-I error in comparison to

[4]To design the prediction methods, the same procedure used in Sect. 3.2 was followed. Then, they were evaluated based on a 10-cross validation and using the abovementioned evaluation metrics.

the single ELM.[5] Moreover, it is important to mention that the prediction performance of the methods in this data set is inferior to the previous one. Thus, it is more arduous to differentiate failed firms from healthy ones in the initial steps of failure, when firms just experience financial distress. The literature documented that firms have shown a certain resilience for a long time, even though their financial situation resembles to a bankrupt one (Iftikhar et al., 2021). In contrast, firms that seem completely sound may suddenly fail. Therefore, the inability to know whether the echoes of financial distress may result in corporate failure makes it difficult to capture distinguishable factors that might reinforce model accuracy. That is why the performance of models is lower when corporate failure is represented as financial distress than when it is defined as bankruptcy.

## 5 Conclusion

In this study, we propose to evaluate several ensemble methods applied to corporate failure prediction in order to improve the classification performance of ELM. An ensemble strategy that combines the predictions of individual models is more performance-based than relying on the prediction capacity of a single model. Our results confirm that the Extreme Learning Machine-based ensemble is more accurate and robust than the "individual best" ELM model using two real financial datasets. In particular, the ensemble methods used in this study increase, on average, the classification accuracy estimated for the single ELM by 1.6 and 2.1 percentage points for the bankruptcy data and financial distress data, respectively. An increase in prediction performance of these magnitudes may seem modest, but the readers need to understand that financial institutions and banks can save a huge amount of the limited financial resources with decision technology that can increase the prediction power by 2%.

As Bagging ELM and Boosting ELM give similar results – there is some evidence that the bagging strategy is more effective for the prediction of corporate failure using ELM – it is arduous to make a design recommendation for which method is more optimal. However, we do notice that both methods, which operate by taking a base learner and invoking it multiple times using different training sets, are most effective in the ensemble ELM prediction method. We also notice that bagged ELM is more computationally efficient, as it requires 40–50 ensemble members, while 60–70 members as necessary for the boosting ensemble.

**Acknowledgments** We sincerely thank Prof. Abedin and Prof. Hajek for their assistance.

[5] The Appendix 2 shows graphically the testing results with different hidden nodes (Fig. 4) and the average classification error of ELM-ensemble methods as a function of ensemble members (Fig. 5).

# Appendices

## *Appendix 1*

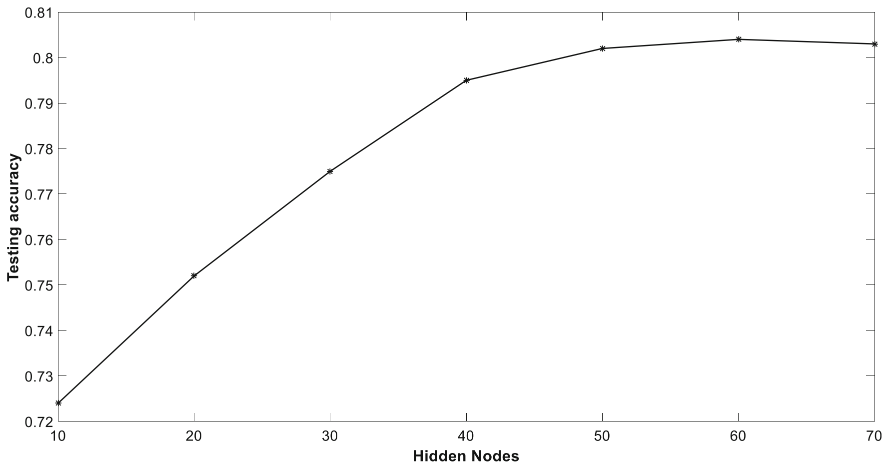



**Fig. 2** Testing results for different hidden nodes in ELM for bankruptcy data

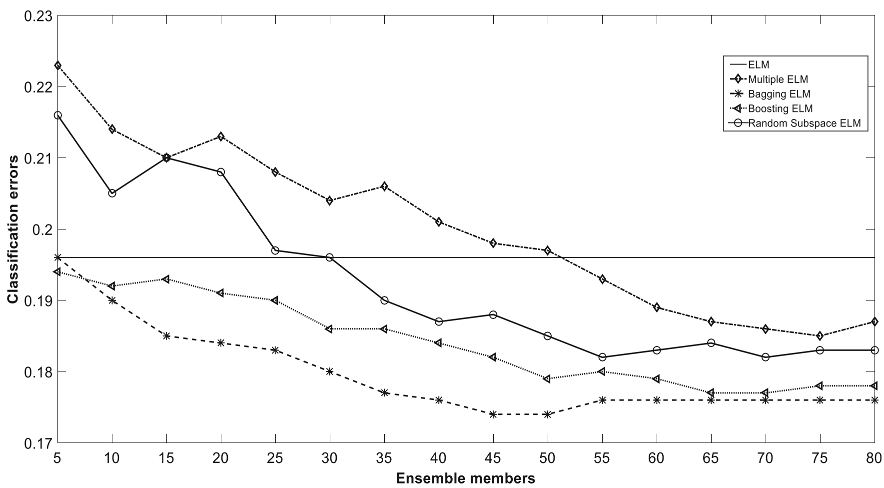



**Fig. 3** Average classification errors of the Ensemble ELM methods by ensemble members for bankruptcy data

## Appendix 2

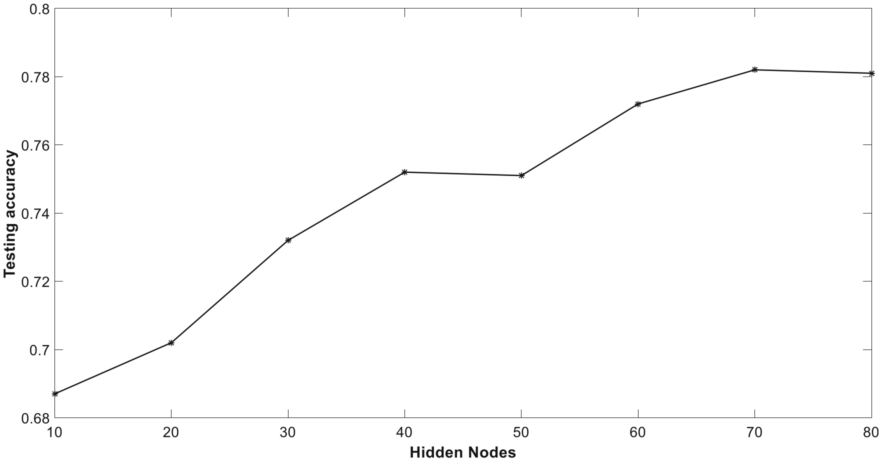



**Fig. 4** Testing results for different hidden nodes in ELM for financial distress data

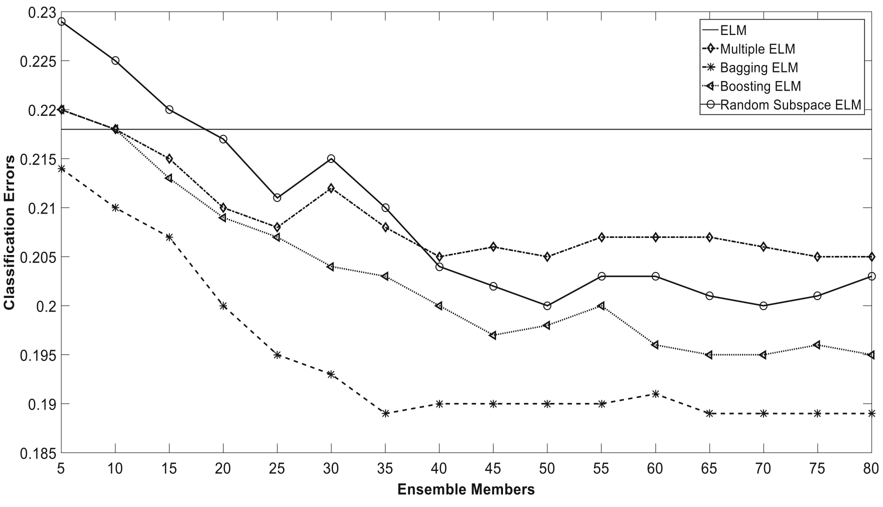



**Fig. 5** Average classification errors of the Ensemble ELM methods by ensemble members for financial distress data

## References


Abedin, M. Z., Chi, G., Colombage, S., & Moula, F. E. (2018). Credit default prediction by using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27.

Abedin, M. Z., Hassan, M. K., Petr, H., & Uddin, M. M. (2021). Machine learning in finance and accounting. In *The essentials of machine learning in finance and accounting,* Taylor & Francis.

Abedin, M. Z., Chi, G., Hajek, P., & Tong, Z. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00614-4

Akusok, A., Veganzones, D., Miche, Y., Björk, K. M., Du Jardin, P., Severin, E., & Lendasse, A. (2015). MD-ELM: Originally mislabeled samples detection using OP-ELM model. *Neurocomputing, 159*, 242–250.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.

Balcaen, S., Manigart, S., & Ooghe, H. (2011). From distress to exit: Determinants of the time to exit. *Journal of Evolutionary Economics, 21*, 407–446.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71–111.

Bi, Y. (2012). The impact of diversity on the accuracy of evidential classifier ensembles. *International Journal of Approximate Reasoning, 53*(4), 584–607.

Brédart, X., Séverin, E., & Veganzones, D. (2021). Human resources and corporate failure prediction modeling: Evidence from Belgium. *Journal of Forecasting, 40*(7), 1325–1341.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises: An empirical analysis of Italian firms. *Journal of Business Research, 68*(5), 1012–1025.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research, 7*, 1–30.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning* (pp. 148–156). IEEE.

Garcia, S., & Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research, 9*(12), 2677–2694.

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion, 64*, 205–237.

Grandvalet, Y. (2004). Bagging equalizes influences. *Machine Learning, 55*(3), 251–270.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006a). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1), 489–501.

Huang, G. B., Chen, L., & Siew, C. K. (2006b). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks, 17*(4), 879–892.

Iftikhar, A., Purvis, L., & Giannoccaro, I. (2021). A meta-analytical review of antecedents and outcomes of firm resilience. *Journal of Business Research, 135*, 408–425.

Kainulainen, L., Miche, Y., Eirola, E., Yu, Q., Frénay, B., Séverin, E., & Lendasse, A. (2011). Ensembles of local linear models for bankruptcy analysis and prediction. *Case Studies in Business, Industry and Government Statistics, 4*(2), 116–133.

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. *European Journal of Operational Research, 180*(1), 1–28.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.

Moula, F. E., Chi, G., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management, 19*(2), 158–187.

Ouenniche, J., & Tone, K. (2017). An out-of-sample evaluation framework for DEA with application in bankruptcy prediction. *Annals of Operations Research, 254*(1), 235–250.

Rao, C. R., & Mitra, S. S. K. (1971). *Generalized inverse of matrix and its application* (Wiley Series in Probability and Mathematical Studies). Wiley.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1), 1–39.

Séverin, E., & Veganzones, D. (2021). Can earnings management information improve bankruptcy prediction models? *Annals of Operations Research, 306*(1), 247–272.

Shahriare S, Khair A, Abedin MZ (2021, December 19–21). Performance analysis of machine learning techniques that predict hotel booking cancellations in hospitality industry. In *ICCIT 2020: 23rd International Conference on Computer and Information Technology, Dhaka*.

Shajalal, M., Abedin, M. Z., & Uddin, M. M. (2021). Handling class imbalance data in business domain. In: *The essentials of machine learning in finance and accounting*. Taylor & Francis.

Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems, 57*, 41–56.

Veganzones, D., & Severin, E. (2020). Corporate failure prediction models in the twenty-first century: A review. *European Business Review, 33*(2), 204–226.

Verikas, A., Kalsyte, Z., Bacauskiene, M., & Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey. *Soft Computing, 14*(9), 995–1010.

Xiao, T., Zhu, J., & Liu, T. (2013). Bagging and boosting statistical machine translation systems. *Artificial Intelligence, 195*, 496–527.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using extreme learning machine and financial expertise. *Neurocomputing, 128*, 296–302.

# Assessing and Predicting Small Enterprises' Credit Ratings: A Multicriteria Approach

**Baofeng Shi**

**Abstract** Credit ratings play a key role in helping financial institutions to make loan decisions and to reduce the financial constraints on small and medium-sized enterprises. However, small enterprises have made it difficult for financial institutions such as commercial banks to accurately determine their credit risk, creating salient loan difficulties, due to the short duration, high frequency, urgent demand for credit, and small amount of their loans. In order to alleviate the difficulties of financing small businesses, this paper develops a new approach for the assessment of credit risk in small enterprises by combining high-dimensional attribute reduction methods with fuzzy decision-making methods. Based on 687 small enterprises in a regional commercial bank of China, we find 17 indicators that have a significant impact on the default risk of small enterprises. Then, it utilizes TOPSIS together with fuzzy C-means to grade the credit ratings of enterprises requesting loans. The standard discrimination and ROC curve dual tests resulted in the prediction accuracy of the standard indicator system reaching 85.40 percent and 90.09 percent, respectively, indicating the strong default discrimination of this rating system and its practicability in commercial banks and other financial institutions.

**Keywords** Credit rating · Default risk · Fuzzy C-means · Small enterprises

## 1 Introduction

China is the world's largest developing country, and small and medium-sized enterprises have developed rapidly. According to statistics, in 2021, Chinese SMEs contribute more than 80% of national employment, 60% of gross domestic

B. Shi (✉)
College of Economics and Management, Northwest A&F University, Xianyang, Shaanxi, China

Research Center on Credit and Big Data Analytics, Northwest A&F University, Xianyang, Shaanxi, China
e-mail: shibaofeng@nwsuaf.edu.cn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_8

product (GDP) (iResearch, 2021). Yet, small and medium-sized enterprises generally struggle to obtain financing, especially loans, severely restricting their development, due to unreliable financial information, loans of enormous volume but for low amounts, and diverse risks (Lu et al., 2022; Abedin et al., 2021; Ciampi & Gordini, 2013; Shi et al., 2016; Chi & Zhang, 2017; Ruan et al., 2018; Sun et al., 2022). To alleviate these financial difficulties, the Chinese Banking Regulatory Commission and other agencies requested the establishment of an "Inclusive Finance Business Division," provide financial services to small and micro businesses, and address issues affecting agriculture, rural areas, and farmers, and strengthen credit risk identification, surveillance, early warning and assessment of borrowers (CBRC, 2015; SCPRC, 2016, 2017).

Many scholars have conducted useful studies on the best way to assess the credit risk of loan-granting enterprises, in terms of the establishment of credit scoring, credit rating, and other systems. Dimensionless processing of statistics is typically necessary before a rating system can be implemented (Shi et al., 2015). In reality, the quantifiable financial data of small enterprises are less and more text-based non-financial data. As a result, researchers often use subjective Delphi method or analytic hierarchy process (AHPs) to process data without dimensions (Liang, 2007; Shi et al., 2018).

Regarding the development of indicator systems, Altman constructed Z-score and ZETA models based on financial indicators such as return on assets and pretax margins of asset interest to assess the probability of lender default (Altman, 1968; Altman et al., 1977). Gu et al. (2017) combined (AHP) with data envelopment analysis (DEA), using indicators such as the cash ratio, inventory turnover, and accounts receivable turnover ratio from the perspective of financial status, enterprise development, credit status, and internet financial status to predict defaults by enterprises that take out loans. This research has great reference value for creating a credit rating indicator system for small enterprises, but little of it studies wholesale and retail enterprises and uses distinctive default variables to forecast the credit outlook of loan customers.

Credit scoring models can be constructed using three methods: metrological statistics, fuzzy systems, and artificial intelligence. Metrological statistics consist of discriminant analysis, logistic regression, and linear regression (Reichert et al., 1983; Yurdakul & Iç, 2015; Iç & Yurdakul, 2010). Artificial intelligence methods include artificial neural nets (Marcano-Cedeño et al., 2011; Rui & Mendes, 2017; Chi et al., 2017), support vector machine (Hens & Tiwari, 2012; Harris, 2015; Tomczak & Zięba, 2015; Abedin et al., 2018; 2019a, b), a decision tree (Zhu & Hu, 2013; Florez-Lopez & Ramon-Jeronimo, 2015; Bahnsen et al., 2015; Zhang et al., 2017; Chern et al., 2021), ensemble learning (Abedin et al., 2022), and so forth. Recently, some academics have combined these methods with fuzzy evaluations and subsequently developed credit rating systems. Akkoç (2012) combined fuzzy evaluation and artificial intelligence to develop a credit rating system using a hybrid adaptive neuron fuzzy inference system predicting the risk of default of credit card holders in Turkey. The empirical research shows that this model is better at correctly averaged classification and wrongly estimated classification cost than liner discriminant analysis, logistic regression, and artificial neural nets. Bai et al. (2019) calculate the risk of default for farm lenders in a hybrid model using fuzzy C-means (FCM) and fuzzy rough sets. This study reveals the determinants of loan defaults, without grading their credit or including any decision function in their evaluation results.

To address this problem, some scholars have begun to divide consideration of credit ratings of loan customers into three credit rating models: scoring intervals of customer credit, establishing the threshold of default probability, and the loss given default (LGD) of loan customers. The Industrial and Commercial Bank of China (ICBC) (2005) divided the credit scores of its loan customers among 10 credit ratings into AA, AA-, so forth. Florez-Lopez (2007) estimated the default probability (PD) of loan applicants using statistical and artificial intelligence methods and classified the applicants into five rating categories. Chi and Zhang (2017) employed nonparametric models to construct a credit rating system specifically designed for small enterprises. They evaluate the credit ratings of loan customers according to their LGD. Therefore, credit rating models based on credit scoring intervals for customers give different results than models based on the threshold of default probability, so different loan approvers may give different results of credit rating for loan customers with those credit scores. The reason is that scoring intervals and the threshold of default probability are given ahead of time, and this increases the subjectivity of the ratings. With regard to the credit rating method based on LGD, a prerequisite is that the default loss of each customer must be known. However, default loss data are not available for some small enterprises that have only recently applied for loans, making this rating method infeasible.

Through our literature review, we find that there is no existing research that has a suitable rating indicator system to measure credit risk based on the loan characteristics of small wholesale and retail enterprises. In fact, industry differences among small enterprises lead to obvious heterogeneity in their estimation of loan and credit risks. For example, the statistics on credit at commercial banks show that the average maximum value of loans given to small enterprises in real estate development and operations is as much as 17 million Yuan (about USD 2.50 million) and that of small enterprises in wholesale and retail only amount to 0.41 million Yuan (Bank of Dalian, 2014). When comparing these two types of companies in the same credit risk system, even if the default model false positive is very low, the bank will suffer completely different losses. Therefore, different credit rating models are required for different industries, based on the fact that they are small enterprises, to distinguish their credit risk from that of other kinds of enterprises.

In view of the foregoing, this paper makes three contributions to the literature. First, in the category of credit rating, it adds to the literature by focusing on Chinese small wholesale and retail enterprises. Second, by establishing suitable credit rating models for small wholesale and retail enterprises, it offers a decision-making reference for credit rating by commercial banks, microcredit organizations, and these enterprises. Third, we propose a credit scoring measurement process by using triangular fuzzy numbers for non-financial data at small wholesale and retail enterprises, which helps to avoid the subjectivity and randomness caused by expertise scoring and makes the quantified processed qualitative indicator more accurate.

The paper is organized as follows. Section 2 introduces credit rating models for small enterprises. Section 3 builds the rating system based on credit data for 687 small wholesale and retail enterprises seeking loans from an urban commercial bank in China. Section 4 offers our main conclusion and lists the innovative aspects of this paper.

## 2 Methodology

First, we set up an assessment system based on the characteristics of small wholesale and retail loans. Second, TOPSIS is used to obtain credit scores based on the indicator weights computed as entropy weights. Finally, fuzzy C-means is used to evaluate the credit ratings of loan customers. The framework can be seen in Fig. 1.

### *2.1 Establishment of a Credit Rating System*

The establishment of this credit rating system is done in two steps. Firstly, initial data must be standardized to eliminate incompatibility between different measurement measures. Second, probit regression and partial correlation analysis are combined to create quantitative screening to reduce the number of indicators.

**Pre-Processing of Indicator Data**

1. *Pre-Processing of Qualitative Indicator*

   Qualitative indicators cannot be directly quantified but, rather, are described narratively. For instance, the indicator for education background has five possible values: "Primary school diploma," "junior high school diploma," "senior high school diploma," "junior college diploma," and "bachelor's degree or above." Qualitative indicators have an advantage similar to that of triangular fuzzy numbers in how they process data with diverse characteristics. To quantify the qualitative indicators, they must be transformed to triangular fuzzy numbers according to their semantics; then, defuzzification is used, that is, triangular fuzzy numbers are transformed to fixed values.

   Let $A$ be a fuzzy set for $x \in U$, if $\mu_A(x) \in [0, 1]$, then $\mu(x)$ is the membership of $x$ to $U$, and $\mu_A$ represents the membership function of $x$. Further, le $l$ and $u$ be the lower and upper limit of the fuzzy number, respectively, and let $m$ be the median value, then the fuzzy number $(l,m,u)$ can be shown in Fig. 2. Its membership function $\mu_A$ is presented in Eq. (1) (Promentilla et al., 2008). Typically, three, five, and seven triangular fuzzy numbers are used (Cheng et al., 2008; Khalili-Damghani et al., 2013; Wang et al., 2016), as illustrated in Figs. 3, 4, and 5 (Chai et al., 2019).

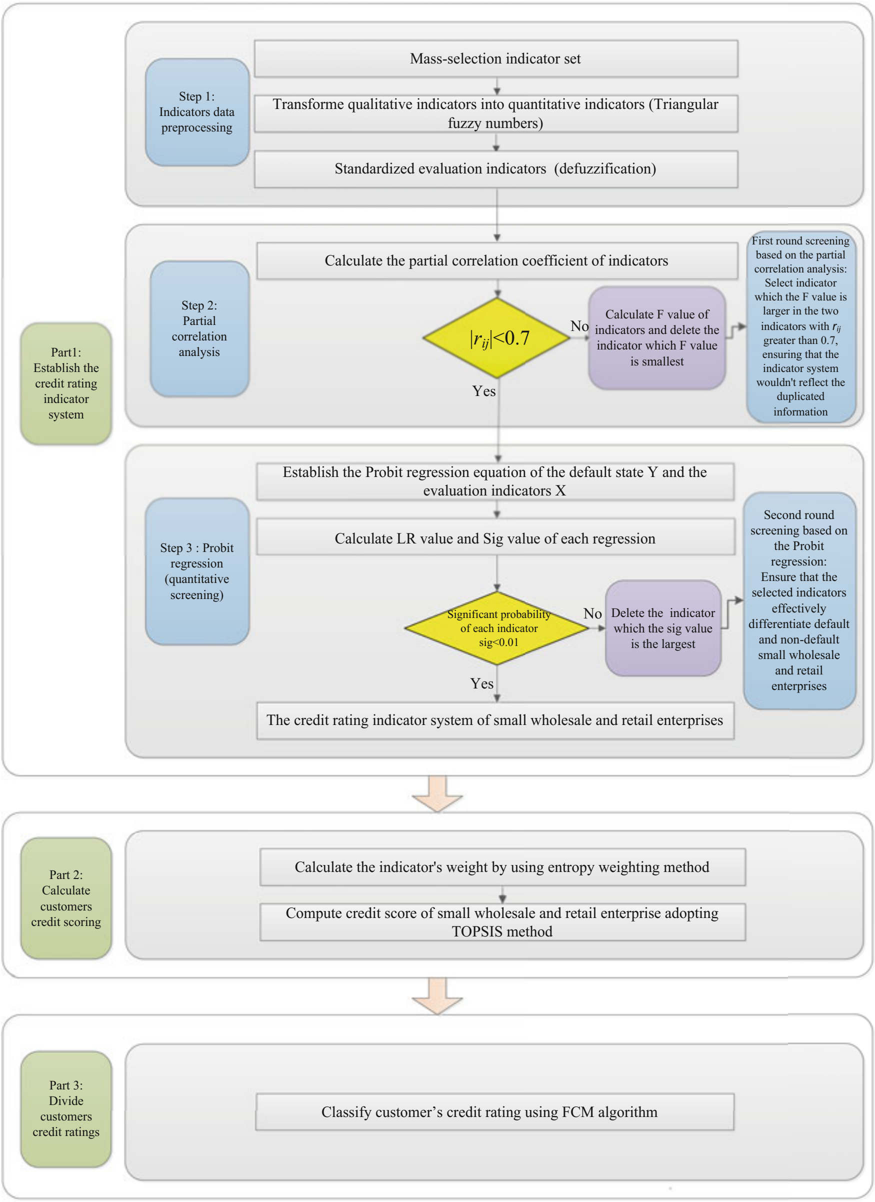



**Fig. 1** Framework of the credit rating model

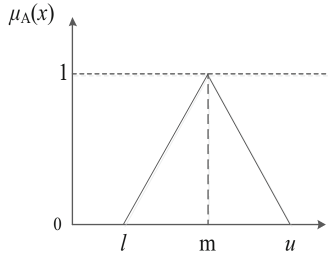



**Fig. 2** Triangular fuzzy numbers (TFNs)

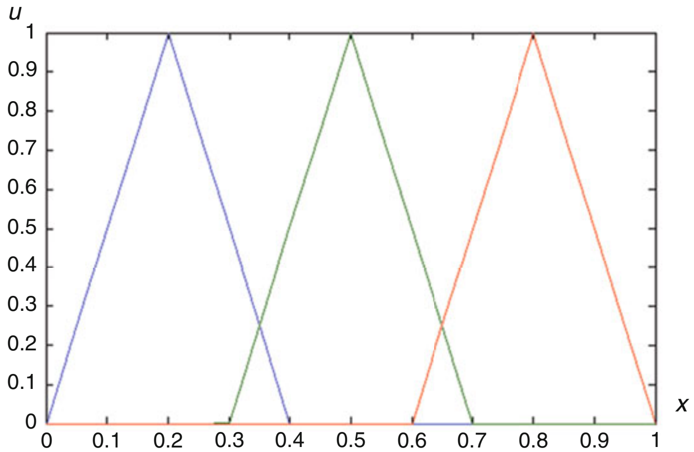



**Fig. 3** TFNs with three classifications

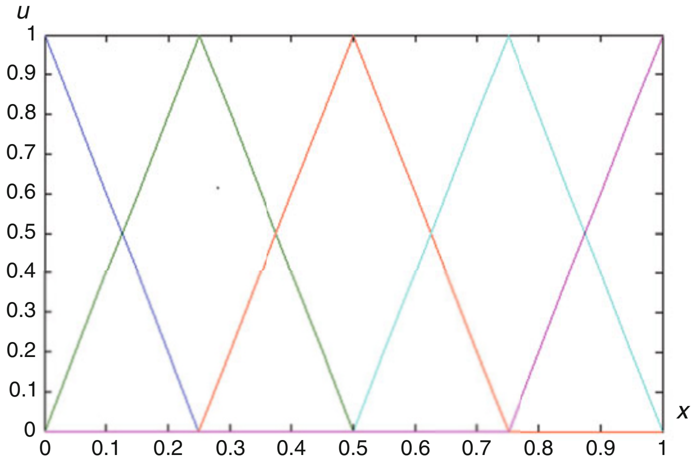



**Fig. 4** TFNs with five classifications

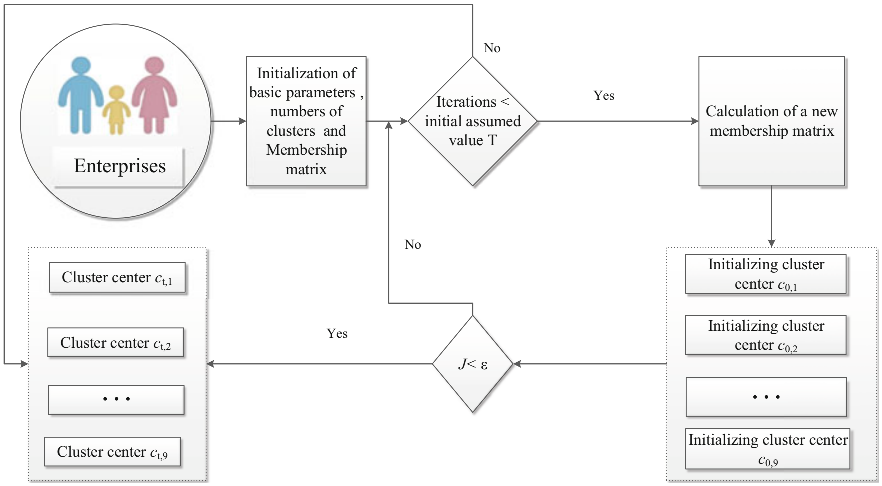



**Fig. 5** The framework for dividing credit ratings using the FCM method

$$\mu_A(x) = \left\{ \begin{array}{ll} 0 & x < l, \\ \dfrac{x-l}{m-l} & l < x < m, \\ \dfrac{u-x}{u-m} & m \leq x \leq u, \\ 0 & x > u, \end{array} \right\}. \tag{1}$$

Let $A_{\max}$ be the defuzzified value, then when combined with Eq. (1), $A_{\max}$ is given as follows (Wu et al., 2016):

$$A_{\max} = (l + m + u)/3. \tag{2}$$

2. *Pre-Processing of Quantitative Indicator*
   Quantitative indicators usually include four types of indicators, namely positive, negative, interval, and moderating indicators. We can use the max-min standardization for the indicators (Chi & Zhang, 2017; Shi et al., 2018; Abedin et al., 2019a, b); to avoid repetition, it is not described here.

**Reduction of Attributes**

1. *The First Indicator Screening Based on Partial Correlation Analysis*
   In the same standard layer, partial correlation analysis (PCA) is used to remove redundant indicators. Let $x_{ij}$ be the value of indicator $i$ for enterprise $j$, $r_{ik}$ be the correlation coefficient between indicators $i$ and $k$, then $r_{ik}$ is defined as follows:

$$r_{\mathrm{ik}} = \frac{\sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i\right)\left(x_{ij} - \bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i\right)^2} \sqrt{\sum_{j=1}^{n} \left(x_{ij} - \bar{x}_k\right)^2}}, \tag{3}$$

where $n$ is the number of enterprises, and $\bar{x}_i$ is the average value of indicator $i$. Suppose that $R$ is the correlation matrix composed of $r_{ik}$, and $m$ is the number of variables at the criterion level. The correlation matrix $R$ is given as follows:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r \end{bmatrix}. \tag{4}$$

The inverse matrix $C$ of the correlation matrix $R$ is:

$$C = R^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c \end{bmatrix}. \tag{5}$$

Then, the partial correlation coefficient of indicator $i$ and indicator $k$ can be obtained:

$$r'_{ik} = \frac{-c_{ik}}{\sqrt{c_{ii}c_{kk}}}. \tag{6}$$

The larger the partial correlation coefficient $r'_{ik}$, the stronger the relativity between indicators $i$ and $k$. When $\left|r'_{ik}\right| > 0.7$, $F$ test (Nami & Shajari, 2018) is employed to perform the evaluation of the two indicators. Subsequently, the indicator with a lower $F$ value is removed.

2. *The Second indicator Screening Based on Probit Regression*

   In the same standard layer, the maximum likelihood function is employed to obtain the probit regression coefficients between the $m$ indicators and the default $y_j$, and to determine the LR statistics of each indicator. Using $\chi^2$, we remove the indicator with the largest sig but that shows the least remarkable effects on defaults among the indicator with a significance probability (Sig $>$ 0.01), and complete the screening of the first indicator. The remaining $m - 1$ indicators, will be screened in the same manner as above until the corresponding significance probability of each indicator fails to exceed 0.01, i.e., Sig $\leq$ 0.01. Then the indicator screening is done. Now, the remaining indicators can all significantly distinguish the defaults of small enterprises. The specific resolution equation is as follows.

Let $\boldsymbol{X_j} = (x_{1j}, x_{2j}, \ldots, x_{mj})$ be the row vector of enterprise $j$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ be the regression coefficient vector of indicators; $m$ denotes the number of indicators; $\varphi(z_j)$ is the standardized normal cumulative distribution function, $P(Y_j = 1)$ indicates the probability of default; and $z_j = \alpha + \boldsymbol{X_j\beta}$. Then,

$$P\left(Y_j = 1\right) = \phi\left(z_j\right) = \int_{-\infty}^{z_j} \frac{1}{\sqrt{2\pi}} \ell^{\left(\frac{-s^2}{2}\right)} ds. \tag{7}$$

The maximum likelihood method can be used to predict the indicators in the probit model. Its log-likelihood function is defined as follows:

$$\max\ ln\, L = \sum\nolimits_{j=1}^{n} \left[y_j\, ln\left(\phi\left(z_j\right) + \left(1 - y_j\right) \ln\left(1 - \phi\left(z_j\right)\right)\right]. \tag{8}$$

In Eq. (8), the larger the log-likelihood function LnL, the more accurate estimate of default $Y_j$.

Suppose that $\mathrm{LR}_k$ is the LR statistic value for indicator $k$, $\sigma_{\beta k}$ is the standard error of regression coefficient $\beta_k$, $\tilde{\beta}_k$ is the estimated parameter value, $\widehat{\sigma}_{\beta k}$ is the standard error of the estimated parameter value, and $\widehat{\beta}_k$ as well as $\widehat{\sigma}_{\beta k}$ are independently the estimated value and standard error beyond constraints. Then:

$$\mathrm{LR}_k = -2\left[\log L\left(\tilde{\beta}_k,\ \tilde{\sigma}_{\beta k}^2\right) - \log L\left(\widehat{\beta}_k,\ \widehat{\sigma}_{\beta k}^2\right)\right]. \tag{9}$$

## 2.2 *Solution to Credit Scoring*

Entropy weight is a method of describing the differences in information between indicators based on entropy in information in evaluated statistics; it has often been used in evaluation of complex systems (Chi & Zhang, 2017; Bai & Zhao, 2022). In this section, entropy is used to calculate the evaluation indicator weight $W = (w_i)$ in the first place; then TOPSIS is used to obtain credit scores (Yurdakul & Iç, 2015; Iç & Yurdakul, 2010; Wang & Leng, 2021). The procedure is presented as follows:

*Step 1*: Obtain the best and worst scores of the indicators.

Suppose that $b_i^+$ and $b_i^-$ are the best and worst scores of indicator $i$, respectively, and $b_{ij}$ is the score for enterprise $j$; so

$$b_i^+ = \begin{cases} \max\left(b_{ij}\right), i \text{ denotes the } i\text{th positive indicator} \\ \min\left(b_{ij}\right), j \text{ denotes the } j\text{th negative indicator} \end{cases}. \tag{10}$$

$$b_i^- = \begin{cases} \min\left(b_{ij}\right), i \text{ denotes the } i\text{th positive indicator} \\ \max\left(b_{ij}\right), j \text{ denotes the } j\text{th negative indicator} \end{cases}. \tag{11}$$

*Step 2*: The standardized score is obtained, and the difference between the best and worst scores are calculated. Suppose that $d_j^+$ (and $d_j^-$ ) are the differences between the best (worst) score and the actual score of enterprise $j$. Then,

$$d_j^+ = \sqrt{\sum\nolimits_{i=1}^{m} \left(w_i b_{ij} - w_i b_i^+\right)^2}, d_j^- = \sqrt{\sum\nolimits_{i=1}^{m} \left(w_i b_{ij} - w_i b_i^-\right)^2}. \tag{12}$$

*Step 3*: Independently solve for the difference between the best and worst scores and the relative closeness of the credit scores. Suppose that $c_j$ is the relative closeness of the score, and $P_j$ be the credit score:

$$P_j = c_j = \frac{d_j^-}{d_j^- + d_j^+}. \tag{13}$$

*Step 4*: The credit score $P_j$ in Eq. (13) range from 0 to 1, which are not consistent with the customary scoring regulations on a scale of 100. In view of this, we standardize $P_j$ to render it in a period from 0 to 100.

$$S_j = \frac{P_j - \min\left(P_j\right)}{\max\left(P_j\right) - \min\left(P_j\right)} \times 100, \tag{14}$$

where $S_j$ is the standardized credit score of enterprise $j$.

This paper employs default discrimination and a ROC curve to evaluate the predictive performance of the system for small enterprises as follows: if the credit score of a rating system meets the requirement that "all the credit scores of non-defaulting small enterprises are higher than those of small defaulting enterprises" the stronger the evaluation ability of the indicator system on the defaults of loan enterprises becomes, the fewer the losses of financial institutions such as banks. In agreement with Chi and Zhang (2017), the rationality of the indicator system is determined.

$$S_c^1 = \frac{1}{m} \sum\nolimits_{j=1}^{m} S_j^1, \tag{15}$$

$$S_c^0 = \frac{1}{n} \sum\nolimits_{j=1}^{n} S_j^0, \tag{16}$$

$$s_c = \frac{\frac{1}{m} \sum\nolimits_{j=1}^{m} S_j^1 + \frac{1}{n} \sum\nolimits_{j=1}^{n} S_j^0}{2}, \tag{17}$$

where $S_c^0$ and $S_c^1$ denote the average value of the credit scores of non-defaulting and defaulting samples, respectively, $S_c = (S_c^1 + S_c^0)/2$.

ROC was first applied by Sobehart and Keenan (2001) to evaluate the accuracy of credit ratings. First, the sensitivity and specificity of the credit rating system are obtained. Given that the number of correctly determined defaulting samples ($y_j = 1$)

is TP (true positive); the number of incorrectly determined defaulting samples is FN (false negative); the number of correctly classified non-defaulting samples ($y_j = 0$) is TN (true negative); the number of incorrectly non-defaulting samples is FP (false positive), sensitivity and specificity can be calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{18}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}. \tag{19}$$

Then, sensitivity and specificity can be used to draw the ROC curve of the system. The larger the area under the ROC curve, the stronger the system's capacity to recognize defaulting samples.

## *2.3 Dividing Credit Ratings of Loan Customers*

In contrast to conventional cluster algorithms, fuzzy cluster algorithms do not require strict identification of objects belonging to specific classes, demonstrating flexible attribute requirements. Thus, it fits the special requirement that the initial indicator information is a value of a triangular fuzzy function. Therefore, this paper follows Bai et al.'s (2019) fuzzy C-means (FCM) algorithm, in rating the credit of small enterprises. The principle is shown in Fig. 5.

The FCM compares each sample with all clusters using real values $u_{ij}$, ranging from 0 to 1, reflecting the degree of membership of indicator $j$ in category $i$.

FCM divides the $m$ vectors $S_j(j = 1,2,\cdots,m)$ into $c$ fuzzy clusters, and calculates the center of each cluster so that non-similarity objective function is minimized. Its objective function $J(U, c_1, \cdots, c_c)$ (Yu et al., 2010) is:

$$J(U, c_1, \cdots, c_c) = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j=1}^{m} \left(u_{ij}\right)^n d^2\left(x_j, c_i\right), \tag{20}$$

where $d(S_j, c_i)$ is the Euclidean distance of the clustering center $c_i$ in the sample $S_j$; $n \in [1, \infty)$ is the weighting indicator, controlling the shared degree of the classified objects in the fuzzy category.

Its structure is shown as the following objective function $\bar{J}(U, c_1, c_2 \cdots, c_c, \lambda_1, \cdots \lambda_m)$ (Sun et al., 2022):

$$\begin{aligned} \bar{J}(U, c_1, c_2 \cdots, c_c, \lambda_1, \cdots \lambda_m) &= J(U, c_1, c_2 \cdots, c_c) + \sum_{j=1}^{m} \lambda \left( \sum_{i=1}^{c} u_{ij} - 1 \right) \\ &= \sum_{i=1}^{c} \sum_{j=1}^{m} \left(u_{ij}\right)^n d_{ij}^2 + \sum_{j=1}^{m} \lambda \left( \sum_{i=1}^{c} u_{ij} - 1 \right) \end{aligned} \tag{21}$$

In this equation, $\lambda_j$ is the Lagrange multiplier; $c_i$ and $u_{ij}$ are defined as follows (Demircan & Kahramanli, 2016):

$$c_i = \frac{\sum_{j=1}^{m} \left(u_{ij}\right)^n S_j}{\sum_{j=1}^{m} \left(u_{ij}\right)^n}. \tag{22}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{n-1}}}. \tag{23}$$

The basic steps of the FCM cluster algorithm are as follows under these two conditions:

1. The number of clusters $c$ is given, $1 < c \leq m$, and $m$ is the number of samples. Given that $T$ is the maximum number of iterations, $\varepsilon$ is the threshold, and $\omega$ is the fuzzy number; the indicator setting iterative counter $t = 0$.
2. Rectify partition matrix $U^{(\mathrm{t})}$ using Eq. (21).
3. Obtain the new cluster center $c_c(t)$ using Eq. (20).
4. $t \leftarrow t + 1$; repeat steps 2 and 3 until $t \geq T$ or$|U^{(t)} - U^{(t-1)}| \leq \varepsilon$.

# 3 Empirical Analysis

## 3.1 Sample Selection and Data Sources

This paper uses credit statistics on 687 small retail and wholesale enterprises, representing customers of a Chinese commercial bank, to validate the model developed in Sect. 2. Further details about the credit rating indicators and default status of these 687 small wholesale and retail enterprises are as follows. We select credit rating indicators first using the standard variables of rating agencies such as Standard & Poor, Moody, and Fitch (Standard and Pool's Services, 2011; Fitch Ratings, 2013; Dagong, 2010), and second from papers on credit rating (Mijid & Bernasek, 2013; Hai et al., 2013; Shi & Chi, 2014; Shi et al., 2016; Abedin et al., 2018, 2019a, b; Sun et al., 2022). That is, a total of 107 indicators are selected on repayment ability and willingness to repay, and so forth. These indicators cover seven secondary standard layers such as financial factors, non-financial factors, and the personal situation of the legal representative of small wholesale and retail enterprises. Furthermore, we eliminated 26 indicators for which statistics are unavailable, leaving 81 indicators, as shown in Table 1.

## 3.2 Credit Rating of Small Wholesale and Retail Enterprises

1. *Establishment of a Credit Risk Evaluation Indicator System*
   The original and standardized data on 687 small enterprises are shown in Tables 2 and 3, respectively.

**Table 1** Screening criteria for indicators of small enterprise credit rating

| (1) No. | (2) First criterion level | (3) Second criterion level | (4) Third criterion level | (5) Indicators | (6) Type | (7) Screening result |
|---|---|---|---|---|---|---|
| 1 | Repayment ability | Financial factors | Solvency | Debt asset ratio | Negative | Probit delete |
| ... | | | | ... | ... | ... |
| 28 | | | | Source of repayment | Qualitative | Unobservable |
| ... | | | ... | ... | ... | ... |
| 55 | | | Growth capacity | Revenue growth | Positive | Pass |
| ... | | | | ... | ... | ... |
| 63 | | | | Wages, welfare growth rate | Positive | Unobservable |
| 64 | | External macroeconomic conditions | | Industry sentiment index | Positive | Pass |
| ... | | | | ... | ... | ... |
| 72 | | | | Economic environment | Qualitative | Unobservable |
| 73 | | Internal non-financial factors | | Years of relevant industry | Qualitative | Probit delete |
| ... | | | | ... | ... | ... |
| 86 | Willingness to repay | Legal person situation | | Education background | Qualitative | Pass |
| ... | | | | ... | ... | ... |
| 98 | | | | Owner qualities | Qualitative | Unobservable |
| 99 | | Enterprise credit situation | | Registered capital classification | Qualitative | Partial correlation analysis delete |
| ... | | | | ... | ... | ... |
| 103 | | Commercial reputation | | Tax records | Qualitative | Partial correlation analysis delete |
| 104 | | | | Legal disputes | Qualitative | Probit delete |
| ... | | | | ... | ... | ... |
| 106 | | | | No. of breaches of contract | Qualitative | Probit delete |
| 107 | Pledge guarantee factor | | | Mortgage/pledge/guarantee | Qualitative | Probit delete |

**Table 2** Original data for a sample of small retail and wholesale enterprises

| | | | Original data | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 681 non-defaulting enterprises | | | 6 defaulting enterprises | | |
| (a) No. | (b) Criterion level | (c) Indicators | (1) C001 | … | C681 | (682) C682 | … | (687) C687 |
| 1 | $C_1$ internal non-financial factors | $X_1$ years of relevant industry | 8 | … | 10 | 8 | … | 10 |
| … | | … | … | … | … | … | … | … |
| 10 | $C_2$ legal person situation | $X_{10}$ education background | Junior diploma | … | Bachelor's degree | N/A | … | Bachelor's degree |
| … | | … | … | | … | … | … | … |
| 20 | | $X_{20}$ the value of car and real estate of legal representatives | 1000 | … | 1000 | N/A | … | 100 |
| 21 | $C_3$ Enterprise credit situation | $X_{21}$ registered capital classification | Found | … | Found | 0.917 | … | 0.917 |
| … | ... | … | … | … | … | … | … | … |
| 27 | $C_5$ operating capacity | $X_{27}$ accounts receivable turnover rate | 5.00 | … | 13.19 | 0 | … | 9.17 |
| … | | … | … | … | … | … | … | … |
| 36 | | $X_{36}$ cash conversion cycle | −3973.69 | … | 7.50 | N/A | … | 2.72 |
| 37 | $C_6$ profitability | $X_{37}$ rate of return on common stockholders' equity | 0.078 | … | 0.003 | 0.000 | … | 0.280 |
| … | | … | … | … | … | … | … | … |
| 49 | | $X_{49}$ operating activities generate cash inflows | 112,458,001 | … | 625,800,630 | 0.000 | … | 26,139,847.75 |
| 50 | $C_7$ growth capacity | $X_{50}$ operating income growth rate | 0.000 | … | 0.023 | 0.00 | … | 1.36 |
| … | | … | … | … | … | … | … | … |
| 54 | | $X_{54}$ retained revenue growth rate | 0.076 | … | 1.251 | 0.510 | … | 0.507 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 55 | $C_8$ solvency | $X_{55}$ debt asset ratio | 6.84 | . . . | 0.56 | 0 | . . . | 0.604 |
| . . . | | . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| 74 | | $X_{74}$ EBITDA/total debt ratio | 0.043 | . . . | 0.003 | −0.04 | . . . | 0.49 |
| 75 | $C_9$ external macro-economic conditions | $X_{75}$ industry sentiment index | 137.45 | . . . | 139.50 | 137.45 | . . . | 127.20 |
| . . . | | . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| 80 | | $X_{80}$ Engel coefficient | 39.4 | . . . | 37.0 | 39.40 | . . . | 37.90 |
| 81 | $C_{10}$ pledge guarantee factor | $X_{81}$ mortgage/pledge/guarantee | The guarantee amount is 5 million yuan | . . . | No guarantee | The guarantee amount is 18.9 million yuan | . . . | The guarantee amount is 3 million yuan |
| 82 | —— | Default | 0 | . . . | 0 | 1 | . . . | 1 |

**Table 3** Standardized data

| (*a*) No. | (*b*) Criterion level | (*c*) Indicator | Standardized Data | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 681 non-default enterprises | | | 6 default enterprises | | |
| | | | C001 | … | C681 | C682 | … | C687 |
| 1 | $C_1$ internal non-financial factors | $X_1$Years of relevant industry | 0.917 | … | 0.917 | 0.917 | … | 0.083 |
| … | | … | … | … | … | … | … | … |
| 9 | | $X_9$ the proportion of total amount of money returned by enterprises through the bank | 0.667 | … | 1.000 | 0.000 | … | 0.000 |
| … | … | … | … | … | … | … | … | … |
| 81 | $C_{10}$ pledge guarantee factor | $X_{81}$ mortgage/pledge/guarantee | 0.650 | … | 0.000 | 0.000 | … | 0.700 |
| 82 | —— | Default | 0 | … | 0 | 1 | … | 1 |

Taking C1 enterprise's internal non-financial factors as an example, the process of partial deleting correlation indicator is illustrated (see Table 3). We put data on nine indicators related to "internal non-financial factors at enterprise C1" in Table 3 into Eqs. (3)–(6), so as to calculate $r_{kj}$, the partial correlation coefficient of the indicators. We respectively calculate the F-statistic of the indicator pairs whose partial correlation coefficients are over 0.7. Then we delete an indicator with a smaller F-statistic and retain the other one. The result is shown in Table 4. The rest can be done in the same manner. Using PCA, this paper removes 14 indicators with redundant information.

After deleting some indicators with PCA, we screen the remaining indicators in all standard layers through probit regression, and select the indicators with remarkable discriminatory power on defaulting status. Then we put the remaining 67 indicator data screened by partial correlation in Table 3 into Eqs. (7)–(9) and screen them using Stata. The 17 remaining screened indicators are in Table 5.

2. *Solution to Credit Scoring of Small Wholesale and Retail Enterprises*

The weight of 17 variables is calculated by the entropy weight in Table 5. With Eqs. (10)–(13), it is easy to calculate the credit scores of the enterprises. The result is presented in Table 6.

Then, we put the credit scores of these enterprises in Eqs. (14)–(16) and subsequently obtain the prediction accuracy of 85.40%. The result of the model classification is presented in Table 7, and the corresponding ROC curve is presented in Fig. 6, where the area under ROC curve (AUC) is 0.909, suggesting the strong predictive accuracy of the defaulting status of small enterprises obtained using the screened 17 indicators.

**Table 4** Partial correlation deletion indicator related to "Internal non-financial factors"

| (1) No. | Indicators with a partial correlation coefficient greater than 0.7 | | | | (6) Partial correlation coefficient | (7) Deleted indicator |
|---|---|---|---|---|---|---|
| | (2) Indicator 1 | (3) F-statistic of indicator 1 | (4) Indicator 2 | (5) F-statistic of indicator 2 | | |
| 1 | $X_{55}$ debt asset ratio | 2.370 | $X_{63}$ shareholder equity ratio | 2.392 | 0.993 | $X_{55}$ debt asset ratio |
| 2 | $X_{56}$ current liabilities operating ratio | 1.284 | $X_{73}$ Total debt operating activity net cash flow ratio | 0.907 | 0.967 | $X_{73}$ Total debt operating activity net cash flow ratio |
| 3 | $X_{57}$ quick ratio | 0.079 | $X_{68}$ cash ratio | 0.753 | 0.809 | $X_{68}$ cash ratio |

**Table 5** Credit indicators weights for small wholesale and retail enterprises

| (a) No. | (b) Indicators | (c) Weight | Standardized data (1) C001 | ... | (687) C687 |
|---|---|---|---|---|---|
| 1 | $X_{10}$ education background | 0.025 | 0.500 | ... | 0.700 |
| 2 | $X_{13}$ gender | 0.003 | 1.000 | ... | 1.000 |
| 3 | $X_{14}$ age | 0.006 | 0.970 | ... | 0.848 |
| 4 | $X_{18}$ family monthly income | 0.172 | 0.071 | ... | 0.071 |
| 5 | $X_{19}$ time in current position | 0.047 | 0.250 | ... | 0.250 |
| 6 | $X_{20}$ the value of car and real estate of legal representatives | 0.095 | 0.917 | ... | 0.917 |
| 7 | $X_{31}$ fix capital ratio | 0.197 | 0.003 | ... | 0.029 |
| 8 | $X_{50}$ operating income growth rate | 0.033 | 0.197 | ... | 0.201 |
| 9 | $X_{51}$ profit growth rate | 0.001 | 0.494 | ... | 0.530 |
| 10 | $X_{52}$ Total asset growth rate | 0.027 | 0.271 | ... | 0.298 |
| 11 | $X_{53}$ capital accumulation rate | 0.001 | 0.496 | ... | 0.496 |
| 12 | $X_{54}$ retained revenue growth rate | 0.017 | 0.510 | ... | 0.518 |
| 13 | $X_{75}$Industry sentiment index | 0.001 | 0.633 | ... | 0.833 |
| 14 | $X_{77}$ per capita disposable income of urban and rural residents at the end of the year | 0.001 | 0.300 | ... | 0.002 |
| 15 | $X_{78}$ residential price index | 0.000 | 0.817 | ... | 0.988 |
| 16 | $X_{79}$ per capita disposable income of urban residents | 0.007 | 0.155 | ... | 1.000 |
| 17 | $X_{80}$ Engel coefficient | 0.001 | 0.576 | ... | 0.821 |

**Table 6** Credit scoring of small enterprises

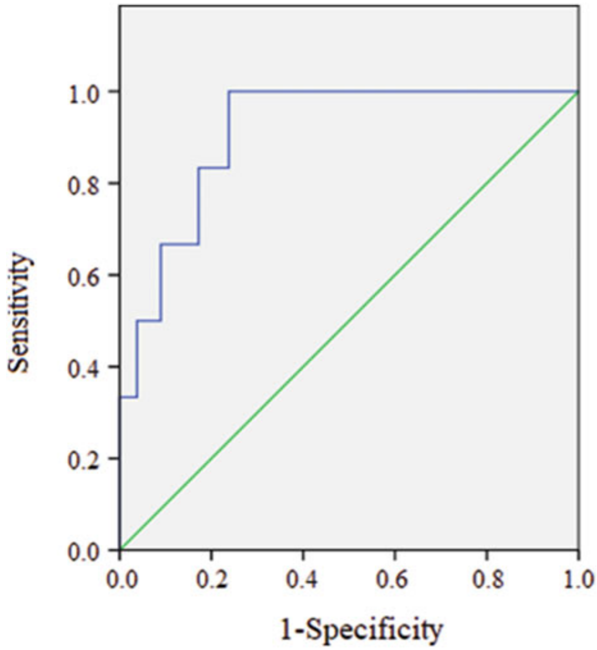
| (1) No. | (2) Loan No. | (3) Original credit score $P_j$ | (4) Standardized credit score $S_j$ |
|---|---|---|---|
| 1 | 200410270004 | 0.391 | 48.846 |
| 2 | 200412150123 | 0.243 | 0.759 |
| ... | ... | ... | ... |
| 687 | X2012060800099 | 0.453 | 89.149 |

**Table 7** Classification of credit rating system

| Actual default status | Model prediction result 1 (Default) | 0 (Non-default) | Sum |
|---|---|---|---|
| 1 (default) | 4 | 2 | 6 |
| 0 (non-default) | 96 | 585 | 681 |
| Sum | 100 | 587 | 687 |

3. *Credit Rating of Small Wholesale and Retail Enterprises*
   According to credit rating procedures, first we set the number of credit rating clusters to 9; the maximum number of iterations $T = 1000$; the threshold $\varepsilon = 1\text{E}{-5}$; and the fuzzy number $\omega = 2$ (Zhong et al., 2014; Robillard et al., 2014). Then, we use the vector $S_j$ of credit scores in MATLAB to get the corresponding data distribution and classification into clusters, as shown in Figs. 7 and 8; the changing trends in the objective functions are shown in

Fig. 6 ROC curve (AUC = 0.909)



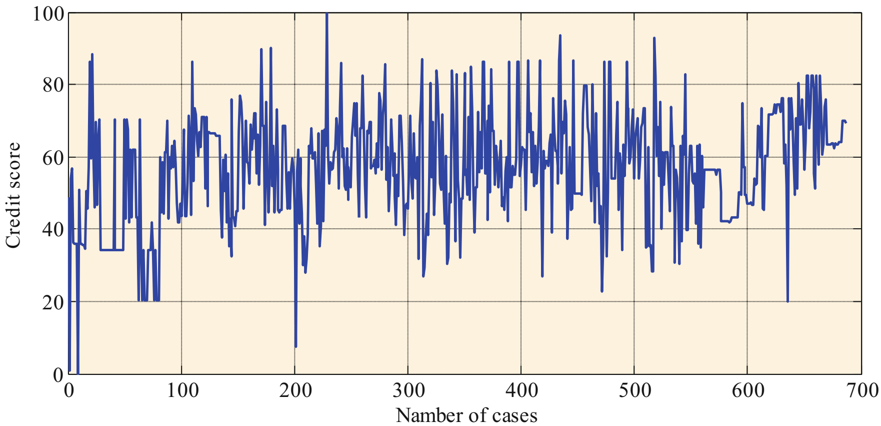




**Fig. 7** Distribution of credit score data for 687 small wholesale and retail enterprises

Fig. 9. Finally, the credit scores of cluster centers are presented in Table 8 to obtain nine corresponding ratings (AAA, AA, . . ., C). Using the upper and lower limits of credit scores, the credit score intervals can be obtained for customers in different clusters (Table 8).

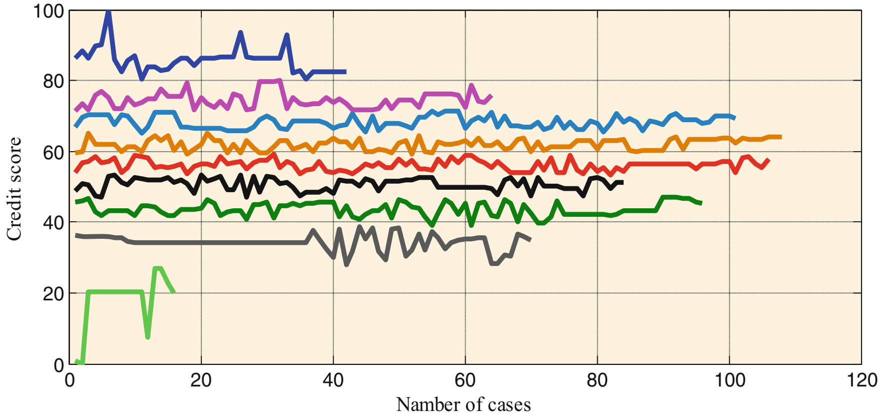
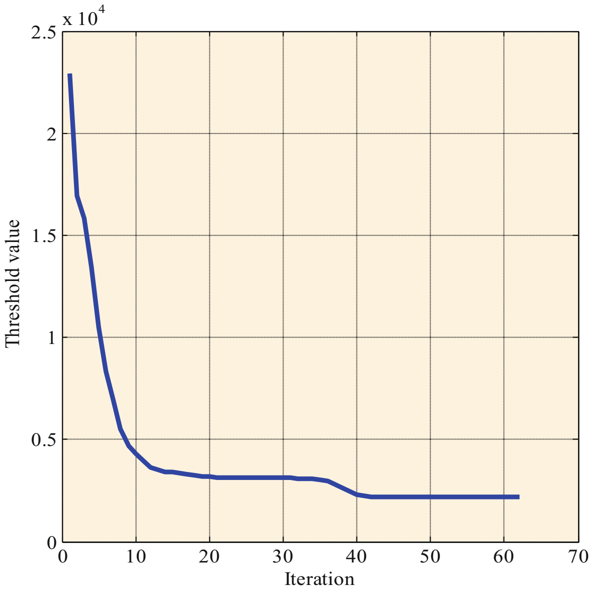



**Fig. 8** The classification of nine cluster centers

**Fig.9** The changing trend of credit rating division objective function




## 4 Conclusion

Small and medium-sized enterprises are important for the economic development of China. However, because of imperfect financial information, urgent demand for loans but small amount of loan business, dispersed risks, and the absence of necessary guarantees, small enterprises have made it difficult for financial institutions such as commercial banks to depict their credit risks precisely, thus bringing about salient loan difficulties in terms of financing and high loan prices. This paper

**Table 8** The credit rating for small enterprises

| (1) No. | (2) Cluster center of credit score | (3) Credit rating | (4) Credit score interval | (5) Number of cases |
|---|---|---|---|---|
| 1 | 85.497 | AAA | [80.447, 100] | 32 |
| 2 | 74.423 | AA | [71.347, 80.447) | 60 |
| 3 | 68.251 | A | [65.264, 71.347) | 54 |
| 4 | 62.147 | BBB | [59.232, 65.264) | 68 |
| 5 | 56.153 | BB | [53.468, 59.232) | 120 |
| 6 | 50.746 | B | [47.179, 53.468) | 73 |
| 7 | 43.464 | CCC | [39.083, 47.179) | 79 |
| 8 | 34.279 | CC | [27.826, 39.083) | 68 |
| 9 | 19.883 | C | [0, 27.826) | 124 |

uses a sample of 687 small enterprises to develop a credit rating system for these enterprises using a combination of metrological statistics and fuzzy decision. To begin with, we use partial correlation analysis to eliminate indicators with repeated information and Probit regression to screen indicators that markedly influence the defaulting status of small enterprises, establishing a credit risk evaluation indicator system composed of 17 indicators such as "X18 family monthly income" and "X20 the value of car and real estate of legal representatives" for these enterprises. Second, the credit scores of loan enterprises are calculated using the entropy-weighting TOPSIS method. Finally, a fuzzy C-means (FCM) algorithm is used to evaluate the credit ratings of small enterprises. The proposed system, through defaulting state testing, shows the predictive accuracy of 85.40% and 90.09%, respectively, confirming a high default predictive capacity, which can be useful for commercial banks.

This study is innovative in the following three respects. Firstly, the study proposes a credit rating system consistent with the credit characteristics of small retail and wholesale enterprises. It is an effective complement to existing credit rating literature and can act as a decision-making reference for commercial banks and small wholesale and retail enterprises in their credit rating. Second, triangular fuzzy numbers are introduced into the scoring process, leading to the objective arbitrariness. Third, the empirical research in this study shows that, for small retail and wholesale enterprises, non-financial indicators are more important for the prediction of default risks than financial factors. According to Fig. 5, among the 17 influential rating indicators, the sum of the weights of non-financial factors and external micro indicators is 0.752, which is much higher than 0.248, the weight of internal financial indicators. Thus, non-financial factors and external microeconomic conditions are more important factors in influencing small and medium-sized wholesale and retail credit ratings; non-financial factors should be investigated in terms of the prediction of small enterprises' default.

The study progressed in the development of credit rating systems for small wholesale and retail companies, but there were still some limitations. Due to the difficulty of getting real default losses data from loan companies, this paper uses default status $y_i$ only as a dependent variable. This rating method has difficulty in explaining the objective reality that two different customers who default at the same time cause different losses to the same bank. With the accumulation of default data and the advance of data analysis technology, further breakthroughs and research on these problems can be produced.

**Acknowledgments** The study was supported by the National Natural Science Foundation of China (Nos: 71873103, 72173096, 71503199 and 71731003), the Social Science Foundation of Shaanxi Province, China (No. 2018D51), the Tang Scholar Program of Northwest A&F University, China (No. 2021-04).

## References

Abedin, M. Z., Chi, G. T., Colombage, S., & Moula, F. E. (2018). Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27.
Abedin, M. Z., Chi, G. T., Moula, F. E., Zhang, T., & Hassan, M. K. (2019a). An optimized support vector machine intelligent technique using optimized feature selection methods: Evidence from Chinese credit approval data. *Journal of Risk Model Validation, 13*(2), 1–46.
Abedin, M. Z., Guotai, C., Moula, F. E., Azad, A. S. M. S., & Khan, M. S. U. (2019b). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*, 474–507.
Abedin, M. Z., Guotai, C., Uddin, M. M., Shahriare, S., Imran, K., & Hajek, P. (2021). Tax default prediction using feature transformation-based machine learning. *IEEE Access, 9*, 19864–19881.
Abedin, M. Z., Chi, G. T., Hajek, P., & Tong, Z. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00614-4
Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three-stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research, 222*(1), 168–178.
Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 589–609.
Altman, E. I., Haldeman, R., & Narayanan, P. (1977). ZETA analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance, 1*, 29–54.
Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications, 42*(19), 6609–6619.
Bai, X. P., & Zhao, Z. C. (2022). An optimal credit scoring model based on the maximum default identification ability for Chinese small business. *Discrete Dynamics in Nature and Society, 2022*, 1551937.
Bai, C. G., Shi, B. F., Liu, F., & Joseph, S. (2019). Banking credit worthiness: Evaluating the complex relationships. *Omega, 83*, 26–38.
Bank of Dalian. (2014). *Dalian Bank small business credit system*. Dalian Bank.
Chai, N. N., Wu, B., Yang, W. W., & Shi, B. F. (2019). A multicriteria approach for modeling small enterprise credit rating: Evidence from China. *Emerging Markets Finance and Trade, 55*(11), 2523–2543.
Cheng, J., Feng, Y., Tan, J., & Wei, W. (2008). Optimization of injection mold based on fuzzy moldability evaluation. *Journal of Materials Processing Technology, 208*(1), 222–228.

Chern, C. C., Lei, W. U., Huang, K. L., et al. (2021). A decision tree classifier for credit assessment problems in big data environments. *Information Systems and e-Business Management, 19*, 363–386.

Chi, G. T., & Zhang, Z. (2017). Multi criteria credit rating model for small Enterprise using a nonparametric method. *Sustainability, 9*(10), 1–23.

Chi, G. T., Abedin, M. Z., & Moula, F. E. (2017). Modeling credit approval data with neural networks: An experimental investigation and optimization. *Journal of Business Economics and Management, 18*(2), 224–240.

China Banking Regulatory Commission (CBRC). (2015). *Supervision by law, supervision of the people, risk supervision: The China Banking Regulatory Commission implements the reform of the regulatory framework*. Accessed from http://www.cbrc.gov.cn/chinese/home/docView/67163D0D8293499BA499D2A9705C61CD.html

Ciampi, F., & Gordini, N. (2013). Small Enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. *Journal of Small Business Management, 51*(1), 23–45.

Dagong Global Credit Rating Co. (2010). *Credit rating methodology framework of Dagong Global Credit Rating Co., Ltd.*

Demircan, S., & Kahramanli, H. (2016). Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. *Neural Computing & Applications, 29*(8), 1–8.

Fitch Ratings. (2013). *Fitch Ratings global corporate finance 2012 transition and default study*. Credit Market Research, Fitch Ratings.

Florez-Lopez, R. (2007). Modelling of insurers' rating determinants: An application of machine learning techniques and statistical models. *European Journal of Operational Research, 183*(3), 1488–1512.

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: A correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*(13), 5737–5753.

Gu, W., Meheli, B., Zhang, C., & Li, R. W. (2017). A unified framework for credit evaluation for internet finance companies: Multi-criteria analysis through AHP and DEA. *International Journal of Information Technology & Decision Making, 3*(3), 597–624.

Hai, L., Shi, B. F., & Peng, G. (2013). A credit risk evaluation index system establishment of petty loans for farmers based on correlation analysis and significant discriminant. *Journal of Software, 8*(9), 2344–2351.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications, 42*(2), 741–750.

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications, 39*(8), 6774–6781.

İç, Y. T., & Yurdakul, M. (2010). Development of a quick credibility scoring decision support system using fuzzy TOPSIS. *Expert Systems with Applications, 37*(1), 567–574.

Industrial and Commercial Bank of China. (2005). *Notice on printing and distributing the measures for the evaluation of credit Ratings of small business Enterprises of Small Industrial Enterprises of Industrial and Commercial Bank of China*. Industrial and Commercial Bank of China, no. 78.

iResearch. (2021). *Report on Chinese Financing Development of Micro, Medium and Small Enterprises in 2021*. iResearch Inc.

Khalili-Damghani, K., Sadi-Nezhad, S., Lotfi, F. H., & Tavana, M. (2013). A hybrid fuzzy rule-based multi-criteria framework for sustainable project portfolio selection. *Information Sciences, 220*(1), 442–462.

Liang, X. C., Chen, S. F., & Liu-Yan. (2007). The study of small enterprises credit evaluation based on incremental AntClust. In *IEEE International Conference on Grey Systems and Intelligent Services* (pp. 294–298). IEEE Xplore.

Lu, Y., Yang, L., Shi, B., Li, J., & Abedin, M. Z. (2022). A novel framework of credit risk feature selection for SMEs during industry 4.0. *Annals of Operations Research, 2022*, 1–28.

Marcano-Cedeño, A., Marin-De-La-Barcena, A., Jimenez-Trillo, J., Piñuela, J. A., & Andina, D. (2011). Artificial Metaplasticity neural network applied to credit scoring. *International Journal of Neural Systems, 21*(4), 311–317.

Mijid, N., & Bernasek, A. (2013). Gender and the credit rationing of small businesses. *Social Science Journal, 50*(1), 55–65.

Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and K-nearest neighbors. *Expert Systems with Applications, 110*, 381–392.

Promentilla, M. A., Furuichi, T., Ishii, K., & Tanikawa, N. (2008). A fuzzy analytic network process for multi-criteria evaluation of contaminated site remedial countermeasures. *Journal of Environmental Management, 88*(3), 479–495.

Reichert, A., Cho, C. C., & Wagner, G. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business & Economic Statistics, 1*(2), 101–114.

Robillard, M. P., Maalej, W., Walker, R. J., & Zimmermann, T. (2014). An approach on fault detection in diesel engine by using symmetrical polar coordinates and image recognition. *Advances in Mechanical Engineering, 2014*(4), 1–9.

Ruan, J. H., Wang, Y., Chan, F. T. S., Hu, X., Zhao, M., Zhu, F., Shi, B. F., Shi, Y., & Lin, F. (2018). A life-cycle framework of green IoT based agriculture and its finance, operation and management issues. *IEEE Communications Magazine, 2018*. https://doi.org/10.1109/MCOM.2018.1800332

Rui, L., & Mendes, R. V. (2017). Detecting and quantifying ambiguity: A neural network approach. *Soft Computing, 22*(8), 1–9.

Shi, B. F., & Chi, G. T. (2014). A model for recognizing key factors and applications thereof to engineering. *Mathematical Problems in Engineering, 2014*(1), 368–381.

Shi, B. F., Wang, J., Qi, J., & Cheng, Y. (2015). A novel imbalanced data classification approach based on logistic regression and fisher discriminant. *Mathematical Problems in Engineering, 2015*(6), 1–12.

Shi, B. F., Chen, N., & Wang, J. (2016). A credit rating model of microfinance based on fuzzy cluster analysis and fuzzy pattern recognition: Empirical evidence from Chinese 2,157 small private businesses. *Journal of Intelligent & Fuzzy Systems, 31*(6), 3095–3102.

Shi, B. F., Meng, B., Yang, H. F., Wang, J., & Shi, W. L. (2018). A novel approach for reducing attributes and its application to small enterprise financing ability evaluation. *Complexity, 2018*, 1–17. https://doi.org/10.1155/2018/1032643

Sobehart, J., & Keenan, S. (2001). Measuring default accurately. *Risk, 14*(3), 31–33.

Standard & Poor's Ratings Services. (2011). *S&P's study of China's top corporates highlights their significant financial risks*. Standard & Poor's.

State Council of the People's Republic of China (SCPRC). (2016). *Notice of the State Council on Printing and Promoting Inclusive Financial Development Plan (2016–2020)*. Accessed from http://www.gov.cn/zhengce/Content/2016–01/15/content_10602.htm

State Council of the People's Republic of China (SCPRC). (2017). *Notice of the General Office of the State Council on the Establishment of the "Made in China 2025" National Demonstration Zone*. Accessed from http://www.gov.cn/zhengce/content/201711/23/content_5241727.htm

Sun, Y., Chai, N. N., Dong, Y. Z., & Shi, B. F. (2022). Assessing and predicting small industrial enterprises' credit ratings: A fuzzy decision making approach. *International Journal of Forecasting, Forthcoming., 38*, 1158.

Tomczak, J. M., & Zięba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications, 42*(4), 1789–1796.

Wang, Y.T. and Leng, H. Y, 2021. Credit decision of SMEs based on improved TOPSIS and decision tree. In *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 489–492). https://doi.org/10.1109/AEMCSE51986.2021.00106

Wang, J., Ding, D., Liu, O., & Li, M. (2016). A synthetic method for knowledge management performance evaluation based on triangular fuzzy number and group support systems. *Applied Soft Computing, 39*, 11–20.

Wu, Z., Ahmad, J., & Xu, J. (2016). A group decision making framework based on fuzzy VIKOR approach for machine tool selection with linguistic information. *Applied Soft Computing, 42*, 314–324.

Yu, F., Xu, H., Wang, L., & Zhou, X. (2010). An improved automatic FCM clustering algorithm. In *International Workshop on Database Technology and Applications* (pp. 1–4). IEEE. https://doi.org/10.1109/DBTA.2010.5659043

Yurdakul, M., & Iç, Y. T. (2015). Development of a performance measurement model for manufacturing companies using the AHP and TOPSIS approaches. *International Journal of Production Research, 43*(21), 4609–4641.

Zhang, J. H., Liu, H. Y., Zhu, R., & Liu, Y. (2017). Emergency evacuation of hazardous chemical accidents based on diffusion simulation. *Complexity, 2017*, 1–16.

Zhong, Y., Zhang, L., Xing, S., Li, F., & Wan, B. (2014). The big data processing algorithm for water environment monitoring of the three gorges reservoir area. *Abstract and Applied Analysis, 2014*(5), 1–7.

Zhu, P., & Hu, Q. (2013). Rule extraction from support vector machines based on consistent region covering reduction. *Knowledge-Based Systems, 42*(2), 1–8.

# Part III
# Financial Time-Series Forecasting

# An Ensemble LGBM (Light Gradient Boosting Machine) Approach for Crude Oil Price Prediction

**Sad Wadi Sajid, Mahmudul Hasan, Md. Fazle Rabbi, and Mohammad Zoynul Abedin**

**Abstract** Crude oil is considered one of the most important resources in the world today. Most of the fuel used today is refined from crude oil. Fuel also has a great impact on the global economy. The crude oil market is liquid and uncertain. The prediction of the crude oil market price has become a necessity of every second for governments, industries, and individuals. Predicting the price of crude oil can help to achieve a sustainable economy. The goal of this study is to forecast crude market prices as accurately as possible using machine learning and ensemble learning methodology. In this study, we propose the prediction of crude oil using Light Gradient Boosting (LGBM), Random Forest ensemble machine learning algorithm, Lasso Regression, and Decision Tree machine learning algorithm. The BRENT time series crude oil data are used for analysis and form a prediction model that gives less error and more accuracy. We have compared the prediction result of LBGM with Lasso Regression, Random Forest Regression, and Decision Tree regression analysis. A comparison curve is used for introducing the result, turns out LBGM gives the most accurate and efficient prediction result. We have validated our result by evaluating the root mean square error (RMSE), mean absolute percentage error (MAPE), mean squared error (MSE), mean absolute error (MAE), and the results obtained by the proposed model are significantly close and superior.

S. W. Sajid
Department of Electronics and Communication Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

M. Hasan · M. F. Rabbi
Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
e-mail: rabbi@hstu.ac.bd

M. Z. Abedin (✉)
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_9

**Keywords** Crude oil price prediction · Ensemble learning · Machine learning · Time series analysis

## 1 Introduction

Crude oil is essentially the primary resource of major oils and fuels available today. Crude oil is a type of petroleum. It is composed of natural hydrocarbon deposits and other organic materials. Crude oil is found as a liquid substance in underground reservoirs (Ashour et al., 2011). If the crude oil price increases generally the expected rate of economic growth decreases. This essentially lowers the economic growth prospects, in turn, decreases the expected earnings of companies, resulting in a dampening effect on stock prices. Rather than that, volatilities in the price of crude oil have a huge impact on other economic activities too, as crude oil is the largest source in the energy market (Zhao et al., 2017). Oil price prediction is very useful and important for companies, industries, researchers, governments, and individuals. Because crude oil has a great impact on the world economy and stability (Chen & Huang, 2021; Abedin et al., 2021a). Like the FOREX market, the crude oil market is very volatile, so it has been an interesting field for researchers. There are already many methods that have been developed to predict crude oil prices. Many of them use convolutional neural networks (CNN), artificial neural networks (ANN), deep neural networks (DNN) (Abedin et al., 2021a,b; Rahman et al., 2021; Kaur et al., 2013). From EIA we get to know that US oil production has increased from 4.96 million barrels per day to 5.59 million barrels per day in just the last five years. OPEC's recent agreement is causing volatility in the oil price. For this reason, the environment of the oil market is changing and influencing factors are becoming more and more complex and diverse (Lu et al., 2021). Therefore, forecasting the price of crude oil has become more difficult for researchers; they are applying new and more efficient approaches such as stream learning, CNN model, ANN, vector autoregressive model, etc. (Chen & Huang, 2021; Abedin et al., 2021a; Rahman et al., 2021; Kaur et al., 2013). Authors of those study included many factors, different approaches. Among all of them, "Ensemble Machine Learning" has been shown to give the most desired result. Authors of this study intend to get the best possible forecasting result; authors started with machine learning approaches Lasso Regression, Decision tree regression and Bootstrap Aggregation (Bagging) ensemble Random Forest Analysis. Both of them gave a good result, but why not analyze it using a better and more efficient forecasting system for the crude oil market. So, the authors use the stochastic boosting ensemble model named "Light Gradient Boosting Machine (LBGM)," which gives the best possible forecasting result. Although authors have found that "Random Forest Analysis" provides better results than "Lasso Regression" as the crude oil price is a nonlinear time series data. The prediction model the authors have built is promising and it will provide an upcoming fluctuation in the price of crude oil. Different types of error measurement techniques are used to measure the performance of the algorithms are shown in tabular format. Also, the error is represented by a line chart that clearly indicates that the performance of Light GBM is better than others. Later parts of this study have reviewed on related work, methodology, performance measurement, result and discussion, conclusion, and future work that the authors intend to do.

## 2 Literature Review

As already mentioned in recent years, many remarkable works have been done on economic predictions. A study proposed a model based on Bidirectional Long Short-Term Memory (Bi-LSTM) for oil price forecast. This proposed framework has two modules (Vo et al., 2020). Zhongpei Chen approached the crude oil price prediction method with Long Short-Term Memory (LSTM) deep learning. They proposed a creative algorithm named data transfer with prior knowledge. The study has also compared the price forecast performance with three other training models, but LSTM gave the most desired result (Cen & Wang, 2019). A novel algorithm was introduced by the authors of the study to predict the variation in the price of crude oil of the West Texas Intermediate (WTI) which is based on soft computing. This study implemented a simple but effective way to predict the price using a data filtering algorithm (Ghaffari & Zare, 2009). A novel network "Random Wavelet Neural Network" combined with effective random time function is developed by the authors to improve the prediction accuracy of fluctuations in crude oil price. This study predicted both WTI and BRE crude oil prices using a custom-developed model (Huang & Wang, 2018). The prediction selection method was introduced, rather than widely used regressors, resulting in improvements in prediction accuracy close to 10% relative to the benchmark. The authors pointed out that the well-known Welch and Goyal's dataset leads to more consistent and remarkable accuracy gains relative to other alternative approaches (Nonejad, 2021; Welch & Goyal, 2008). Various types of deep learning approaches have been applied to predict the exchange rate during the COVID-19 pandemic, and the authors here worked with a few interesting parameters to prioritize the effect of the pandemic on the economy (Shajalal et al., 2021). LSTM and GRU are widely used recurrent neural networks that are used to predict various phenomena. GA Busari has shown the comparison between Adaboost-LSTM and Adaboost-GRU, and the empirical result of that study shows that Adaboost-GRU performs better than Adaboost-LSTM in predicting the price of crude oil (Busari & Lim, 2021). Predicting a phenomenon has always been a favorite for researchers. There are many approaches to do so, but a combination of traditional and modern artificial intelligence has been shown to provide more accurate and efficient results. The authors of this study have proposed a "Hybrid Model" to predict credit risk (Chi et al., 2019). Complex and volatile financial markets are well suited to gray analysis environments. So, the authors proposed a gray prediction model that significantly improved performance (Norouzi & Fani, 2020). Yanhui Chan proposed a new deep learning-based hybrid crude oil price prediction model, which improved the forecasting accuracy of previous works (Chen et al., 2017). The more accurate oil price can be predicted, the more stable the market will be. Real-time prediction is rare in the case of crude oil price forecasting; Yuan Zhao proposed a new hybrid model that can provide online real-time price prediction (Zhao et al., 2021). As the crude oil market is highly volatile, it is like an imbalance of time series data. A novel ensemble approach was suggested by the authors to predict an imbalance dataset (Abedin et al., 2019). Yifan Yang found that divide-and-conquer strategy gains a better prediction performance. They have come up with a hybrid approach based on K-means + KPCA + KELM based (Yang et al., 2021). Many researchers have worked on predicting the price of crude oil. Autoregressive moving average (ARMA) models and vector autoregressive (VAR) models with diverse data input each time (Kulkarni & Haidar, 2009). If the crude oil price data are strongly nonlinear, then these nonlinear models can produce more accurate results (Bashiri Behmiri & Pires Manso, 2013). On the crude oil market, the uncertainty of the price is a factor, as the value depends on many parameters. The machine learning method based on adaptive Cuckoo search algorithm (AGWOCS) is proposed to predict the volatile market price of crude oil. The effectiveness of the proposed system, daily and weekly Brent oil prices, are modeled as a case study (Wang et al., 2020). Binrong Wu proposed a novel text-based and big-data-driven model, which utilized a convolutional neural network (CNN) to automatically scrap crude oil news updates. This case study collected 4837 and 3883 news headlines to develop a text-based crude oil forecasting system (Wu et al., 2021). Based on this analysis, in this study, we use one of the latest ensemble algorithms called light gradient boosting machine (LGBM) to predict the price of crude oil.

## 3 Research Methodology

The traditional approach of machine learning analysis is used to predict the price of crude oil. Data are preprocessed before being split into training and testing sets. We randomly split the dataset into 80:20 ratio for training and testing data. The analysis model is built by machine learning and ensemble algorithms trained by the training data, and after training the predicted values come out using the testing values as input. A block diagram of our proposed methodology is given in Fig. 1.

### *3.1 Dataset*

The dataset that was used for the analysis is Crude Oil Prices: Brent – Europe data. It is taken from the US Energy Information Administration. It releases as spot prices, and units is Dollar per barrel. Data frequency is daily, but not seasonally adjusted. It is a time series data from May 20, 1987 to September 10, 2021, and the total number of observations is 8954. Figure 2 represents the information about the dataset.

The price of crude oil was stable during the period 1987 to 2000. After this time, the price increases by a rate. In 2008–2009 it was the maximum and then the price goes down. At the time 2011 to 2015 the price was in a stable situation and after the

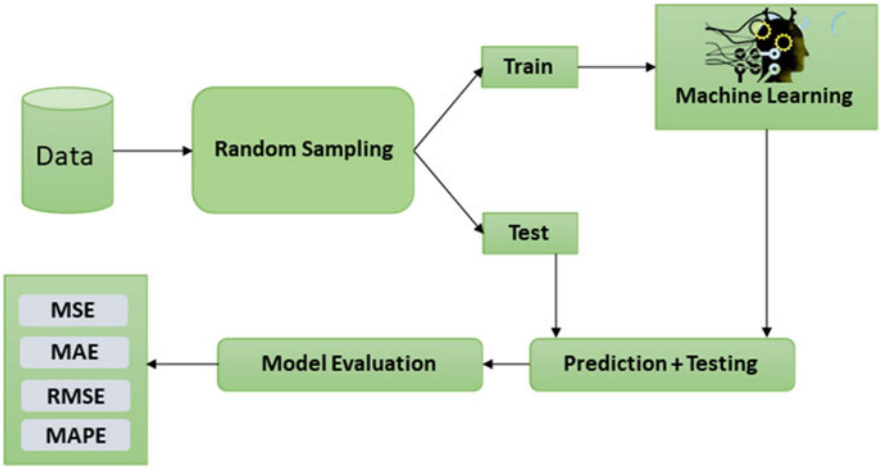



**Fig. 1** Proposed methodology for predicting the price of crude oil

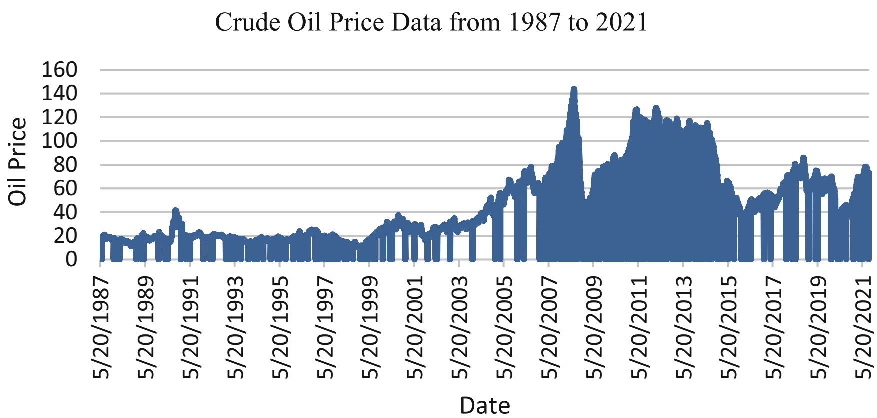



**Fig. 2** Representation of crude oil price data from 1987 to 2021

period it started falling. In 2020 the price of crude oil fell due to the Covid-19 pandemic. The situation is going to be good now and the price is also increasing. The plot clearly indicates that there is a great impact of Covid-19 on the price of crude oil. The above discussion indicates that market of crude oil is not fully stable. Many variables are responsible for varying this price. The prediction of this market is really hard and requires a special and deep analysis. The numerical description of the data set is given in Table 1.

The standard deviation of the crude oil price is 32.01776, and it is not too many scatters. The price of crude oil is increasing day by day and is maintaining a rate. But in the last three months of 2008 the price of crude oil was the highest, because the stock was primarily caused by physical disruptions of supply and the strong demand facing stagnating world production (Ratti & Vespignani, 2013).

**Table 1** Descriptive statistics of Brent Crude Oil data

| Mean | Standard Deviation | Min | Max |
|---|---|---|---|
| 46.75337 | 32.01776 | 9.10000 | 143.95000 |

## *3.2 Description of the Algorithms Used in Analysis*

Two ensemble machine learning algorithms named Light Gradient Boosting and Random Forest Regression as well as Lasso and Decision Tree machine learning algorithm, are used for this analysis. The short description of the algorithms is given below.

**Lasso Regression** The lasso is a type of linear regression and it is a shrinkage method like a ridge. There is a little difference between them. LASSO stands for Least Absolute Shrinkage and Selection Operator. The cost function for the lasso regression can be defined as follows:

$$\sum_{i=1}^{M} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{M} \left( y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j| \text{ for some } t > 0, \sum_{j=0}^{p} \times |w_j| < t, \tag{1}$$

The main difference between the ridge and the Lasso regression cost function equation is that magnitudes are considered in the Lasso regression instead of the square coefficient. This normalization (L1) can result in zero coefficients, i.e., some properties are completely ignored for output evaluation. As a result, Lasso regression not only reduces overfitting, but also helps select features that facilitate the interpretation of models.

**Random Forest** Random Forest is an ensemble classifier that creates a number of separate and non-identical decision trees using randomization (Datta et al., 2021). This algorithm, which is a mixture of tree predictors, is used for both classification and regression. Each decision tree includes a random vector as a parameter, determines the feature of the samples at random, and chooses the training data set at random from either a subset of the data set or the entire data set (Bradter et al., 2013). The error rates are comparable to Ad boost when a random selection of features is employed to divide each node, but they are more resilient in terms of turbulence (Shakoor et al., 2017). Random Forest is a very flexible and simple machine learning technique that, in most cases, gives excellent results even without hyper-parameter adjustment. Based on our need, we employed Random Forest for the regression portion of our technique in this study. Utilizing random forest regression, we were able to get very high accuracy for our dataset. SK-learn offers a useful tool for this that quantifies the significance of a feature by looking at how much error is reduced on all trees in the forest by tree nodes using that feature (Grange & Hand, 1987). Overfitting is a problem with deep decision trees; however, overfitting is rarely a problem with Random Forest. It generates random subsets of the characteristics and uses these selections to form smaller trees that it then merges.

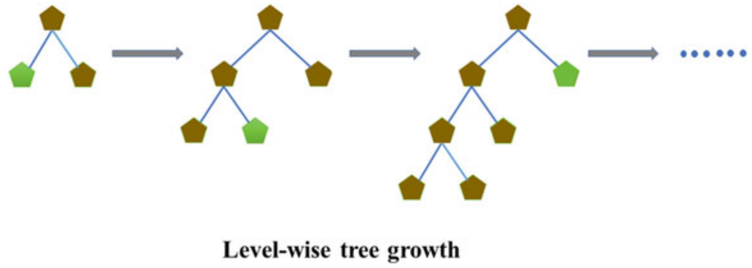



**Fig. 3** Leaf-wise tree growth of Light Gradient Boosting Machine

**Decision Tree Regression** For supervised learning, a decision tree is a common practical technique. It allows both classification and regression estimates to be made. The root node, inner node, and leaf node are the three types of nodes in a decision tree, which is a tree-structured classifier. The root node is the first node, which represents the entire sample and can be divided into other nodes. The core nodes reflect the characteristics of the dataset, whereas the branches represent decision rules. Finally, the root nodes represent the result. A decision tree is executed for a specific data point, True/False questions are answered until they reach the leaf node. The average value of the dependent variable at that particular leaf node is used to produce the final prediction. Through several iterations, the tree is able to predict an appropriate value for the data point. Decision trees are useful because they are simple to grasp, need minimal data cleansing, do not suffer from non-linearity, and have a small number of hyper-parameters to tune.

**Light Gradient Boosting Machine** Light GBM is a tree-based learning algorithm-based gradient boosting framework (Rufo et al., 2021). It is intended to be dispersed and efficient and provides the following advantages: reduced memory utilization, increased training efficiency and speed, and better accuracy. This algorithm uses two novel techniques called Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bunding (EFB), which makes it faster. The Light GBM approach is built on a histogram that organizes continuous feature values into discrete bins to accelerate the training process. Lower memory utilization: Continuous values are replaced with discrete bins, resulting in lower memory usage. It makes this algorithm faster than the others. The tree-based structure of this algorithm is given in Fig. 3.

## *3.3 Performance Measures*

Machine learning and predictive analytics are indeed prone to a variety of errors. We use four mostly used error measurement techniques and compare them using both tabular and graphical forms. Here is a short overview of the errors with the parameters:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - y|, \mathrm{MSE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y - yi)^2, \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - yi)^2}, \mathrm{MAPE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{(y - yi)}{y}, \tag{3}$$

where $n$ is the number of samples, Σ is the summation symbol (which means "add them all together"), $y$ is the actual value, and $yi$ is the predicted value.

*MAE* means Mean absolute error. Absolute errors are defined as absolute values that differ from prediction to actual values. MAE indicates the average error expected from forecasts.

*MSE* means Mean square error. The average square error of the regression line shows the distance to the point set. This is done by dividing the distance between points and regression lines (these distances are the "errors"). Squaring is needed to eliminate any negative signs.

*RMSE* means Root Mean Square Error, which is the standard deviation of the residuals (prediction errors). Residuals are used to measure the distance between data points and the regression lines; RMSE measures the distribution of these residuals. In other words, it reveals how strongly the data is aggregated around the line of best fit.

*MAPE* means Mean absolute percentage error. One of the most widely used KPIs for evaluating predictive performance is MAPE. MAPE is calculated by dividing the total absolute error by the desired quantity (each period is separately). This is calculated on an average percentage error.

## 4 Results and Discussion

In this paper, we use four different models to predict the price of crude oil. After developing the model, we test by the test value and generate the actual vs. predicted curve. The curves of the three methods are given in Figs. 4 and 5.

Figure 4 indicates the actual vs predicted curve of crude oil price prediction. The blue color indicates the actual values and the red color indicates the predicted values of the Brent oil price data. The curve shows that the performance of Lasso is good and after evaluation we get 0.01730 MAE, 0.00046 MSE, 0.02143 RMSE and 0.40613 MAPE error, which are tabulated in Table 2. This curve also indicates that the price of crude oil does not maintain any specific rules. It can fall at any time and increase at any time. The results of the remaining algorithms are shown and discussed below one by one.

Figure 5 indicates the actual vs. predicted curve of crude oil price prediction. The blue color indicates the actual values and the red color indicates the predicted values of the Brent oil price data. The curve shows that the performance of Random Forest

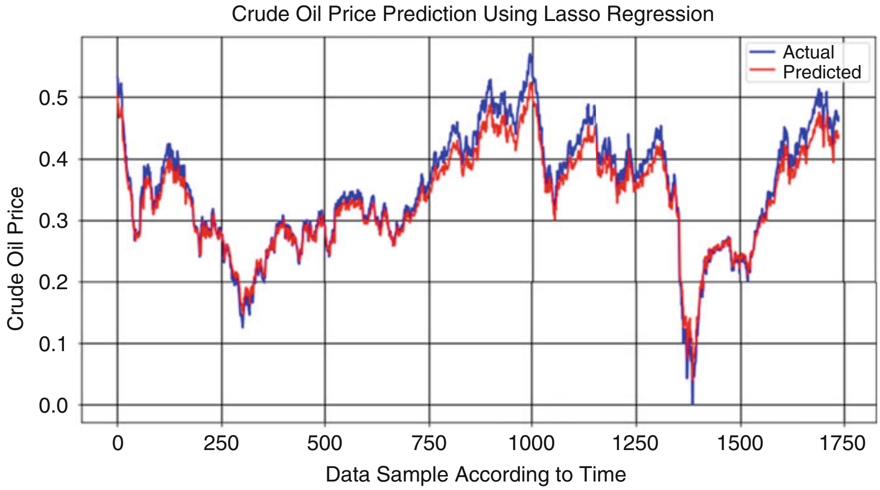



**Fig. 4** Actual vs. predicted using Lasso

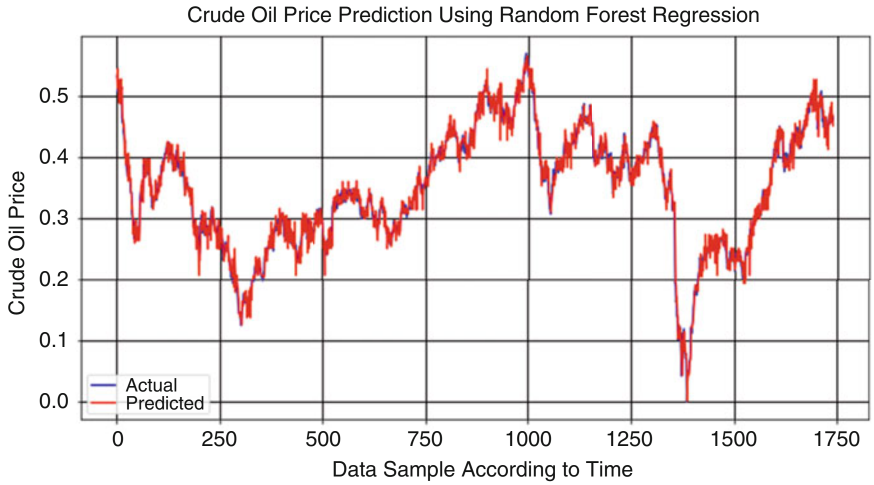



**Fig. 5** Actual vs. predicted using Random Forest

**Table 2** Performance measurement of different algorithms for the prediction of crude oil price

| Method | MAE | MSE | RMSE | MAPE |
|---|---|---|---|---|
| Lasso regression | 0.01730 | 0.00046 | 0.02143 | 0.40613 |
| Random Forest regression | 0.01076 | 0.00020 | 0.01416 | 0.26699 |
| Decision tree | 0.01065 | 0.00019 | 0.01393 | 0.27218 |
| Light gradient boosting | **0.00732** | **0.00009** | **0.00998** | 0.26201 |

Bold values: the most minimum error rate that signifies the best model performance

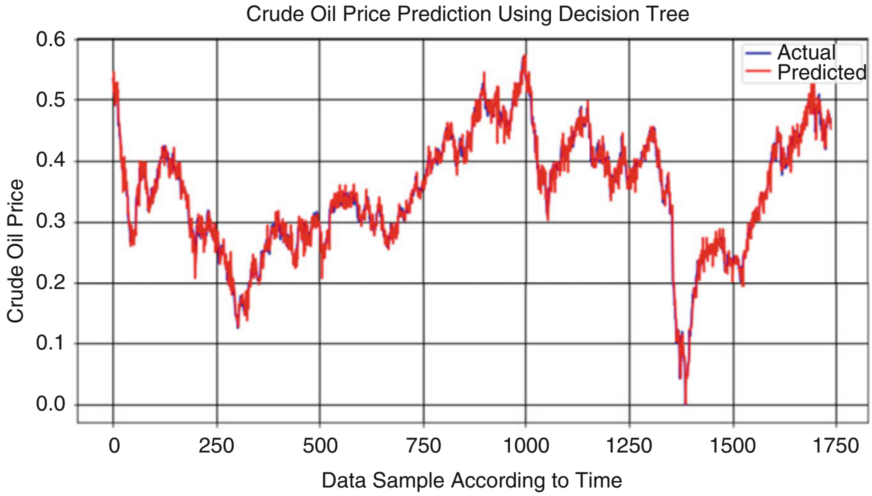



**Fig. 6** Actual vs. predicted using Decision Tree

is good and after evaluation we get 0.01076 MAE, 0.00020 MSE, 0.01416 RMSE and 0.26699 MAPE error, which are presented in Table 2.

Figure 6 indicates the actual vs. predicted curve of crude oil price prediction. The blue color indicates the actual values and the red color indicates the predicted values of the Brent oil price data. The curve shows that the performance of Decision Tree is good and after evaluation we get 0.01065 MAE, 0.00019 MSE, 0.01393 RMSE and 0.27218 MAPE errors, which are presented in Table 2.

Figure 7 indicates the actual vs. predicted curve of crude oil price prediction. The blue color indicates the actual values and the red color indicates the predicted values of the Brent oil price data. The curve shows that the performance of LGBM is good and after evaluation we get 0.00732 MAE, 0.00009 MSE, 0.00998 RMSE, and 0.26201 MAPE error, which are tabulated in Table 2.

Table 2 represents the MAE, MSE, RMSE, and MAPE error values of Lasso Regression, Random Forest Regression, and Light Gradient Boosting. It clearly indicates all kinds of error in Light Gradient Boosting are less than others. It means that the prediction of Light Gradient Boosting is better than the other two algorithms. For clear understanding, we represent the errors in a line chart in Fig. 7.

Figure 8 represents MAE, MSE, RMSE, and MAPE of three models. The yellow color represents the errors of the Light Gradient Boosting algorithm, the gray color represents the Decision Tree, the orange color represents the errors of the Random Forest Regression, and the blue color represents the errors of Lasso Regression. The numeric values 1, 2, 3, and 4 represent MAE, MSE, RMSE, and MAPE consecutively. The figure clearly indicates that the error rate of Light Gradient Boosting is less than others.

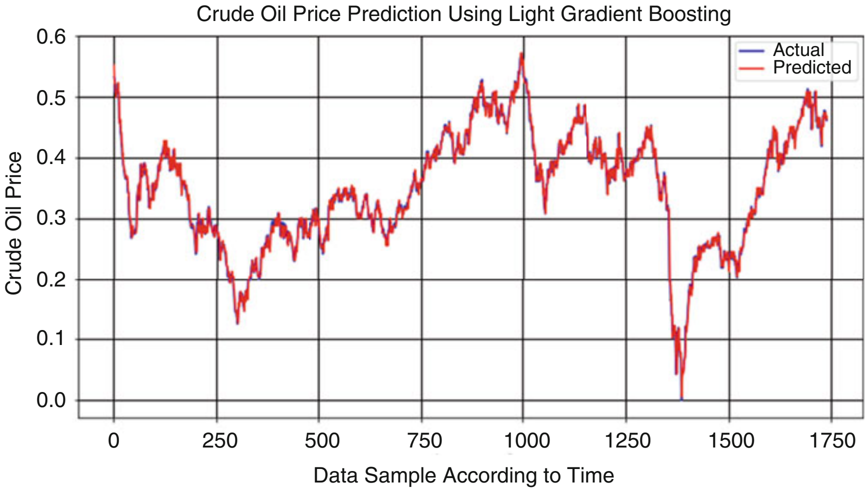



**Fig. 7** Actual vs. predicted using LGBM

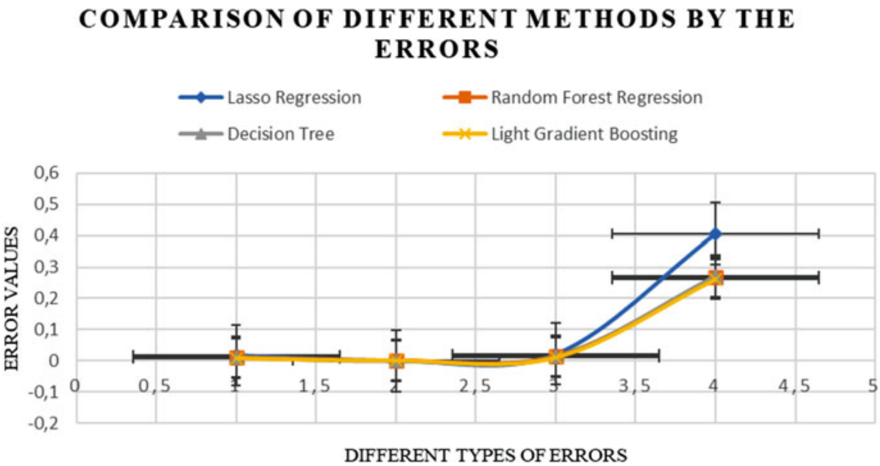



**Fig. 8** Comparison of different methods by errors

## 5 Conclusion and Future Work

This research focuses on the prediction of Brent crude oil price. We apply two machine learning algorithms and two ensemble algorithms for analysis. Overall performance of Light Gradient Boosting Machine algorithms is better than others. All the measurements are shown in both tabular and graphical form. The performance of the other algorithms is also satisfying and error is low. This analysis helps all those related to this field take the challenging decisions that are directly and indirectly depend on the price of crude oil.

In the future, we want to build an API that shows the prediction of crude oil real-time price. The authors want to add more parameters to the input, and to minimize the complexity of the space and time of the model to ensure accurate prediction. The authors also want to prepare an application software that anyone can use to obtain the real-time predictions.

## References


Abedin, M.Z., Guotai, C. and Moula, F.E., 2019, June. Weighted SMOTE-ensemble algorithms: Evidence from Chinese imbalance credit approval instances. In *2019 2nd International Conference on Data Intelligence and Security (ICDIS)* (pp. 208–211). IEEE Xplore.

Abedin, M. Z., Hasan, M. M., Hassan, M. K., & Hajek, P. (2021a). Deep learning-based exchange rate prediction during the COVID-19. *Annals of Operations Research*, 1–52.

Abedin, M. Z., Hassan, M. K., Khan, I., & Julio, I. F. (2021b). Feature transformation for corporate tax default prediction: Application of machine learning approaches. *Asia-Pacific Journal of Operational Research, 39*, 2140017.

Ashour, I., Al-Rawahi, N., Fatemi, A., & Vakili-Nezhaad, G. (2011). Applications of equations of state in the oil and gas industry. *Thermodynamics Kinetics of Dynamic Systems, 1*, 165–178.

Bashiri Behmiri, N., & Pires Manso, J. R. (2013). Crude oil price forecasting techniques: A comprehensive review of literature. Available at SSRN 2275428.

Bradter, U., Kunin, W. E., Altringham, J. D., et al. (2013). Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution, 4*(2), 167–174.

Busari, G. A., & Lim, D. H. (2021). Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Computers & Chemical Engineering, 155*, 107513.

Cen, Z., & Wang, J. (2019). Crude oil price prediction model with long short-term memory deep learning based on prior knowledge data transfer. *Energy, 169*, 160–171.

Chen, Y. C., & Huang, W. C. (2021). Constructing a stock-price forecast CNN model with gold and crude oil indicators. *Applied Soft Computing, 112*, 107760.

Chen, Y., He, K., & Tso, G. K. (2017). Forecasting crude oil prices: A deep learning-based model. *Procedia computer Science, 122*, 300–307.

Chi, G., Uddin, M. S., Abedin, M. Z., & Yuan, K. (2019). Hybrid model for credit risk prediction: An application of neural network approaches. *International Journal on Artificial Intelligence Tools, 28*(05), 1950017.

Datta, R. K., Sajid, S. W., Moon, M. H., & Abedin, M. Z. (2021). Foreign currency exchange rate prediction using bidirectional long short-term memory. In *The big data-driven digital economy: Artificial and computational intelligence* (pp. 213–227). Springer.

Ghaffari, A., & Zare, S. (2009). A novel algorithm for prediction of crude oil price variation based on soft computing. *Energy Economics, 31*(4), 531–536.

Grange, R., & Hand, D. (1987). A review of the effects of atmospheric humidity on the growth of horticultural crops. *Journal of Horticultural Science, 62*(2), 125–134.

Huang, L., & Wang, J. (2018). Global crude oil price prediction and synchronization-based accuracy evaluation using random wavelet neural network. *Energy, 151*, 875–888.

Kaur, J., Nathani, N., & Kaur, M. (2013). Oil price fluctuations and FOREX market: Evidence from OPEC countries. *International Journal of Research in Business and Technology, 3*(3), 215–235.

Kulkarni, S., & Haidar, I. (2009). Forecasting model for crude oil price using artificial neural networks and commodity futures prices. arXiv preprint arXiv: 0906.4838.

Lu, Q., Sun, S., Duan, H., & Wang, S. (2021). Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model. *Energy Informatics, 4*(2), 1–20.

Nonejad, N. (2021). Predicting equity premium by conditioning on macroeconomic variables: A prediction selection strategy using the price of crude oil. *Finance Research Letters, 41*, 101792.

Norouzi, N., & Fani, M. (2020). Black gold falls, black plague arise - An Opec crude oil price forecast using a gray prediction model. *Upstream Oil and Gas Technology, 5*, 100015.

Rahman, M. M., Guotai, C., Das Gupta, A., Hossain, M., & Abedin, M. Z. (2021). Impact of early COVID-19 pandemic on the US and European stock markets and volatility forecasting. *Economic Research-Ekonomska Istraživanja, 2021*, 1–18.

Ratti, R. A., & Vespignani, J. L. (2013). Why are crude oil prices high when global activity is weak? *Economics Letters, 121*(1), 133–136.

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics, 11*(9), 1714.

Shajalal, M., Hajek, P., & Abedin, M. Z. (2021). Product backorder prediction with deep neural network on imbalance data. *International Journal of Production Research*, 1–18.

Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017, July). Agricultural production output prediction using supervised machine learning techniques. In *2017 1st International Conference on Next Generation Computing Applications (Next Comp)* (pp. 182–187).

Vo, A. H., Nguyen, T., & Le, T. (2020). Brent oil price prediction using Bi-LSTM network. *Intelligent Automation and Soft Computing, 26*(6), 1307–1317.

Wang, J., Niu, T., Du, P., & Yang, W. (2020). Ensemble probabilistic prediction approach for modeling uncertainty in crude oil price. *Applied Soft Computing, 95*, 106509.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 21*(4), 1455–1508.

Wu, B., Wang, L., Lv, S. X., & Zeng, Y. R. (2021). Effective crude oil price forecasting using new text-based and big-data-driven model. *Measurement, 168*, 108468.

Yang, Y., Guo, J. E., Sun, S., & Li, Y. (2021). Forecasting crude oil price with a new hybrid approach and multi-source data. *Engineering Applications of Artificial Intelligence, 101*, 104217.

Zhao, Y., Li, J., & Yu, L. (2017). A deep learning ensemble approach for crude oil price forecasting. *Energy Economics, 66*, 9–16.

Zhao, Y., Zhang, W., Gong, X., & Wang, C. (2021). A novel method for online real-time forecasting of crude oil price. *Applied Energy, 303*, 117588.

# Model Development for Predicting the Crude Oil Price: Comparative Evaluation of Ensemble and Machine Learning Methods

**Mahmudul Hasan, Ushna Das, Rony Kumar Datta, and Mohammad Zoynul Abedin**

**Abstract** The crude oil market is unstable, and its price is highly volatile. Due to the Covid-19 pandemic, the price of crude oils goes up and down in a short period of time. Future plans and projects' policies depend directly and indirectly on the future price of crude oil. So, the aim of this study is to predict the price of crude oil by using machine learning and ensemble algorithm, as well as to show the comparison of performance of Ada Boost, Bagging Lasso and Support Vector Regression model. The study uses crude oil price time series data for analysis and to form a model to predict future price. The actual vs. predicted curve is used to show the performance of each algorithm individually. Analysis shows that the ensemble AdaBoost algorithm displays better performance than other algorithms. The result is validated using mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), two accuracy score function variance score, and $R^2$ score. This study will help the stakeholders of the crude oil industry in making decisions and formulating policies based on forecasted crude oil prices.

M. Hasan
Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

U. Das
Department of Electrical and Electronic Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

R. K. Datta
Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
e-mail: rony.datta@hstu.ac.bd

M. Z. Abedin (✉)
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_10

**Keywords** Crude oil · Price prediction · Ensemble learning · Machine learning

## 1 Introduction

Crude oil, also known as liquid petroleum, accumulates in porous rock formations in the Earth's crust and is used as fuels or for the processing of chemical compounds. Crude oil is not only yellowish-black oil; it is a quarry of the golden possibilities that form as a result of the decomposition of organic material within the crust of the Earth. In global warming and intense impacts of environmental issues, roughly discouraging to reduce the usage of crude oil, almost all of us rely on renewable energy in order to save the environment and protect the future generation. But crude oil is the most valuable energy resource in the present world. Crude oil is essential for various chemical industrial products, including plastics, solvents, fertilizers, and pesticides (Dhifqaui et al., 2022). The price of crude oil and the global economy are interrelated and depend on each other. If for any reason the price of crude oil fluctuates, there will be a massive change in the activities of the global economy (Baumeister & Kilian, 2016). The influence factors of crude oil price include supply and demand, finance factor, and technology are directly influencing the change of the interior and exterior environment of the crude oil market. Day by day, the influencing factors become perplexing and diverse. So, accurate crude price forecasting is a really tough process nowadays (Hamilton, 2009; Kilian & Murphy, 2014; Zhang et al., 2015; Wang et al., 2015; Tang et al., 2012). Many researchers have applied various machine learning methods to predict the price of crude oil. In this way, the Support Vector Machine and the Neural Network are generally used (Zhao et al., 2017). In addition to the Multi-Recurrent Network (Orojo et al., 2019), LSTM (Dhifqaui et al., 2022; Hajek & Abedin, 2020), ARIMA (Abdollahi & Ebrahimi, 2020) and the Deep Belief Network (Chen et al., 2019) have been used to predict the price of crude oil. The high prediction accuracy of the crude oil price is beneficial in asset assignment, to mitigate risks for investors and financial policy adjustment for policy makers. It is working as a safeguard for national security and to naturalize the economic growth of the country (Abedin et al., 2019; Guotai et al., 2017). Data processing and a suitable model selection have been splayed the possibility of obtaining a high prediction accuracy (Abedin et al., 2021). In this research, the authors applied AdaBoost, Bagging Lasso, and Support Vector Regression machine learning methods to predict the crude oil price with reliable. And finally, the authors compare all three methods with various error measurement techniques and reach a decision that AdaBoost is better than other algorithms. It is a highly suitable method to predict the price of crude oil.

## 2 Related Literature

Three factors such as supply and demand, finance, and technology are influencing the price of crude oil (Lu et al., 2021). Considering the three factors, we have to clarify the best congruent forecast scheme among others. Time series models, econometric models, qualitative models, and artificial intelligence models are immensely operable for oil price forecasting and modeling. In recent era, anticipating the price of crude oil is a great blessing for many large and tiny industries, individuals and countries (Abedin et al., 2020). To predict the price of oil, many economists and analysts use autoregressive moving average models and vector regression models (Chai et al., 2022). Artificial intelligence methods and traditional econometric models are two highly responsive methods to predict the price of crude oil at present. In the accuracy issue, artificial intelligence methods are more compatible than traditional econometric models (Song et al., 2020). Ensemble probabilistic prediction is given more efficiently than deterministic prediction. The deterministic prediction contains prediction errors that create a discrepancy in financial decision-making in the crude oil market. But the ensemble probabilistic method attempts to overcome all difficulties and mitigate all risks (Satu et al., 2020). There is a hectic relationship between global economics and crude oil prices. For crude oil market indices throughout the world, West Texas Intermediate Crude oil and Brent Crude oil are the most important (Li et al., 2021). The forecasting level is increased by a good data length. Moreover, the length of the data on a daily basis gives a good forecasting level compared to weekly and monthly (Zhao et al., 2021). Due to economic crises, geopolitics, and unforeseen occurrences, the price of crude oil is immensely impacted. The model collocation influences the prediction ability of the model. The validity of crude oil price forecasts would be affected by erroneous model collocation (Yu et al., 2016). Linearity, non-linearity, hysteresis, structural discontinuities, and instability are all aspects of crude oil time series. The decomposing algorithm may be used to create sub-series or components with linearity, non-linearity, and instability (Yu et al., 2016). In the crude oil price, to detect the latent nonlinear features, traditional methods may not be feasible. As a result, a new technique is required to overcome the drawbacks of conventional methods. According to prior studies, artificial intelligence models with robust self-learning capabilities, such as support vector machines (SVMs), artificial neural networks (ANNs), and other intelligence algorithms, have become increasingly popular for crude oil price predictions. Empirical evidence indicates that they outperformed traditional methods. AI models admit its radical limitations such as time consuming, slack convergence, and local minima (Yang et al., 2021). For analyzing tangle and anomaly data, the "decomposition and ensembled" principle is deliberated as an excellent tool (Datta et al., 2021). Data preparation, which includes data cleaning, data transformation, and data reduction, is a critical stage whose main purpose is to generate final data sets that are appropriate and precise for future predictions. In the forecasting literature, there are a variety of strategies for data reduction, including feature selection and future extraction. Feature selection can detect and eliminate as many redundant and unnecessary characteristics as possible. Most crude price forecasting research employs feature selection for data reduction because features maintain their original characteristics, allowing for improved model interpretation. Feature selection only keeps valid variables by defining a threshold, so discarding a lot of important data, whereas feature extraction reduces the original feature space to a simpler one, retaining more data (Abedin et al., 2019).

# 3 Methodology

To predict the price of crude oil, the traditional machine learning analysis technique is applied. Before splitting the data into training and testing, it is preprocessed. For training and testing data, we divided the dataset into an 80:20 ratio at random. Machine learning and ensemble techniques are used to build the analytical model, which is then trained using the training data to provide projected values using the testing values as input. Figure 1 shows a block diagram of our proposed methodology.

## 3.1 Dataset

The dataset that is used for the analysis is the price of the Brent crude oil – Europe data. It is taken from the US Energy Information Administration. It releases as spot prices and its price is in Dollars per barrel. Data frequency is daily, but not seasonally adjusted. It is time series data from May 20, 1987 to September 10, 2021, and the total number of observations is 8954. Figure 2 represents the information about the dataset.

The price of crude oil was stable during the period 1987 to 2000. After this time, the price increased. During 2008–2009 it was maximum, and then the price went down. Between 2011 and 2015, the price was in a stable situation and after this period it started falling. In 2020 the price of crude oil fell due to the Covid-19 pandemic. The situation is going to go well now and the price is also increasing. The plot clearly indicates that there is a great impact of Covid-19 on the price of crude oil. The above discussion indicates that the market for crude oil is not fully stable. Many variables are responsible for varying this price. The prediction of this market is really difficult and needs special and deep analysis. The numerical description of the dataset is given in Table 1.

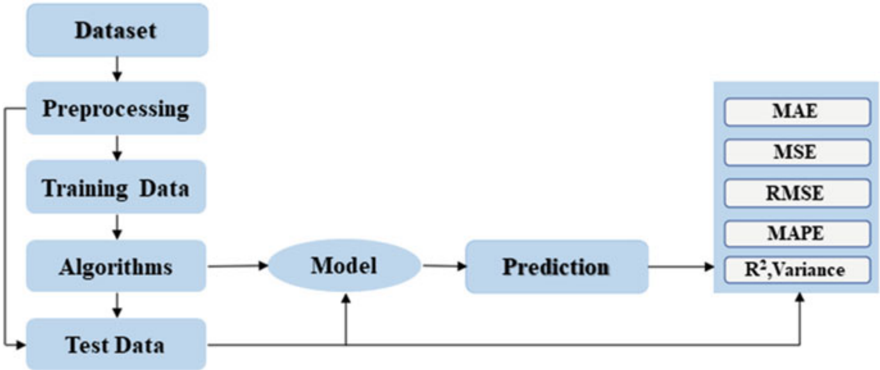



**Fig. 1** Block diagram of the proposed methodology for predicting the price of crude oil

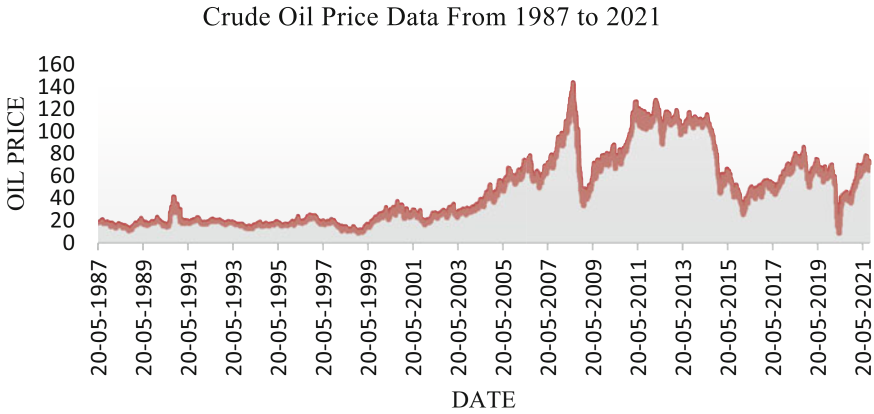



**Fig. 2** Crude oil price data from 1987 to 2021

**Table 1** Statistical description of the Brent crude oil data

| Mean | Standard Deviation | Min | Max |
|---|---|---|---|
| 46.75337 | 32.01776 | 9.10000 | 143.95000 |

## *3.2 Description of the Algorithms*

**AdaBoosting**

Boosting is a kind of ensemble technique that improves prediction accuracy by converting a number of weak learners into strong learners. The Boosting algorithm works on the principle that the first model is developed in the training data set and the second model is constructed to correct the first model errors. This procedure is iterated until the errors are minimized and the data instances are accurately predicted. For each feature, this algorithm generates a weak regressor. Because the weight of effectively calculated samples will be suitably lowered, while the weight of misclassified samples will be appropriately raised, the original classifier does not require a high accuracy if somehow the accuracy is higher than that of random. As a result, the sample distribution is altered. A strong regressor with improved performance may be created by merging the weak samples acquired from each cycle. The features that these powerful classifiers employ are well-classified (Fig. 3).

**Bagging Lasso**

The lasso is a shrinking approach similar to the ridge, but with some key distinctions. The lasso regression cost function may be defined as follows:

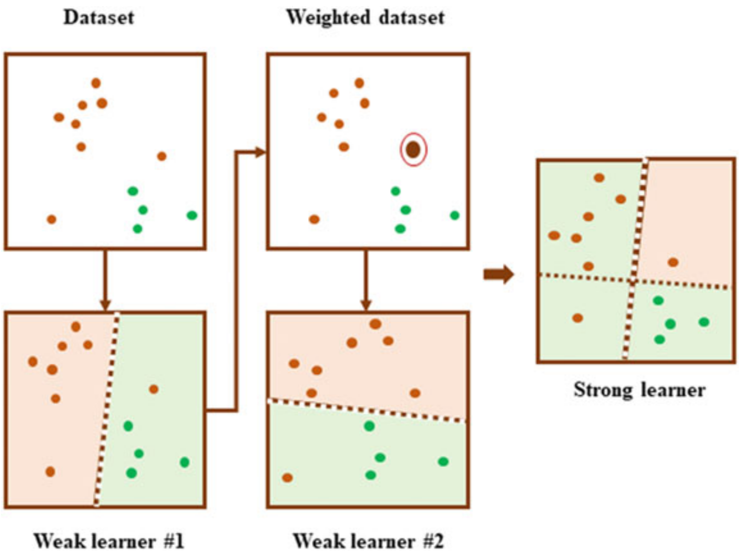



**Fig. 3** Block diagram of the operation of the AdaBoost algorithm

$$\sum_{i=1}^{M}(y_i-\widehat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|\mathrm{w}_j| \text{ for some } t>0, \sum_{j=0}^{p}\times|w_j|<t. \tag{1}$$

The key distinction between the formulations of the cost function of the ridge and lasso regression is that in the lasso regression, instead of calculating the square of the coefficients, the magnitudes are factored into the equation. This method of regularization (L1) might result in a zero coefficient, which means that some characteristics are completely ignored when evaluating the output. As a result, lasso regression not only aids in the reduction of over-fitting but also in the selection of characteristics that make the model easier to understand.

Bagging Lasso is an ensemble algorithm constructed by the bagging ensemble procedure, where Lasso is used as a base algorithm. The data is bagged into different parts and then trained by the Lasso regression. Finally, the final results emerge and give better accuracy than the base Lasso model.

**SVR (Linear, RBF, Polynomial)**

In today's world, the most widely utilized and high-performance algorithm is the support vector machine. This is a supervised machine learning approach that may be used to classify and predict data. However, the authors can employ this learning approach to solve regression problems. The goal of SVM is to build a model (based on the training data). Given only the test data features, the model anticipates the

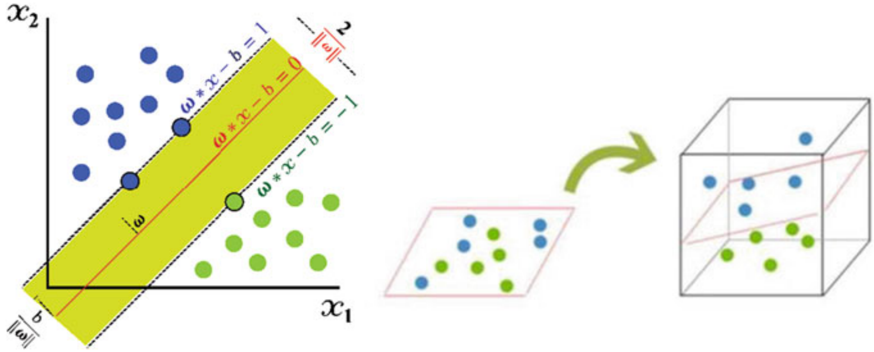



**Fig. 4** Block diagram of Support Vector Machines

output of the target values of the test data. Linear SVM and Kernel SVM are the two forms of SVM that are currently accessible. Linear SVM is an incredibly fast machine learning approach for solving multiclass problems from large datasets (Fig. 4).

SVM implements an exclusive proprietary version of a linear support vector machine design algorithm. This algorithm classifies the data by generating a decision boundary based on the support vector point (Yang et al., 2021). In some instances, the accuracy of SVM is higher than that of other classification algorithms. Kernel SVM is employed for nonlinear data categorization because the data in the real world is not as straightforward as the data in the previous picture. The Kernel SVM is a modified SVM algorithm that may be used to categorize this type of data. SVM's kernel contains a number of arithmetic operations. The functions take data as input and transform them into the format necessary. There are various types of mathematical function. Polynomial, sigmoid, linear, nonlinear, and radial basis functions, for example.

### *3.3 Performance Measures*

*MAE*: It is nothing more than an arithmetic average of the absolute errors. It is the simplest measurement for computing forecast accuracy. It measures the accuracy for a continuous variable as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |yi - y| \,. \tag{2}$$

*MSE*: The Mean Square Error is narrated as an average of the difference between actual and estimated value. In this procedure, all errors are positive. It is highly sensitive to outliers. The small value of this model represents a better model. The *MSE* is defined as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} (y - yi)^2. \tag{3}$$

*RMSE*: The Root Mean Square Error is the root of the mean of the square of all of the errors. It is a standard way to measure the error of a model given as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y - yi)^2}. \tag{4}$$

*MAPE*: The average percentage error for each time period minus genuine values divided by genuine values is how MAPE determines this reliability as a percentage:

$$\mathrm{MAPE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{(y - yi)}{y}. \tag{5}$$

*Variance Score*: The variance is a metric to determine how variable something is. To calculate it, the mean square deviation is used. The dispersion of data collected is measured by variation. The greater the difference in average, the greater the data spread.

$R^2$ *Score*: $R^2$ varies from 0 to 1. It determines how well the data match the regression line. For predictive models, a low $R^2$ value is usually a poor indication. An excellent model may display a little value in some circumstances.

## 4 Results and Discussion

In this paper, two ensemble algorithms are used to predict the crude oil price. The actual vs. predicted values of the algorithms are given in Fig. 5.

Figure 5 indicates the actual vs. predicted curve of crude oil using AdaBoost. The red color indicates the actual values and the blue color indicates the predicted values of the Brent oil price data. The curve shows that the performance of lass is good and after evaluation we get 0.00932 MAE, 0.00015 MSE, 0.01235 RMSE, and 0.24785 MAPE error, which are tabulated in Table 2. From Table 3, we see that the variance score is 0.98 and the $R^2$ score is 0.98. This curve also indicates that the price of crude oil does not have any specific rules. It can fall at any time and increase at any time. The result of the remaining algorithms is shown and discussed below, one by one.

Figures 6, 7, 8, and 9 show the actual vs. predicted curve of crude oil using Bagging Lasso Regression, SVR (Linear Kernel) Regression, SVR (RBF Kernel) Regression, and SVR (Polynomial Kernel) Regression, respectively. The results show that the Bagging Lasso Regression performed best in terms of MAPE error

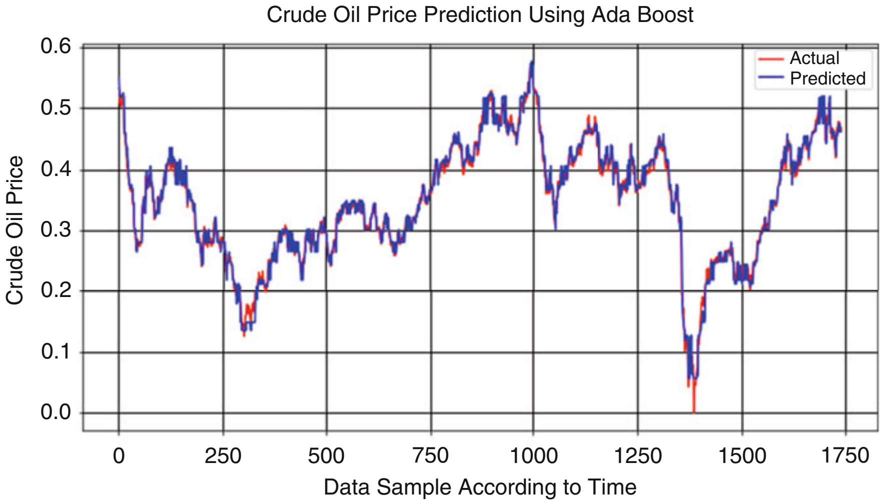



**Fig. 5** Curve of actual vs. predicted crude oil price using AdaBoost

**Table 2** Performance measurement of different algorithms for the prediction of crude oil price

| Method | MAE | MSE | RMSE | MAPE |
|---|---|---|---|---|
| **Ada Boost** | **0.00932** | **0.00015** | **0.01235** | **0.24785** |
| Bagging Lasso Regression | 0.01738 | 0.00046 | 0.02152 | 0.40649 |
| Support Vector Machine (Linear) | 0.01743 | 0.00042 | 0.02059 | 0.45646 |
| Support Vector Machine (RBF) | 0.01663 | 0.00039 | 0.01986 | 0.46593 |
| Support Vector Machine (Polynomial) | 0.02653 | 0.00098 | 0.03131 | 0.59497 |

Note: The best-performing algorithm is in bold

**Table 3** Different types of scores of algorithms for the prediction of crude oil price

| Method | Variance Score | $R^2$ Score |
|---|---|---|
| **AdaBoost** | **0.98** | **0.98** |
| Bagging Lasso Regression | 0.95 | 0.95 |
| Support Vector Machine (Linear) | 0.95 | 0.95 |
| Support Vector Machine (RBF) | 0.96 | 0.96 |
| Support Vector Machine (Polynomial) | 0.90 | 0.90 |

Note: the best-performing algorithm is in bold

(0.40649), while the SVR (Polynomial Kernel) Regression model was superior with respect to MAE (0.01663), RMSE (0.01986), and $R^2$ score (0.96). In fact, the Bagging Lasso Regression overestimates the prices of crude oil, whereas for the SVR models, it is rather the opposite.

Table 2 represents the MAE, MSE, RMSE and MAPE error values of AdaBoost, Bagging Lasso Regression, and different kernel functions of Support Vector Regression. It clearly indicates that all kinds of error in AdaBoost are less than those of the

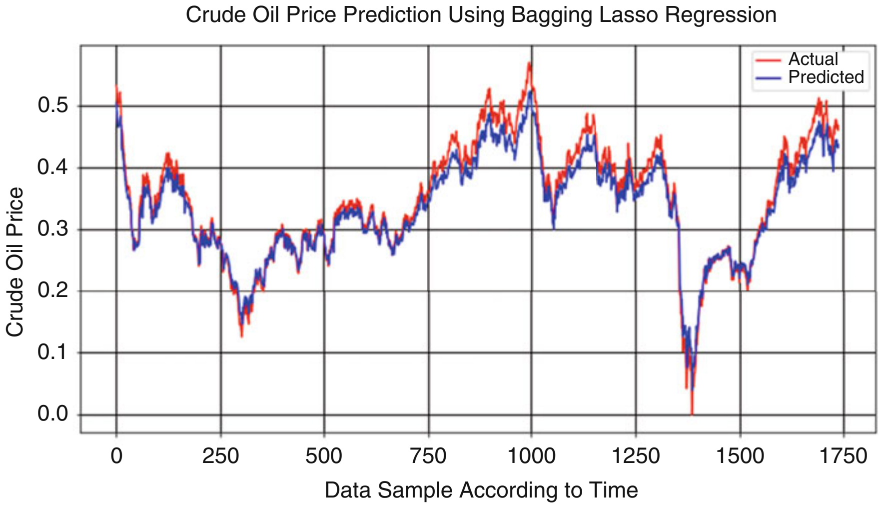



**Fig. 6** Curve of actual vs. predicted crude oil price using Bagging Lasso Regression

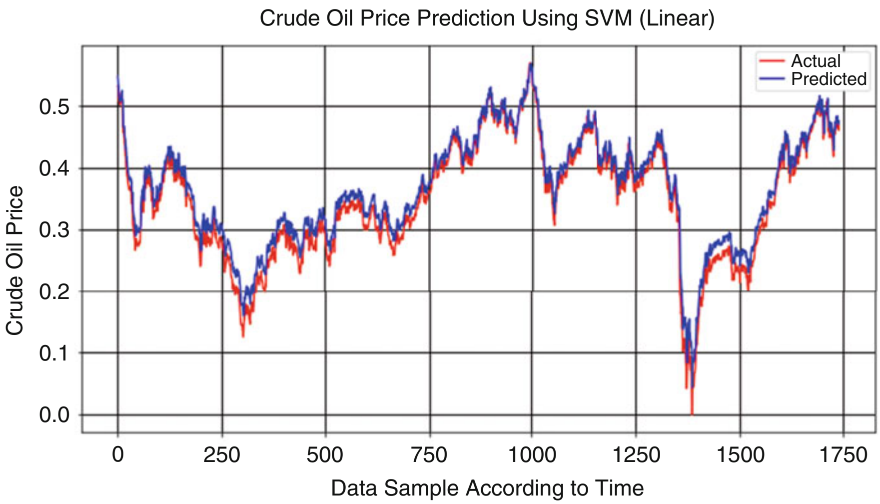



**Fig. 7** Actual vs. predicted curve of crude oil using SVR (Linear Kernel) Regression

other algorithms. It means that the prediction of AdaBoost is better than that of the other algorithms. For clear understanding, we represent the errors in a line chart in Fig. 10.

Figure 10 represents MAE, MSE, RMSE and MAPE of three models. The orange color represents the errors of the AdaBoost model, the yellow color represents the errors of Bagging Lasso, the green color represents the linear SVM, the purple color represents the RBF SVR, and the coffee color represents the Polynomial SVR. The

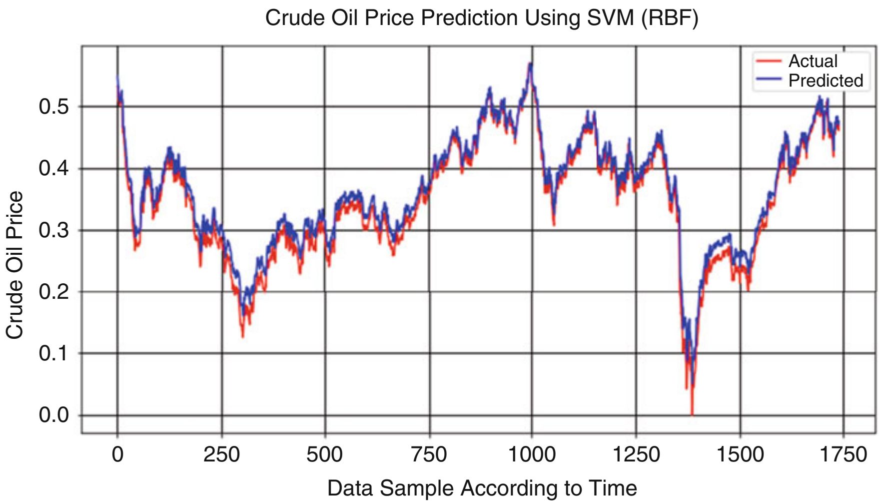



**Fig. 8** Curve of actual vs. predicted crude oil price using SVR (RBF Kernel) Regression

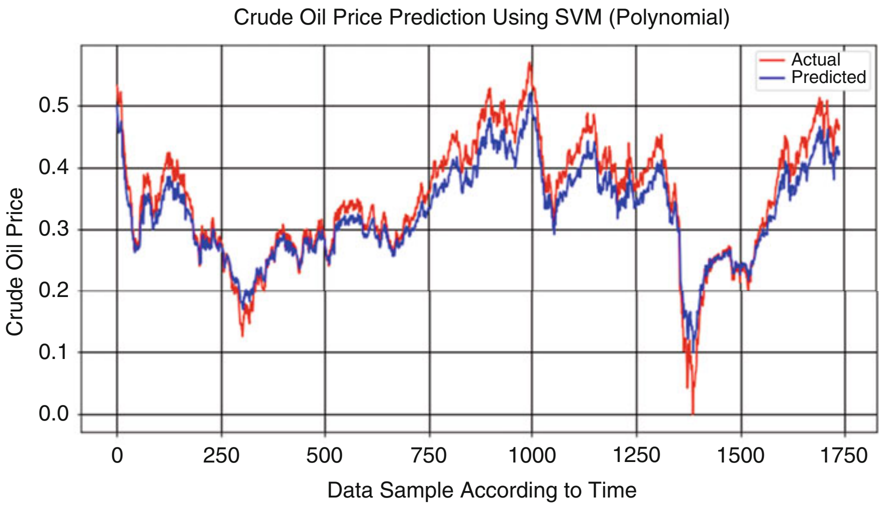



**Fig. 9** Curve of actual vs. predicted crude oil price using SVR (Polynomial Kernel) Regression

numeric values 1, 2, 3, 4 represent MAE, MSE, RMSE, and MAPE consecutively. The figure clearly indicates that the error rate of AdaBoost is less than that of Bagging Lasso.

Table 3 shows the Variance and $R^2$ scores for the compared methods, suggesting that AdaBoost also outperforms the Bagging Lasso Regression and the three Support Vector Regression Model in terms of explained variance, which confirms that the predicted values obtained by AdaBoost fit well the actual oil prices.

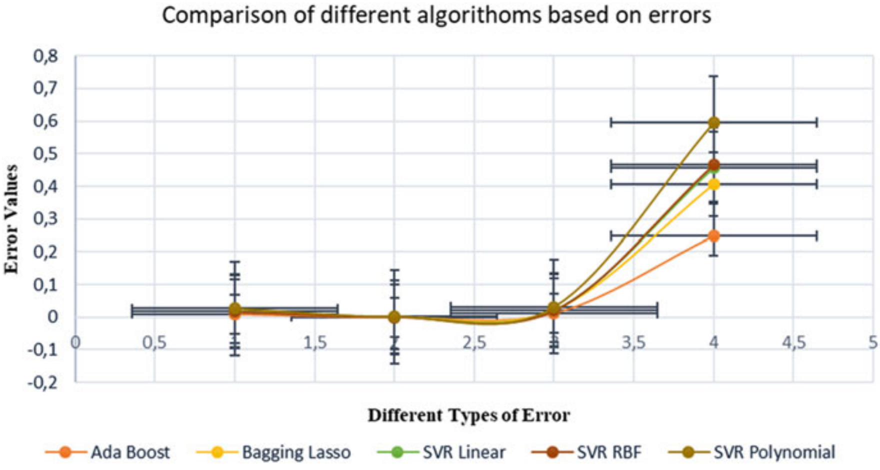



**Fig. 10** Comparison of different methods

## 5 Conclusion and Future Work

The purpose of this study is to forecast the price of Brent crude oil. For analysis, we use the SVR machine learning algorithm and two ensemble techniques Ada Boost and Bagging Lasso Regression. The AdaBoost ensemble machine learning technique outperforms others in terms of overall performance. All the data are presented in tabular and graphical format. The performance of the other algorithms is equally satisfactory, and the error rates are too low. This study helps everyone involved in this industry make difficult decisions that are directly or indirectly influenced by crude oil prices.

In the future, direct and indirect factors can be included, and deep neural network can be used for better prediction. In addition, a website can be developed based on the analysis that can show real-time analysis on the future price of crude oil data.

## References

Abdollahi, H., & Ebrahimi, S. B. (2020). A new hybrid model for forecasting Brent crude oil price. *Energy, 200*, 117520.

Abedin, M. Z., Guotai, C., Moula, F. E., Azad, A. S., & Khan, M. S. U. (2019). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*(1), 474–507.

Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., & Hajek, P. (2020). Tax default prediction using feature transformation-based machine learning. *IEEE Access, 9*, 19864–19881.

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2021). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research, 2021*, 1–52. https://doi.org/10.1007/s10479-021-04420-6

Baumeister, C., & Kilian, L. (2016). Forty years of oil price fluctuations: Why the price of oil may still surprise us. *Journal of Economic Perspectives, 30*(1), 139–160.

Chai, S., Chu, W., Zhang, Z., & Abedin, M. Z. (2022). Dynamic nonlinear connectedness between the green economy, clean energy, and stock price: The impact of the COVID-19 pandemic. *Annals of Operations Research, 2022*, 1.

Chen, J. H., Hao, Y. H., Wang, H., Wang, T., & Zheng, D. W. (2019). Futures price prediction modeling and decision-making based on DBN deep learning. *Intelligent Data Analysis, 23*(S1), 53–65.

Datta, R. K., Sajid, S. W., Moon, M. H., & Abedin, M. Z. (2021). Foreign currency exchange rate prediction using bidirectional long short-term memory. In A. M. A. Musleh Al-Sartawi (Ed.), *The big data-driven digital economy: Artificial and computational intelligence. Studies in computational intelligence* (Vol. 974). Springer. https://doi.org/10.1007/978-3-030-73057-4_17

Dhifqaui, Z., Khalfqoui, R., Abedin, M. Z., & Shi, B. (2022). Quantifying information transfer among clean energy, carbon, oil, and precious metals: A novel transfer entropy-based approach. *Finance Research Letters, 49*, 103138.

Guotai, C., Abedin, M. Z., & Moula, F. E. (2017). Modeling credit approval data with neural networks: An experimental investigation and optimization. *Journal of Business Economics and Management, 18*(2), 224–240.

Hajek, P., & Abedin, M. Z. (2020). A profit function-maximizing inventory backorder prediction system using big data analytics. *IEEE Access, 8*, 58982–58994.

Hamilton, J. D. (2009). Understanding crude oil prices. *The Energy Journal, 30*(2), 179–206.

Kilian, L., & Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics, 29*(3), 454–478.

Li, R., Hu, Y., Heng, J., & Chen, X. (2021). A novel multiscale forecasting model for crude oil price time series. *Technological Forecasting and Social Change, 173*, 121181.

Lu, Q., Sun, S., Duan, H., & Wang, S. (2021). Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model. *Energy Informatics, 4*(2), 1–20.

Orojo, O., Tepper, J., McGinnity, T.M., & Mahmud, M. (2019, December). A multi-recurrent network for crude oil price prediction. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 2940–2945). IEEE.

Satu, M. S., Rahman, S., Khan, M. I., Abedin, M. Z., Mahmud, M., & Kaiser, M. S. (2020, September 19). Towards improved detection of cognitive performance using bidirectional multilayer long short-term memory neural network. In *The proceedings of LNAI: The 13th international conference on brain informatics: Virtual conference*, pp. 297–306.

Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., & Cheng, Z. (2020). Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science and Engineering, 186*, 106682.

Tang, L., Yu, L., Wang, S., Li, J., & Wang, S. (2012). A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. *Applied Energy, 93*, 432–443.

Wang, Y., Wu, C., & Yang, L. (2015). *Forecasting the real prices of crude oil: A dynamic model averaging approach*. Available at SSRN 2590195.

Yang, Y., Guo, J., Sun, S., & Li, Y. (2021). Forecasting crude oil price with a new hybrid approach and multi-source data. *Engineering Applications of Artificial Intelligence, 101*, 104217.

Yu, L., Dai, W., & Tang, L. (2016). A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. *Engineering Applications of Artificial Intelligence, 47*, 110–121.

Zhang, J. L., Zhang, Y. J., & Zhang, L. (2015). A novel hybrid method for crude oil price forecasting. *Energy Economics, 49*, 649–659.

Zhao, Y., Li, J., & Yu, L. (2017). A deep learning ensemble approach for crude oil price forecasting. *Energy Economics, 66*, 9–16.

Zhao, Y., Zhang, W., Gong, X., & Wang, C. (2021). A novel method for online real-time forecasting of crude oil price. *Applied Energy, 303*, 117588.

# Part IV
# Emerging Technologies in Financial Education and Healthcare

# Discovering the Role of M-Learning Among Finance Students: The Future of Online Education

**Armana Hakim Nadi, Syed Far Abid Hossain, Al Mahmud Hasan, Mahbuba Rahman Sofin, Saadman Shabab, Md. Ahmedul Islam Sohan, and Chunyun Yuan**

**Abstract** The chapter aims to explore the role of m-learning among finance students with an additional focus on the future of online higher education. The key reason to conduct the study is to explore the hidden issues of m-learning for the students majoring in finance, especially in the online classroom setting. The study used a qualitative research approach to discover the phenomenon. The authors conducted a thorough literature review of the existing literature and attempted to fulfill the research gap following the qualitative research approach. The result shows that digitalized education provides the opportunity for finance major students to access financial markets using the Internet and gain personal and professional knowledge in a better way rather than traditional learning. The result also discovers a significant positive relationship between m-learning and online educational effectiveness. Only the students of Finance were the participants which may affect the generalizability. The study presents significant implications for education policymakers and practitioners. The study fills the gap in the current literature by discovering the role of m-learning in the online educational setting for finance major students.

A. H. Nadi
Bangladesh University of Professionals, Dhaka, Bangladesh

S. F. A. Hossain (✉)
BRAC Business School, BRAC University, Dhaka, Bangladesh

A. M. Hasan · M. R. Sofin · M. A. I. Sohan
IUBAT University, Dhaka, Bangladesh
e-mail: asohan@iubat.edu

S. Shabab
North South University, Dhaka, Bangladesh

C. Yuan
The School of Economic Management, University of Arts & Sciences, Baoji, China

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_11

**Keywords** m-learning · Finance students · Higher education · Online education · Traditional learning

## 1 Introduction

Owing to technological advancements in the education sector, the significance of Mobile Learning has skyrocketed to a great extent in the modern world. The impact of m-learning on education is a global phenomenon today. Especially in this post-COVID era, m-learning has had an immense impact on scholars and students from around the globe. After the outbreak of the COVID pandemic around the world, education through "traditional learning" that involves direct teaching or an in-person approach in classrooms has become near impossible. In contrast, "e-learning", although proven to be somewhat useful in conducting online sessions, lacks effective retention of the material studied by the students. Smartphones are increasingly serving as the primary source of entertainment, information, communication, and other resources during times of crisis when individuals must stay at home for a longer period of time (Khan et al., 2022). Smartphones are becoming more and more the primary platform for m-learning. M-learning is complementary to both traditional learning and e-learning (Kumar Basak et al., 2018). In recent years, big data science, biomedical computing, and precision medicine have all benefited greatly from the steadily increasing desire to introduce algorithms in machine learning in conjunction with multi-omic data for detecting potential genotype–phenotype links (Khan et al., 2021). M-learning has combined the best of both worlds and introduced mobility in education, thus paving the way for portable education in the process.

Imagine the following scenario: You have some newly recruited employees whom you have been training for a month. At the end of their training, they should be able to perform all the tasks assigned to them when they join the workplace. However, this is highly unlikely. "Research shows that in one hour people will have forgotten an average of 50 percent of the information you presented. Within 24 hours, they have forgotten an average of 70 percent of new information, and within a week they forget to claim an average of 90 percent of it." ("Brain science: The forgetting curve–the dirty secret of corporate training", 2019). So, when your new employees are ready to start working, they would be lucky if they remembered at least half of what they had learned during training, let alone the name of the trainer. However, this situation can be improved through m-learning to some extent. M-learning would help these employees in the hands-on situation with the resources and training they need at that time. Employees can simply pull out their cell phones to get just-in-time training or supporting materials that would allow them to identify and perform the task at hand. Since employees learn the task at the very moment they perform it, they will be able to apply the majority of what they have just learned to the task at hand. Thus, M-learning educates learners by providing bite-sized information, the on-the-go, and just-in-time to perform a task or solve a problem.

Access to financial technology services is relatively well-known to people who are influenced by knowledge of financial services (Hasan et al., 2022). Financial behavior has improved significantly through financial literacy (Wahyudi et al., 2020). However, to achieve financial literacy, it is imperative to implement m-learning in the present world scenario.

In the world of higher education, specifically in the Finance major, scholars are gradually implementing M-learning. It is well known that higher education in Financial Management is heavily focused on mathematical data calculation and analysis. Hence, the traditional education approach in this aspect demands finance students to memorize a myriad of formulae. However, m-learning significantly reduces the inconvenience for students and contributes to efficient learning. The current literature has explored IoT (Internet of Things) in education (Nguyen et al., 2022) with an opportunity to 6G in the near future; class imbalanced prediction (Abedin et al., 2022); deep learning in the contemporary era (Abedin et al., 2021); technological applications (Abedin et al., 2019); the way to achieve education sustainability with advanced technology usage (Hossain et al., 2022) and TPACK development with smartphone usage (Hossain et al., 2021).

In this chapter, we will study the role m-learning plays in students who are majoring in finance in their higher education.

## 2 Literature Review on Mobile Technologies in Teaching

M-learning or Mobile learning has become one of the most talked about topics in the world right now. And due to recent events, m-learning has become a crucial part of daily life for everyone. Starting from teachers to corporate employees, everyone has adopted m-learning as a natural part of their learning routine to stay up-to-date even in this pandemic. Mobile learning, simply put, is any form of education where the central technology used for learning is a handheld or palmtop device. Devices such as smartphones, tablets, and even laptops can be used for m-learning (Guy, 2009). However, there is a common misconception that using personal desktop computers for educational purposes falls under m-learning. This is clearly wrong since m-learning takes advantage of the mobility of handhelds and to provide the user with the opportunity to learn anywhere and at any time. The high success rate of m-learning can mainly be attributed to the high penetration of mobile phones around the world. Even in 2015 studies had shown that mobile phones successfully penetrated 97% of the world (Criollo-C et al., 2018). In recent times, there are almost 8 billion mobile devices in the world. This only goes to show the popularity that mobile devices have gained over the years. In addition to the obvious, m-learning has many advantages. The first obvious advantage, which is evident from the name, is mobility. M-learning has allowed users to learn and teach even when they are away from their homes, schools, offices, and any other conventional learning locations. Another important advantage of M-learning is accessibility. Gone are the days when users would need a full desktop computer to access the Internet. Instead, smartphones give users the ability to access any content on the Internet in the palm of their hand. To add to this, smartphones now have added functionality that desktops cannot provide, like gyro sensors that can be used to view 3D images. Mobile learning helps both students and teachers. Mobile learning enables students to self-regulate their learning and also allows teachers to customize instructions as they see fit (Naciri et al., 2020). Mobile learning is unique because it does not bind students to a certain place or a certain time frame. Students can access the learning materials anytime and anywhere which introduces the idea of training at their convenience (Bazhenov, 2011). Another study shows that university students have a greater ability to learn a foreign language when they do so using a smartphone. This research also shows that although M-learning may not be able to completely replace traditional learning, it can, however, be used to complement to achieve enhanced teaching outcomes (Klimova, 2019). It must also be kept in mind that factors such as information quality and information quality also have an impact on how likely students are to and are satisfied with mobile learning (Almaiah & Alismaiel, 2018). Other factors such as the ability of a teacher to make use of m-learning without sacrificing the quality of education in a traditional or physical classroom can also play a significant role in the adoption of mobile learning by the student (Pedro et al., 2018).

## 3 The Impact of M-Learning on Finance Students

In order to fully realize the potential and the impact of m-learning among Finance students, we must first understand how a Finance student can utilize m-learning. In a general sense, there are three ways mobile learning can be used: educational video content on video portals, mobile apps providing bite-sized lessons on topics, and finally, group study through learning groups on social media. It goes without saying that there is a smartphone in almost every pocket in the world at present. If you own a smartphone, it is safe to say that you are familiar with video portals such as YouTube, Vimeo, Daily Motion, etc. These video portals offer numerous tutorial videos on thousands of topics. A finance student can also easily find tutorials on different topics such as financial ratio calculation, wealth management, corporate finance, investment banking, and many more. It is only a matter of searching for a specific topic and watching a video.

Mobile apps have always been and will continue to be an integral part of smartphones. The variety of apps is endless, to say the least. Engagement with educational apps improves the students' competencies (Camilleri & Camilleri, 2019). Many educational apps prove to be useful to Finance students in their higher education. For example, Android apps in Google Play Store such as "Finance Formulas" and "Financial Ratio Calculator" help students learn and implement a myriad of formulas required in finance education. Social media also play a crucial role in m-learning for finance students. Social media are argued to have the potential to bridge formal and informal learning via a digital culture of participation (Greenhow & Lewin, 2019). Social media can be used for educational purposes in several ways such as enhancing communication and interaction between students and between teachers and students, as well as promoting student engagement as it allows intimidated, shy, or bored students to share thoughts and express his or her opinion more comfortably (Faizi et al., 2013). Furthermore, educational groups and pages on social media platforms such as Facebook facilitate finance students to discuss and perform group studies on various topics of interest. Students with Finance major in higher education can also stay up-to-date on innovations in their field through social media. Last but not least, social media provides networking opportunities to Finance students with successful individuals in their career paths.

## 4 Available Mobile Applications for Online Platforms

Mobile devices and applications to support teaching and learning (m-learning) have received attention in education. In many nations and regions, the spread of Covid-19 has resulted in a rapid shift from traditional to online education platforms. The use of technology in education significantly impacts learning, with universities serving as the primary providers of online education (Aljaaidi et al., 2020). There are different operating platforms, such as Android, iOS and Windows Mobile, that build mobile apps (Hamilton, 2019). Mobile applications make educational information more accessible, and each app has its own set of characteristics that allow it to provide its own set of services. The mobile application also offers online educational services through e-Books, e-Library, informative videos, and games (Jaber et al., 2021). The use of virtual reality (AR) in education has several advantages, including improved engagement and interaction, and can mitigate the negative consequences of face-to-face education disruption (Criollo-C et al., 2021). Learners can access the material anywhere and anytime with learning approaches, with just the touch of a button on the mobile application (Baharum et al., 2020). Therefore, the mobile application is of great benefit to the learner. Mobile-Based Assessment has been increasingly popular in higher education worldwide in recent years; even every learning material is available through a mobile application on the mobile device (Singh et al., 2021). However, the application of technology improves the ability of instructors to reduce digital gaps, improve digital creativity, raise awareness, improve critical thinking, and build reliability on the online platform (Dorouka et al., 2020). According to a study, teachers used Live Video Streaming on numerous social media platforms (such as live social media or linked live) to deliver online instruction to increase student engagement (Chen et al., 2021). Furthermore, mobile learning technologies provide web-based teaching and learning platforms for teachers and learners around the world (Akour et al., 2021). M-learning technology assists teachers in saving time by allowing them to check assignments completed by students, solve numerical methods from the section of calculus for higher mathematics, and use a QR code application to determine whether it is correct or incorrect (Zhylenko et al., 2020). Based on research, students' learning activities and motivation improve after adopting an English game-based Mobile Application (EBMA) in learning (Sofiana & Mubarok, 2020). The revolution technology provides numerous applications available for online learning. Currently, renowned Google Drive applications (docs, spreadsheets, presentations, forms) are gaining popularity and may be utilized efficiently in online education to facilitate communication between academic professionals and students. The learning process, Google Keep, Microsoft Forms, and mural.co designed to construct group work (Llerena-Izquierdo et al., 2020). Day by day, many free online application resources are being updated and new features are also being added for online education. Even the availability of online learning platforms helps students gain different skills, learning activities, and building interest in learning through application. Especially during the COVID-19 epidemic, mobile learning helped students fill in the gaps in their studies (Biswas et al., 2020). There is a great deal of interest in the use of mobile devices and technologies for learning purposes for learners and the need to integrate them more deeply into teacher education in all technological advancements (Connolly et al., 2021).

## 5 Online Platforms for University Students

Although institutions are introducing new areas of study to use the online learning platform, it provides university students with more options to learn. Previously, e-learning, distant education, and correspondence courses were commonly accepted as non-formal education components. However, if current trends continue, it appears that they will gradually supplant the traditional schooling system (Mishra et al., 2020). With more and more university students wanting to study online, online education has become a vital component of modern higher education (Australian Government, 2011). Ted-Ed, Coursera, Google Classroom, Bakpax, Pronto, and Skillshare are examples of some of the most popular online networking sites that will alter the direction and route of the entire educational system in post-COVID-19 scenarios around the world (Mishra et al., 2020). Because online learning will soon become the norm, the government, telecommunication companies, and universities should fund the establishment of technological infrastructure throughout the country (Chung et al., 2020). Furthermore, if students' experiences meet their expectations, they are more likely to feel at ease and continue their studies, and likewise. If students miss classes, want to avoid being absent while filling knowledge gaps, they can attend online training sessions and pass the relevant online tests. The system automatically reports test results to teachers, and when the student is successful, the session evaluation is approved and the student is successfully assessed. To prepare students for the fluctuations of the employment market caused by machine learning and automation, higher education must change and grow quickly and continuously. The communicative online platform system may be linked to a university's student information system, allowing it to modify outreach based on students' actual progress on each required transformation activity. The design of the electronic learning platform, on the other hand, boosts the intellectual and creative qualities of higher education students to help them grow in their careers (Chansanam et al., 2021).

## 6 The Effect of Implementing M-Learning in Education

The revolution in teaching methods expands the possibilities for online education and enhances learners' opportunities through implementing m-learning in education. Individual acceptance of m-learning is crucial for developing countries to extend m-learning successfully (Pratama, 2020). In addition, creating ideal circumstances for women, middle school students, and children in rural areas to use m-learning is critical to education. Implementing online resources is the essential factor for learning (Herrador-Alcaide et al., 2020). Although the acceptance of m-learning in education is effective, proper implementation is a more crucial aspect of learning. The implementation of m-learning in education creates new approaches and educational environments based on the flexible interaction between distance users connect, anonymously or perfectly profiled, and between student-based communities, allowing distance communication between students and teachers; and also between students and machines (Fombona et al., 2020). According to the research, the analysis found that the effect of mobile learning on student learning performance did not vary depending on their educational level or implementation period; however, it did change depending on the course/subject (Talan, 2020). The widespread use of mobile devices in education, as well as the popularity of transferable courses, has resulted in many benefits in terms of the learning process and outcomes, but it has also resulted in several issues. When looking at these issues in general, they may be classified into the following categories: technology-related hardware and software issues, internet and infrastructure issues, mobile device screen, keyboard, and battery issues (Kacetl & Klímová, 2019). Regarding the education study, the authors explore that informal learning contexts are most frequent in m-learning education, followed by formal contexts and both (Aaron & Lipton, 2017). The potential for effective teaching and learning is growing due to the implementation of m-learning in education (Abidin & Tho, 2018). To maximize the effectiveness of M-learning, attention should be paid to designing suitable courses to save time and improve learning efficiency, increase student mobility, and offer the flexibility of the course system to learners through a variety of channels (Trinh et al., 2021). Simultaneously, thousands of apps are available today that are challenging and problematic for both teachers and educators (Papadakis & Kalogiannakis, 2017). Although m-learning has been implemented at a very early stage in other countries around the world, m-learning still solves the problems. Moreover, m-learning implementation brings advantages to education, but it is also vital to expose learners to the convenience of courses through various methods. One of the recent studies shows that the implementation of m-learning positively enhances students' enthusiasm for learning kinematics as well as their self-confidence (Laurens Arredondo & Valdés Riquelme, 2021). While the revolutionary adaptation of m-learning can be noticed in education or other sectors, its implementation takes time in some nations. Based on research, students' attitudes toward utilizing m-learning and their behavioral intents positively influence their long-term viability in higher education (Al-Rahmi et al., 2021).

## 7 A Projection of the Available Digital Online Contents in the Future

In the modern era, the contributions of digital online content to financial education are increasing. At present, there are comprehensive digital libraries that allow students to dive even further into vital financial topics. Websites such as teachbanzai.com, oecd.org, and everfi.com provide finance students with digital delivery of courses that include important topics in financial management. This digital education motivates finance students to engage themselves in an in-depth discussion through a combination of face-to-face interaction along with online learning. In the future, there is a huge scope for development in this sector by generating up-to-date and innovative ideas in the field of financial management. Provisions can be made for on-the-go lessons on financial software used by organizations around the world. Developments can be made by sharing lessons not only through online media but also through interactive sessions that will allow users to acquire first-hand experience of the use of the financial software.

Furthermore, with the rapid growth of technological advancements in the modern world, it can be said without a doubt that Finance education will be heavily impacted by innovative technologies in the future. New and improved technologies can positively contribute to both direct education and m-learning. For example, Virtual Reality (VR) is believed to play a crucial role in the transformation of learning and teaching in higher education. New developments and complete immersion in the virtual environment will undoubtedly increase students' attention (Slavova & Mu, 2018). Mark Zuckerberg announced that Facebook would change its name to Meta, reflecting the new focus on creating a metaverse: a vast and integrated online world that would cover the entire digital society and economy (Oremus, 2021). If the possibility of this situation or in other words, the virtual universe becomes a reality, the implementations can only be imagined as limitless. Students around the world can gather at the same place to attend a virtual classroom session, as well as utilize virtual educational material at the same time. Virtual libraries can even be created where students can study educational material.

Moreover, mobile apps for financial education are in abundance at the moment. Introducing more bite-sized lessons as well as downloadable material that can be accessed offline will surely increase the interest of finance students in M-learning. However, most of these apps only provide learning material to study. But the number of apps that provide interactive solutions for students to practice on is near zero. Therefore, there is scope to make these apps more interactive and enriched with updated information. Interactive apps will allow students to learn about the stock market and challenging apps that encourage them to implement managerial decision skills.

Last but not least, it is true that there is a huge amount of video content on YouTube and other video portals that provides tutorials on different topics of financial management. Unfortunately, very little video content shows the use of financial management software used by organizations. In the future, more video creators can contribute to this case.

## 8 The Development in Education by Virtue of M-Learning

Mobile learning or m-learning has become ever so popular in recent years. The spread of mobile devices plays a vital role in this popularity. Since 95% of the human population lives in an area covered by mobile networks and most adults own more than one mobile device, it is easy to understand the role and importance of m-learning in the world today (Crompton & Burke, 2018).

One of the not-so-obvious implementations of m-learning is in education, and this has become evident in recent times. Before the pandemic, the general population mostly thought education to be in-person learning. However, the ability of humans to adapt to any situation has proved this idea to be incorrect and the biggest contributor which helped prove this concept wrong is m-learning. Not only has m-learning enabled students to learn from the comfort of their homes or even when they are on the move, it has also had positive impacts on the students as well. Studies have found that students perceive collaborative learning positively while learning through mobile technology (Heflin et al., 2017). Other than this, there are many other implications that mobile technology has on students. The use of mobile technology has been associated with higher academic performance for students. On top of this, using mobile technologies for learning can also bring psychological comfort to students who use their mobile devices all the time. Mobile technology even has social implications for students, such as integrating education into their lives as a natural process and not as a training process (Shyshkanova et al., 2017). In general, mobile technologies increase peer-to-peer engagement and also increase participation in learning activities (Fabian et al., 2015). M-learning has helped develop not only the way students perceive education, but also how teachers teach. One research suggests that mobile learning has a high level of success in project-oriented education (Hermann & Gruhn, 2018). M-learning has changed the perspectives of students and teachers alike, since each new topic presents a new opportunity to learn from a new angle. One such example of this is the use of mobile technology and augmented reality to learn Descriptive Geometry (Criollo-C et al., 2018).

## 9 The Affordability and Availability for Pursuing Studies as a Finance Student

In the modern age, any student can pursue their studies on the vast global online education platform that makes education more available and affordable. Finance students and instructors can access educational materials using digital technology anytime, anywhere. Students who use information technology no longer have difficulty obtaining learning resources, which are now widely available on the Internet (Hendra Divayana & Sanjaya, 2017). Finance students and educators also benefit from online learning platforms because they pursue studies with simple and quick access to high-quality educational materials; previously, it was only available in libraries. Since the online application is rapidly developing, students can access finance courses, and even finance students may use affordable mobile devices for learning. And, according to the research, accessing learning materials from a mobile device is essential for 64% of learners.

Furthermore, 89% of smartphone users download apps, and 50% of students use apps for educational courses, including finance courses (Klimova & Polakova, 2020). In foreign language classes, some students may use their mobile devices to look up terms in translations, either installed or web-based dictionaries. As mobile devices are effective educational platforms, students can quickly access mobile devices that provide adequate support for standard Internet technologies. Finance students can use available and affordable websites to acquire knowledge and calculate necessary transactions through the Internet. Due to the new corona virus disease, students generally face the problem of textbook affordability, but online platforms offer a huge opportunity to access e-books. Some Open Education Resource (OER) sites specialize in a specific source type, such as textbooks; even the Open Stax and the Open Textbook Library are two notable textbook available sites (Murphy & Shelley, 2020). However, finance students can bring books from the mentioned sites. The authors suggest that a lack of understanding may hamper the development of m-learning in Higher Learning Institutions, accessibility to technology tools, and affordability (Kamaghe et al., 2020). Online education faces various obstacles, including technological availability and affordability, even when well planned, including obstacles such as learning differences, as well as the instructors' and students' technological skills. The growing popularity of mobile applications requires the banking industry to have a broader view of the market and collaborate with the FinTech sector (Waliszewski & Warchlewska, 2021).

## 10 Conclusion

The practical experience is challenging for finance major students as it includes financial affairs. This chapter ensures the necessity of financial classes being conducted online with effective teaching materials. Numerous personal finance software and apps are available online, making finance students more efficient at managing money and meeting long-term financial goals. As a result, a finance student must know about mobile applications available in the market through the available online courses to prepare for the job market. In addition, finance students can improve their financial management at home using the right tools. Although it is affordable to broadcast lectures on a website for many students, online courses with meaningful interaction among students and instructors are not cost effective

(Baum & Hai, 2020). Digital education provides an opportunity for finance students to access financial markets using the Internet and gain personal and professional knowledge. Also, online teaching and learning have been internationalized. For instance, a well-reputed finance teacher from the USA may conduct a class online with the students of another university situated in Asia. As a result, the chapter ensures that m-learning is significant for finance major students.

**Acknowledgments** The Research on Developing Cultural Industry Chain in Guanzhong-Tianshui Economic Zone (11JK0070), project of the Education Department of Shannxi Provincial Government. Studies on Gradient Development of Cultural Industry in Shannxi Province (11E067), project supported by Social Science Foundation of Shannxi Province, China, The Research on Overall Design of Cultivating Talents throughout Higher Education (SGH10107), Education Sciences Planning Project of Shannxi Province, China.

## References

Aaron, L. S., & Lipton, T. (2017). Digital distraction: Shedding light on the 21st-century college classroom. *Journal of Educational Technology Systems, 46*(3), 1–16. https://doi.org/10.1177/0047239517736876

Abedin, M. Z., Guotai, C., Moula, F.-E., Azad, A. S. M. S., & Khan, M. S. U. (2019). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*(1), 474–507. https://doi.org/10.1002/ijfe.1675

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2021). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research, 2021*, 1. https://doi.org/10.1007/s10479-021-04420-6

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems.* https://doi.org/10.1007/s40747-021-00614-4

Abidin, N. Z., & Tho, S. (2018). The development of an innovative resonance experiment using smartphones with free mobile software applications for tertiary education. *International Journal of Education and Development using ICT, 14*(1), 164–174. Retrieved from https://www.learntechlib.org/p/183553/

Akour, I., Alshurideh, M., Al Kurdi, B., Al Ali, A., & Salloum, S. (2021). Using machine learning algorithms to predict people's intention to use mobile learning platforms during the COVID-19 pandemic: Machine learning approach. *JMIR Medical Education, 7*(1), e24032. https://doi.org/10.2196/24032

Aljaaidi, K. S., Bagais, O. A., & Sharma, R. B. (2020). Factors influencing usage of university mobile application among university students. *The Journal of Asian Finance, Economics, and Business, 7*(10), 1129–1136. https://doi.org/10.13106/jafeb.2020.vol7.no10.1129

Almaiah, M. A., & Alismaiel, O. A. (2018). Examination of factors influencing the use of mobile learning system: An empirical study. *Education and Information Technologies, 24*(1), 885–909. https://doi.org/10.1007/s10639-018-9810-7

Al-Rahmi, A. M., Al-Rahmi, W. M., Alturki, U., Aldraiweesh, A., Almutairy, S., & Al-Adwan, A. S. (2021). Exploring the factors affecting mobile learning for sustainability in higher education. *Sustainability, 13*(14), 7893. https://doi.org/10.3390/su13147893

Australian Government. (2011). *Higher education standards framework*. Department of Industry, Innovation, Science, Research and Tertiary Education. Retrieved from https://www.legislation.gov.au/Details/F2012l00003/Html/Text#_Toc311791709

Baharum, A., Wan, L. Y., Yahya, F., Nazlah, N. H., Nor, N. A., Ismail, I., & Noor, N. A. (2020). Mobile learning application: Flipped classroom. *Indonesian Journal of Electrical Engineering and Computer Science, 17*(2), 1084–1090. https://doi.org/10.11591/ijeecs.v17.i2.pp1084-1090

Baum, T., & Hai, N. T. T. (2020). Hospitality, tourism, human rights and the impact of COVID-19. *International Journal of Contemporary Hospitality Management, 32*(7), 2397–2407.

Bazhenov, R. I. (2011). *Intelligent information technologies* (p. 176). PSU named after Sholem Aleichem.

Biswas, B., Roy, S. K., & Roy, F. (2020). Students perception of mobile learning during COVID-19 in Bangladesh: University student perspective. *Aquademia, 4*(2), ep20023. https://doi.org/10.29333/aquademia/8443

Brain science: The forgetting curve–the dirty secret of corporate training. (2019). Retrieved from https://learningsolutionsmag.com/articles/1379/brain-science-theforgetting-curvethe-dirty-secret-of-corporate-training

Camilleri, A. C., & Camilleri, M. A. (2019). Mobile learning via educational apps. In *Proceedings of the 2019 5th International Conference on Education and Training Technologies-ICETT 2019*. https://doi.org/10.1145/3337682.3337687

Chansanam, W., Tuamsuk, K., Poonpon, K., & Ngootip, T. (2021). Development of online learning platform for Thai University students. *International Journal of Information and Education Technology, 11*(8), 348–355. https://doi.org/10.18178/ijiet.2021.11.8.1534

Chen, X., Chen, S., Wang, X., & Huang, Y. (2021). "I was afraid, but now I enjoy being a streamer!" Understanding the challenges and prospects of using live streaming for online education. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1–32. https://doi.org/10.1145/3432936

Chung, E., Subramaniam, G., & Christ Dass, L. (2020). Online learning readiness among university students in Malaysia amidst COVID-19. *Asian Journal of University Education, 16*(2), 45. https://doi.org/10.24191/ajue.v16i2.10294

Connolly, C., Hijón-Neira, R., & Grádaigh, S. Ó. (2021). Mobile learning to support computational thinking in initial teacher education. *International Journal of Mobile and Blended Learning, 13*(1), 49–62. https://doi.org/10.4018/ijmbl.2021010104

Criollo-C, S., Lujan-Mora, S., & Jaramillo-Alcazar, A. (2018). Advantages and disadvantages of M-learning in current education. In *2018 IEEE World Engineering Education Conference (EDUNINE)*. https://doi.org/10.1109/edunine.2018.8450979

Criollo-C, S., Abad-Vásquez, D., Martic-Nieto, M., Velásquez-G, F. A., Pérez-Medina, J., & Luján-Mora, S. (2021). Towards a new learning experience through a mobile application with augmented reality in engineering education. *Applied Sciences, 11*(11), 4921. https://doi.org/10.3390/app11114921

Crompton, H., & Burke, D. (2018). The use of mobile learning in higher education: A systematic review. *Computers & Education, 123*, 53–64. https://doi.org/10.1016/j.compedu.2018.04.007

Dorouka, P., Papadakis, S., & Kalogiannakis, M. (2020). Tablets and apps for promoting robotics, mathematics, STEM education and literacy in early childhood education. *International Journal of Mobile Learning and Organisation, 14*(2), 255. https://doi.org/10.1504/ijmlo.2020.106179

Fabian, K., Topping, K. J., & Barron, I. G. (2015). Mobile technology and mathematics: Effects on students' attitudes, engagement, and achievement. *Journal of Computers in Education, 3*(1), 77–104. https://doi.org/10.1007/s40692-015-0048-8

Faizi, R., El Afia, A., & Chiheb, R. (2013). Exploring the potential benefits of using social media in education. *International Journal of Engineering Pedagogy (iJEP), 3*(4), 50. https://doi.org/10.3991/ijep.v3i4.2836

Fombona, J., Pascual, M. A., & Pérez Ferra, M. (2020). Analysis of the educational impact of M-learning and related scientific research. *Journal of New Approaches in Educational Research, 9*(2), 167. https://doi.org/10.7821/naer.2020.7.470

Greenhow, C., & Lewin, C. (2019). Social media and education: Reconceptualizing the boundaries of formal and informal learning. *Social Media and Education, 2019*, 6–30. https://doi.org/10.4324/9781315121697-2

Guy, R. (2009). The evolution of mobile teaching and learning. *Informing Science*.

Hamilton, E. (2019). *Why mobile apps are important for your business?* Retrieved from https://www.techtimes.com/brandspin/242588/20190504/why-mobile-apps-areimportant-for-your-business.htm

Hasan, M., Noor, T., Gao, J., Usman, M., & Abedin, M. (2022). The impact of financial knowledge on Fintech access: Client perspective evidence from Bangladesh. *Journal of the Knowledge Economy*.

Heflin, H., Shewmaker, J., & Nguyen, J. (2017). Impact of mobile technology on student attitudes, engagement, and learning. *Computers & education, 107*, 91–99.

Hendra Divayana, D. G., & Sanjaya, D. B. (2017). Mobile phone-based cipp evaluation model in evaluating the use of blended learning at school in Bali. *International Journal of Interactive Mobile Technologies, 11*(4).

Hermann, J., & Gruhn, V. (2018). M-learning to support project-oriented higher education in software engineering. In *New trends in intelligent software methodologies, tools and techniques* (Vol. 303, pp. 705–712). https://doi.org/10.3233/978-1-61499-900-3-705

Herrador-Alcaide, T. C., Hernández-Solís, M., & Hontoria, J. F. (2020). Online learning tools in the era of M-learning: Utility and attitudes in accounting college students. *Sustainability, 12*(12), 5171. https://doi.org/10.3390/su12125171

Hossain, S. F. A., Nurunnabi, M., & Hussain, K. (2021). Continuous mobile devices usage tendency in the TPACK-based classroom and academic performance of university students. *Technology, Pedagogy and Education, 30*(4), 589–607. https://doi.org/10.1080/1475939X.2021.1933160

Hossain, S. F. A., Xi, Z., Nurunnabi, M., & Anwar, B. (2022). Sustainable academic performance in higher education: A mixed method approach. *Interactive Learning Environments, 30*(4), 707–720. https://doi.org/10.1080/10494820.2019.1680392

Jaber, K., Abduljawad, M., Ahmad, A., Abdallah, M., Salah, M., & Alhindawi, N. (2021). E-learning mobile application evaluation: Al-Zaytoonah University as a case study. *International Journal of Advances in Soft Computing and its Applications, 13*(3), 89–99. https://doi.org/10.15849/ijasca.211128.07

Kacetl, J., & Klímová, B. (2019). Use of smartphone applications in English language learning—A challenge for foreign language education. *Education Sciences, 9*(3), 179. https://doi.org/10.3390/educsci9030179

Kamaghe, J. S., Luhanga, E. T., & Michael, K. (2020). The challenges of adopting M-learning assistive technologies for visually impaired learners in higher learning institution in Tanzania.

Khan, M. N. R., Yesmin, S., Aktar, M., Chowdhury, K. B. Q., Labeeb, K., & Abedin, M. Z. (2021). Techniques for multi-omics data incorporating machine learning and system genomics. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1524–1528). IEEE.

Khan, M., Nasfikur, R., Shakir, A. K., Nadi, S. S., & Abedin, M. Z. (2022). An android application for university-based academic solution for crisis situation. In *Sentimental analysis and deep learning* (pp. 653–663). Springer.

Klimova, B. (2019). Impact of mobile learning on students' achievement results. *Education Sciences, 9*(2), 90. https://doi.org/10.3390/educsci9020090

Klimova, B., & Polakova, P. (2020). Students' perceptions of an EFL vocabulary learning mobile application. *Education Sciences, 10*(2), 37.

Kumar Basak, S., Wotto, M., & Bélanger, P. (2018). E-learning, M-learning and D-learning: Conceptual definition and comparative analysis. *E-Learning and Digital Media, 15*(4), 191–216. https://doi.org/10.1177/2042753018785180

Laurens Arredondo, L. A., & Valdés Riquelme, H. (2021). M-learning adapted to the ARCS model of motivation and applied to a kinematics course. *Computer Applications in Engineering Education*. https://doi.org/10.1002/cae.22443

Llerena-Izquierdo, J., Barcia-Ayala, O., & Ayala-Carabajo, R. (2020). Faculty training through crowdlearning for emerging online education. In *2020 IEEE ANDESCON*. https://doi.org/10.1109/andescon50619.2020.9272103

Mishra, L., Gupta, T., & Shree, A. (2020). Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *International Journal of Educational Research Open, 1*, 100012. https://doi.org/10.1016/j.ijedro.2020.100012

Murphy, J. A., & Shelley, A. (2020). Textbook affordability in the time of COVID-19. *Serials Review, 46*(3), 232–237.

Naciri, A., Baba, M. A., Achbani, A., & Kharbach, A. (2020). Mobile learning in higher education: Unavoidable alternative during COVID-19. *Aquademia, 4*(1), ep20016. https://doi.org/10.29333/aquademia/8227

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., Niyato, D., Dobre, O., & Poor, H. V. (2022). 6G internet of things: A comprehensive survey. *IEEE Internet of Things Journal, 9*(1), 359–383. https://doi.org/10.1109/JIOT.2021.3103320

Oremus, B. (2021, December 30). *In 2021, tech talked up 'the metaverse.' One problem: It doesn't exist*. Retrieved from https://www.washingtonpost.com/technology/2021/12/30/metaverse-definition-facebookhorizon-worlds/

Papadakis, S., & Kalogiannakis, M. (2017). Mobile educational applications for children: What educators and parents need to know. *International Journal of Mobile Learning and Organisation, 11*(3), 256. https://doi.org/10.1504/ijmlo.2017.085338

Pedro, L. F., Barbosa, C. M., & Santos, C. M. (2018). A critical review of mobile learning integration in formal educational context. *International Journal of Educational Technology in Higher Education, 15*(1). https://doi.org/10.1186/s41239-018-0091-4

Pratama, A. R. (2020). Fun first, useful later: Mobile learning acceptance among secondary school students in Indonesia. *Education and Information Technologies, 26*(2), 1737–1753. https://doi.org/10.1007/s10639-020-10334-w

Shyshkanova, G., Zaytseva, T., & Frydman, O. (2017). Mobile technologies make education a part of everyday life. *Information and Learning Science, 118*(11/12), 570–582. https://doi.org/10.1108/ils-03-2017-0019

Singh, R., Timbadia, D., Kapoor, V., Reddy, R., Churi, P., & Pimple, O. (2021). Question paper generation through progressive model and difficulty calculation on the Promexa mobile application. *Education and Information Technologies, 26*(4), 4151–4179. https://doi.org/10.1007/s10639-021-10461-y

Slavova, Y., & Mu, M. (2018). A comparative study of the learning outcomes and experience of VR in education. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. https://doi.org/10.1109/vr.2018.8446486

Sofiana, N., & Mubarok, H. (2020). The impact of English game-based mobile application on students' reading achievement and learning motivation. *International Journal of Instruction, 13*(3), 247–258. https://doi.org/10.29333/iji.2020.13317a

Talan, T. (2020). The effect of mobile learning on learning performance: A meta-analysis study. *Educational Sciences: Theory & Practice, 20*(1), 79–103. https://doi.org/10.12738/jestp.2020.1.006

Trinh, L. T., Thao, T. T., Hang, T. T., Thanh, N. C., & Trung, T. (2021). Analysis of students' ability to accept M-learning technology: An exploratory study from high schools in Vietnam. *International Journal of Interactive Mobile Technologies (iJIM), 15*(12), 86. https://doi.org/10.3991/ijim.v15i12.22143

Wahyudi, W., Tukan, B. A. P., & Pinem, D. (2020). Analysis of the effect of financial literation, financial technology, income, and locus of control on lecturer financial behavior. *AFEBI Management and Business Review, 5*(1), 37–46.

Waliszewski, K., & Warchlewska, A. (2021). How we can benefit from personal finance management applications during the Covid-19 pandemic? The polish case. *Entrepreneurship and Sustainability Issues, 8*(3), 681.

Zhylenko, T. I., Martynova, N. S., Shuda, I. A., Chykalov, Y. A., & Kuzmuk, D. A. (2020). *Auto checker of higher mathematics – An element of mobile cloud education*. Retrieved from http://elibrary.kdpu.edu.ua/bitstream/123456789/3895/1/paper39.pdf

# Exploring the Role of Mobile Technologies in Higher Education: The Impact of Online Teaching on Traditional Learning

**Syed Far Abid Hossain, Armana Hakim Nadi, Rahma Akhter, Md. Ahmedul Islam Sohan, Faiza Tanaz Ahsan, Mahbuba Rahman Shofin, Saadmann Shabab, Tanusree Karmoker, and Krishna Paul**

**Abstract** The chapter aims to explore the role of mobile technologies in higher education especially the impact of online teaching on traditional learning. The transformation of the educational setting from online to offline draws limited attention from researchers in the post-pandemic era. The key reason for conducting this chapter is to explore the hidden issues of student coping strategies in the offline learning environment. In addition, the chapter explores the opportunities and limitations of technology usage in higher education. The study utilized a qualitative research approach to conduct the chapter with an extensive literature review. The result shows that with the advanced usage of mobile technology, the academic resources are freely available and accessible to all the learners that can ensure effective teaching and learning, however, the study is conducted among a limited number of respondents in a single country. This may affect the generalization of the study.

**Keywords** Mobile technologies · Higher education · Online teaching · Traditional learning

S. F. A. Hossain (✉) · R. Akhter
BRAC Business School, BRAC University, Dhaka, Bangladesh
e-mail: rahma.akhter@bracu.ac.bd

A. H. Nadi
Bangladesh University of Professionals, Dhaka, Bangladesh

M. A. I. Sohan · M. R. Shofin · T. Karmoker
IUBAT University, Dhaka, Bangladesh
e-mail: asohan@iubat.edu

F. T. Ahsan · S. Shabab · K. Paul
North South University, Dhaka, Bangladesh
e-mail: faiza.tanaz@northsouth.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_12

# 1 Introduction

Mobile technologies have enormous potential to transform education if they are developed and applied in a way that is appropriate for the social and cultural environment in which students learn with up-to-date technology. In the era of learning with industry 4.0 (Yang et al., 2022), mobile technology's use, implementation, and design in the higher educational setting provide technological and sociocultural obstacles. Various studies have shown that in nations such as the USA, the United Kingdom, Germany, France, and Japan, there are more 5G internet-capable mobile phones (Nguyen et al., 2022) with comparable functionality than genuine desktop computers (Rmenardi, 2012) that enhanced the learning trajectory such as deep learning (Abedin et al., 2021). Mobile technologies are described as all technological devices that are portable and lightweight (Lai et al., 2022) that can connect to the Internet via wireless connections or data cables, such as smartphones, iPads, and PDAs. M-learning is also defined as a dynamic learning environment enabled by the use of mobile technologies, especially in the fields of education (Keengwe & Bhargava, 2013). Given the digital environment of the twenty-first century, the application of mobile technology to education is essential to investigate how these applications change the social structure of learning environments in different learning environments, as well as how mobile technologies shape learning environments. By addressing different learning styles of learners and providing educational materials to everyone, anywhere, anytime, and in various versatile formats such as podcasts, audio recordings, or videos, mobile technologies can be of great help to education, strategy, organization, and content. Students and trainees working in distant field regions can communicate with their lecturers and obtain information via mobile devices from anywhere and at any time. Patients can benefit from mobile technology when used for notifications, reminders, language acquisition, motivation, and guiding. As a result, mobile technologies can provide a portable, lightweight learning platform that can result in private and spontaneous learning (Traxler, 2005). Mobile phones' IM (Instant Messaging) capabilities can aid in the creation of learning environments that improve knowledge transformation (Kekwaletswe, 2007). We have reached the mobile era, in which people carry their mobile gadgets with them at all times. Mobile technologies offer the potential to promote informal education from anywhere, at any time, and in any context. The major focus should be placed on recognizing that new learning applications arise through interaction and communication among the main participants in the development cycle and that mobile technologies are facilitating technology (Sharples, 2007). The development of modern society requires well-educated people. Mobile technologies have the potential to turn education into a seamless aspect of everyday life, to the point that people no longer identify it as training. The learning process will become natural and easy and the quality of learning will improve (Shyshkanova et al., 2017). The advancement of wireless technology in education, as well as the development of mobile apps, is astounding. Mobile technology in education has become one of the most significant areas of research and application in recent years.

For many educational institutions, mobile learning is becoming a crucial concern. Because new types of devices and apps are transforming education, it is critical to ensure that mobile learning is properly used and implemented (Sattarov & Khaitova, 2019).

Recent literature discovered diversified phenomena such as mobile applications to utilize financial decision support system (Abedin et al., 2019), continuous trend of smartphone usage in collaboration with TPACK-based lesson plan development (Hossain et al., 2021), evaluation of the FinTech opportunity for the organization with updated technological advancement (Hasan et al., 2022), sustainable academic performance in higher education with cutting-edge technology of smartphone in higher education (Hossain et al., 2022), complex and intelligence system development (Abedin et al., 2022), and many more; however, the impact of online teaching and learning (Hossain et al., 2019) on traditional teaching and learning style is still under shadow.

## 2 Literature Review on Mobile Technologies in Teaching

The debate about the use of technology in education dates back at least 2500 years. To better comprehend the role and impact of technology on education, we need to go back in time, because there are always lessons to be learned from history. One of the most comprehensive historical histories is Paul Saettler's "The Evolution of American Educational Technology" in 1990; however, it only covers up to 1989. Since then, a lot has transpired. Teemu Leinonen has a wonderful research article on recent history as well (Leinonen et al., 2010) (Fig. 1).

During the 1990s, the expense of making and dispersing video dropped significantly because of computerized pressure and rapid Internet access. This decrease in the expenses of recording and appropriating video likewise prompted the improvement of talk catch frameworks. The innovation allows understudies to view or audit addresses whenever and place with an Internet association. YouTube began in 2005 and is progressively being utilized for short instructive clasps that can be downloaded and coordinated into online courses. It is also seen that The Khan Academy began using the YouTube platform in 2006 for recorded voice-over addresses involving an advanced chalkboard for conditions and delineations. Apple Inc. made iTunes U in 2007 to turn into a gateway or a webpage where recordings and other computerized materials on college instruction could be gathered and downloaded for nothing by end clients.

Technology puts students on the way to customizing learning by giving them the power to control their studies, make education relevant to their digital lives, and prepare them for their future. Students are driven to become reflective practitioners, collaborators, creators, and critical thinkers as a result of access to technologies and resources outside the classroom. When technology is well integrated into the classroom, students have a lifetime of learning love (Arnold & Sangrà, 2018). Instructors are always working to customize learning for their students. Technology can help

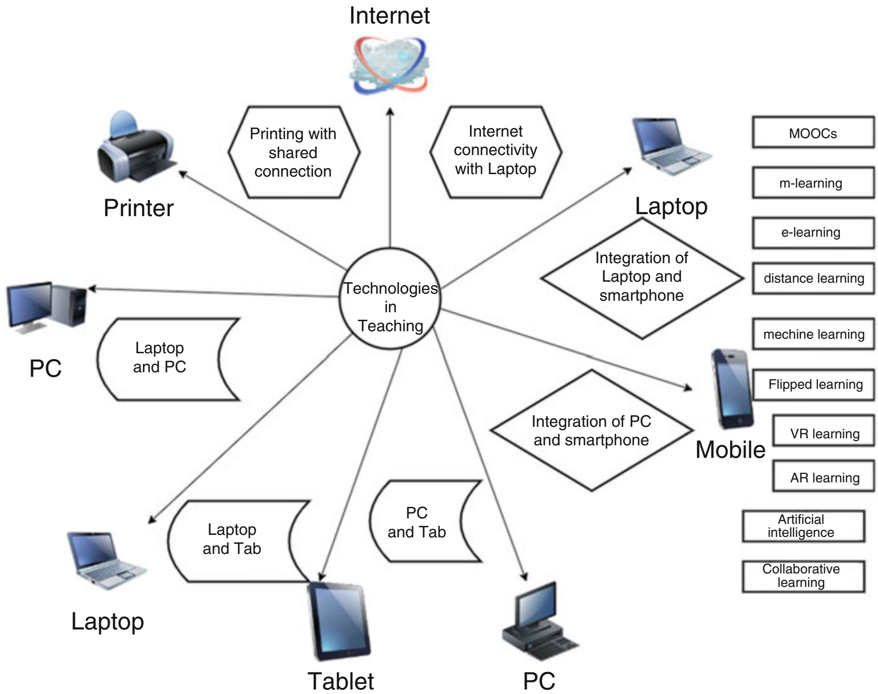



**Fig. 1** Technologies in teaching (Source: Author's own creation)

them reach new heights by accessing student data in real time, transverse information, materials, applications, and more. Software can improve teachers' efforts in creating hybrid learning environments and using digital tools for summative and formative evaluations, introducing new paradigms of teaching and learning into classrooms.

A study by Intel Corporation suggests that digital learning, as well as having the correct devices in students' hands, helps them prepare for the professional life and provide them with technical skills needed by the future workforce. Relevant STEAM learning experiences can provoke creativity, help students apply meaning to learning, and plan future career opportunities and undeveloped careers. Physical computation, coding, programming, and computer thinking skills are common in this profession. Students may learn these techniques while also improving their critical thinking and problem-solving skills for the twenty-first century through the creation.

Design and proper technology can make learning with manufacturers and the environment very stimulating. School and universities face difficulties in deciding which devices and technologies will help them realize their ambitions of changing learning. Working with various stakeholders to evaluate how teachers and students use devices for daily learning, devices should be used to select devices. Stakeholders must consider acceptable content requirements, grade-level curricula, and how devices will be used. It is not a simple chore, but factors like assessment needs, security features, compatible digital curriculum and material, management choices, device performance, and total cost of ownership all play a role in selecting the correct device. The basis of a 360-degree learning experience is a safe and strong IT infrastructure that supports digital material, protects important student data, increases operational efficiency, and ensures safety of the students.

Besides the development and use of virtual classrooms and online-based education platforms, online media are actually a subclass of PC innovation; however, their improvement merits its very own segment throughout the entire existence of instructive innovation. Web-based media cover a wide scope of various advancements, including websites, wikis, YouTube recordings, cell phones like telephones and tablets, Skype, Facebook, and Twitter. Kaplan & Haenlein, (2010) characterize web-based media collectively "*of Internet based applications that permit the creation and trade of client produced content, in light of cooperations among individuals in which they make, offer or trade data and thoughts in virtual networks and organizations*."

The gap in the past literature reviews that we are going to address in this paper is the substitutability of online classes with a physical classroom-based study session due to the prevalence of COVID-19 since early 2020. The tsunami of web-based learning has occurred. Many schools offer on the Web (virtual) learning for understudies as a method of continuing education during the remainder of the school year. Educationalists and directors who hesitantly teach on the Internet have only a few choices to accept the decade-old innovation. Some instructors may encounter fears and fear when moving their home room to the Internet, but most of them do so quickly and within a short period of time; over the long haul, everybody appears to adjust well. The advanced separation is more obvious than ever in recent memory (Guernsey et al., 2020). Children who can bring computers are ready. Educators and showing strategies are a piece behind; notwithstanding, there is confidence in the creation of a new school model. Change can be valuable.

## 3 The Influence of Mobile Technologies in Teaching

In today's world, most of the population keeps smartphones in their possession at a very early age (Han, 2022). It goes without saying that mobile technologies are used for much more than just communication. In fact, mobile technology is one of the most recent tools to support real-world learning (Hashim, 2018).

Like any other technology available in the world in the contemporary era, mobile technologies are no different in terms of influencing users and stakeholders both positively and negatively, especially with hedonic usage (Vujić & Szabo, 2022). Research shows that mobile technologies are associated with a positive perception of students in collaborative learning, but that students are more dissatisfied in class (Heflin et al., 2017). Positive influences of modern technology on education include: globalization and improvements in education and learning without geographical restrictions. In contrast, negative influences include: increasing incidents of cheating, declining writing skills, and lack of focus (Raja & Nagasubramani, 2018).

Currently, especially in this post-COVID era, the usage of mobile technology has become part and parcel of education. Mobile technologies have facilitated improved means of education through increased portability and easy access to the Internet. Now teachers and students can search for a topic and learn on-the-go. Mobile devices enable students to easily access education content from any place and at any time (Criollo-C et al., 2018). Students are able to tutor themselves through video tutorials or downloadable bite-sized lessons from the Internet. Through the means of online education facilitated by smartphones apps, students around the world are now able to familiarize themselves with international contexts. Furthermore, students hailing from different areas of the world are now able to attend online classrooms at the same time through virtual meeting apps such as Zoom, Google Meet, etc., which greatly reduces geographical barriers. Students of the modern era can easily communicate and enhance their network on a global scale through mobile technologies.

However, there are certain negative impacts to this facility. It is true that mobile technologies have facilitated on-the-go learning, but this also means that anyone anywhere can search on topics they want to learn or, if they intend to, copy in their exams. Some students may tend to use unfair means in their examination through mobile technology. Moreover, depending on the texts and material composed by others is also greatly reducing the creativity of students. Instead of coming up with their own ideas, students are becoming dependent on information which is already available through online media. This, in turn, also results in the lack of focus of students and the creation of a mindset among them to depend on online educational materials without giving much concentration in classroom sessions.

## 4 Mobile Technologies Available via an Online Platform

Even in the recent past, smartphones were considered hazardous to the educational well-being of students, and parental control seems very strict according to the existing literature (Hadad et al., 2020). However, this scenario has changed to a great extent at present. Utilizing new innovations in technologies, smartphone apps have contributed significantly in the field of education. At the moment, there are hundreds, if not thousands, of mobile apps providing educational support to students, scholars, and teachers from around the world. Moreover, most of these educational apps are free. Among the most notable free educational apps, the ten most prominent ones are mentioned in the table below (Mindster, 2020) (Table 1).

Apart from the mobile apps mentioned in the table above, other mobile technologies are also available that contribute a lot to the field of education. Figure 2 represents various income groups with life expectancy. The overall income or GDP is an indicator of the use of individually owned technological devices. According to life expectancy data, the use of technology in the classroom may vary significantly. For example: the use of cloud-based Learning Management Systems such as Moodle, Blackboard, etc., in combination with web conferencing platforms such as Zoom, Google Meet, etc. has revolutionized the education sector.

**Table 1** Various mobile apps for educational purposes available on the online platform

| Mobile App | Description | Source |
|---|---|---|
| Google classroom | A virtual classroom that facilitates submitting and grading assignments, sending announcements, starting discussions, creating classes, sharing resources, asking for remarks and answers, and so on. | Tarteer et al. (2021) |
| edX | Educational material from top universities such as Harvard, MIT, Columbia, etc. including compilation of more than 2000 courses like engineering, computer science, linguistics, business studies, and many more. | Shi and Lin (2021) |
| Khan academy | Platform providing lessons in the form of video tutorials. The video shows the drawing recorded on the virtual black board that the narrator shows. Khan academy also offers online courses to prepare standardized exams such as SAT, MCAT, and LSAT. | Massey et al. (2022) |
| Duolingo | Language learning app that facilitates learning of 30+ languages in an interactive way through mini games. The app also tracks the performance of the learner and provides insight. | Ahmed et al. (2022) |
| Remind | Community that helps students learn in groups and stay connected. Remind is also used to message the entire class, submit assignments, share photos and handouts, and clear doubts with friends, individually and in collaboration. | Jones et al. (2022) |
| Photomath | Solves mathematical problems by providing step-by-step explanations and instructions to the learner by utilizing submitted photos; either handwritten or printed. | Long and Bouck (2022) |
| SoloLearn | A platform providing tutorials for learning coding languages such as C++, Java, python, swift, JavaScript, CSS, PHP, HTML, and so on. | Elsisi et al. (2022) |
| Quizlet | Simple tools that help students practice and master any topics they prefer. Quizlet allows learners to design their own sets or gather sets from other contributors and study them. | Senior (2022) |
| Kahoot | Provides ready-made quizzes on any topic of interest. Learners can take individual quizzes or participate in live quizzes with other learners. | Vijayakumar (2022) |
| uDemy | Holds more than 130,000 video tutorials for courses ranging from business and technology to personal development. If the learner feels stuck in a particular lesson, he or she can ask questions to other students and teachers and solve their doubts. | Moslehi et al. (2022) |

Both students and teachers are enjoying the benefits of online education through these services.

## 5 Popular Applications in Higher Education

Higher education institutions have begun to use mobile technologies to improve education quality (Han & Shin, 2016). Although institutions do not only improve the quality of higher education, they also assist students in learning. Numerous tools and

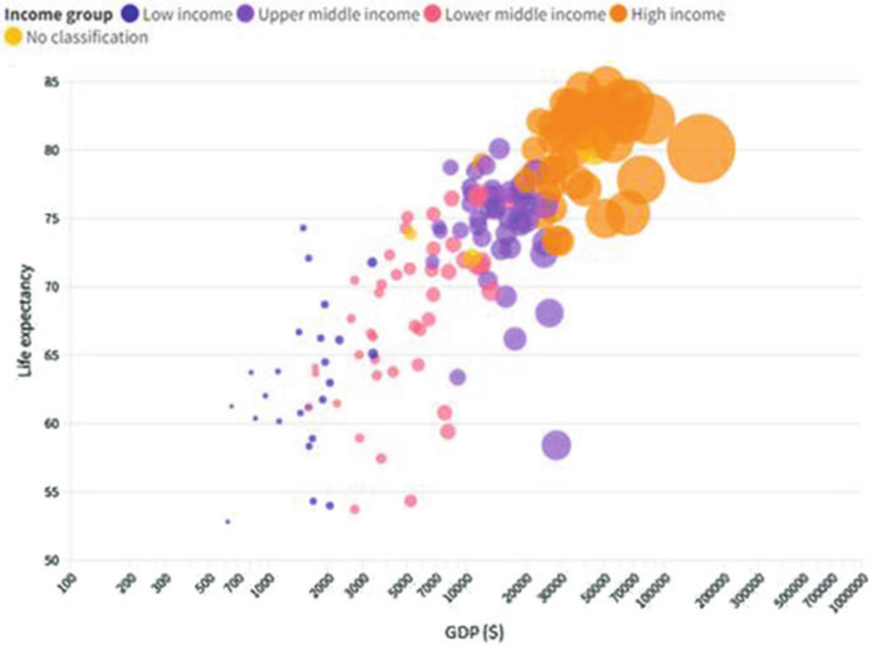



**Fig. 2** Various income groups with life expectancy

applications are available to imply the overall higher education system. The most common applications are such as Virtual reality applications (Radianti et al., 2020), Game-based learning applications (Subhash & Cudney, 2018), Blockchain-based applications (Awaji et al., 2020), and so on.

However, the twenty-first century learning aid is virtual reality (Rogers, 2019). According to one study, after participating in VR activities, students can retain more information and better apply the information they have learned (Krokos et al., 2019). The authors suggest that integrated gameplay features are more efficient in increasing learner engagement (Scholz et al., 2021). The game-based application can help students assess learners and personalized collaboration in group playing, enhancing the level of knowledge (Troussas et al., 2020). Students and educational institutions can also use a blockchain-based application to create more personalized digital agreements that include assignment criteria, time frames, and grading procedures (Chen et al., 2018). Despite that, EduCTX is a blockchain-based decentralized higher education credit network (Turkanovic et al., 2018). These systems are flexible, secure, and reliable due to their global storage capacity and resource sharing (Coulouris et al., 2012). According to research, most higher education institutes offer online courses to their students through websites, learning platforms, video conferencing (Google Meet, Google Duo, Google Hangout, and Zoom), YouTube, Social media (Facebook and Twitter), and several other free websites for blended learning tools (Shahzad et al., 2020). Furthermore, emerging technology improved several aspects of student participation in the three dimensions of engagement, with web conferencing software, digital games, and Facebook seem to be the most significant (Schindler et al., 2017). The features of mobile technology, such as portability and mobility, improve their potential application in education through the use of digitalized library and information access; many higher education institutions worldwide are exploring its possible use in higher education (Yip et al., 2020). It can remove space and time limitations to education, thereby greatly expanding participation in higher education. Therefore, online application has become a necessary and ongoing investment in the modern era, and most educational institutions must gather additional funds to fulfill their expanding needs for technology resources (Baldwin, 2021). There is little doubt that technology will be one of the driving forces in the development of higher education. In existing models, technology should be considered a component that impacts student involvement. Soon, the development of educational apps will be a once-in-a-lifetime opportunity for developers.

## 6 Higher Education from Online to Offline Setting

After the global pandemic has stabilized, the education system will move from online to offline platforms. In an offline setting, professors have spent a significant portion of their class time with students distributing material through lectures and follow-up discussions. As a large group of participants, universities have been forced to take similar preventive measures to minimize the impact of COVID-19 on higher education and beyond (Liguori & Winkler, 2020). The current increase in education is the technology to combine teaching information online and offline. However, when higher education institutions adopted online learning settings, they faced challenges, but when considering online learning to offline, it seems back to the traditional learning process. The authors investigated the high combination of digital technologies and academic teaching to grow students' autonomous learning ability and increase classroom vitality, which is valuable and relevant for improving classroom teaching (Chen et al., 2020). Higher education in online learning has formal and informal approaches; on the other hand, offline learning is formal to increase its role as a learning platform to provide students with various learning opportunities. Offline teaching increases active interaction between teachers and students, even though online learning allows students to study solely via the Internet on their own. In the Covid-19 pandemic and post-epidemic era, the teaching model combines the benefits of online and offline learning in higher education (Gao & Lu, 2021). Currently, online practices are connected to offline instruction to better understand learning materials. According to a study, online teaching is ineffective than face-to-face learning (Liu & Han, 2020). In addition, transformation methods from online to offline enhance student communication with the instructor, active learning, and reduce procrastination.

## 7 The Impacts of Mobile Technologies on University Students

Mobile technologies have come a long way since their inception. In the past, people used mobile phones that would be weighted like a dumbbell. However, thanks to the advancement in mobile technology, people now use phones that fit in the palm of their hands and have bodies that are mostly screens. Mobile technology affects people of all ages but in the recent times the group of people that it has had the most effect on are students. Studies have found that students perceive collaborative learning positively while learning through mobile technology (Heflin et al., 2017). Other than this, there are many other implications that mobile technology has on students. The use of mobile technology has been associated with higher academic performance of students. On top of this, using mobile technologies for learning can also bring psychological comfort to students who use their mobile devices all the time. Mobile technology even has social implications on students, such as integrating education into their lives as a natural process and not as a training one (Shyshkanova et al., 2017). In general, mobile technologies increase peer-to-peer engagement and also increase participation in learning activities (Fabian et al., 2015). However, not all the impacts that mobile technologies have are positive. There are many negative impacts that mobile technology can have on students. Mobile technology, while increasing student engagement, can also be the cause of distraction. Since students access their social lives mainly on their mobile devices, it is quite easy for them to become distracted with social media while trying to engage in any learning activities. Another negative impact mobile technology can have on students is that it can increase the number of students using unfair means on tests and assessments since they have full access to the Internet and no one to supervise them. Another study found that the more mobile technology is used for social interaction, the more it negatively affects the quantity and quality of face-to-face interactions (Elsobeihi & Abu Naser, 2022). In other words, due to too much reliance on mobile technology for social interaction, users become more and more disengaged to in-person social interaction.

## 8 The Impacts of Variation in Assessment in Higher Education

Covid-19 has caused changes in almost all aspects of everyone's lives. From how people work to how people commute. This pandemic has once again proven just how adaptable humans are. However, one of the largest industries that Covid-19 has impacted is the education industry. The education sector has changed enormously toward E-learning and M-learning. Although most of the impact has been positive, there are still some negative impacts that M-learning has had on education. One such area where M-learning has had both a positive and a negative impact is in the assessment process. One study states that lack of preparation and the inherent downsides of remote assessment have proven to be an extraordinary challenge for higher education assessment. Some of these challenges include dishonesty, the lack of proper infrastructure, submission deadline commitment, and so on (Guangul et al., 2020). Since teachers and faculties do not have any way to observe their students live in a controlled environment, there are high chances that students use unfair and dishonest means to complete their assessments. Although assessment technology has come very far in just a couple of years, it is yet to be considered as the solution which can completely eliminate dishonest means in assessments. Again, it is not the case that the students are always at fault. Sometimes students are the one who fall victim to the lack of proper infrastructure, leading to late submission or even missed assessments.

However, not all the impacts of online assessments have been negative. Mobile learning has enabled students and teachers both to access the assessments at their convenience time and place. Teachers also face challenges when it comes to online assessments. Since in M-learning there is no face-to-face interaction, teachers find difficulty in conveying their intentions (Kearns, 2012). Another substantial problem that is common with online assessment is the risk that students will get the assessment responses in advance. This can be caused by various factors like faulty infrastructure, hacking, and even dishonest assistants. Of course, since everyone takes online assessments in their convenient time, thus students can easily share the answers with their peers taking the assessment in a different time than theirs (Rowe, 2004). All in all, online assessment is the part of M- or E-learning that needs to be developed the most, and although there are certain advantages to online assessments, the disadvantages of online assessments overshadow them.

## 9 Traditional, Online, or Blended Learning?

Traditional learning is a face-to-face interaction process; through this way, there is no need to worry about security and confidentiality issues in traditional education, as in the case with online education; as well as in this process, a student of higher education can gain connection, inspiration, availability, structure, and so on (Razeeth et al., 2019). Also, connectivity leads to direct communication between students and professors so that higher education as a consequence students are enabled to expand their collaborative activities and eliminate direct doubts about specific issues promptly, which is different from online learning. Furthermore, most of the time in traditional learning professors present and discuss topics; on the other hand, students pay close attention and try to understand the topics (Azzalis et al., 2009). Various scholarly articles indicate that this way of learning improves students' capacity to recall and grasp new content (Hyun et al., 2017). But in the age of technological advancement, students want to be able to read material from anywhere, and it has become possible through online learning in higher education. Due to this need, online education has become an effective and desirable choice. Online learning is becoming highly popular among students in higher education, as well as they believe that the traditional learning format is rigid, authoritarian, and unsustainable and higher education may now provide efficient classroom instruction through the Web in this advancing age (Paul & Jefferson, 2019).

For higher education, students want to have a better education without having to abandon employment, home life, or transportation costs. In addition, online learning students have the opportunity to contact professors, engage friends and classmates, study documents, and finish all the class tasks through any Internet accessible point, rather than needing to be in a given place at a particular time frame (Richardson & Swan, 2003). As online learning is growing in popularity, various higher education institutions are fond of determining the best way to distribute course content among online students (Dumford & Miller, 2018). As a result, higher education institutions have begun to embrace mobile technologies to meet student requisites (Han & Shin, 2016). The first and foremost reason for learning online these days is the assault of the Covid-19 virus, which has led to large-scale migration from traditional face-to-face learning to online learning. Millions of teaching members across the world began lecturing in front of electronic screens shortly after the start of 2020, while their pupils were required to remain at home and attend courses over the Internet (Bao, 2020). Another thing is that online learning is more flexible than traditional learning. In response to fears about the rapid spread of the coronavirus around the world, a large number of educational institutions around the world have temporarily stopped face-to-face classes to prevent it from spreading, leading universities around the world to shift more toward online learning, and other research authors have also suggested online and distance education as a necessity during social distance with lockdown due to the COVID-19 pandemic (Ali, 2020). The coronavirus has also shown new threats to the entire education system, demonstrating that society needs a reliable and versatile education system to confront an uncertain future. Another learning term is blended learning (BL), which combines traditional face-to-face learning alongside online learning, is a technological advancement that is drastically revolutionizing teaching and learning in higher education, and is becoming more popular in higher education. And blended learning is often used in a combination of phrases that include merged flexible, mixed mode, or hybrid learning (Anthony et al., 2019). Previous research tested the efficacy of blended learning by comparing traditional and online teaching, as there has been tremendous progress in blended learning that has emphasized improving learning and teaching outcomes (Van Laer & Elen, 2020). Online activities such as wordbooks, study guides, online writing tools, discussion forums, web addresses, video tutorials, relevant materials, models, exercises, quizzes, and so on are all part of the layout and execution of blended learning online educational materials (Anthony et al., 2019). Inversely, traditional face-to-face education includes lectures, laboratory activities, face-to-face practice and skills assessment, individual/group presentations, and professor-led discussions to assess students' academic performance (Sun & Qiu, 2017).

According to the results of a previous research paper, blended learning methods improve the acquisition of knowledge, learning engagement, and wisdom because it has a remarkable impact on the consciousness and learning backgrounds of students and emphasizes learning from blended learning (Edward et al., 2018), thus guiding students in becoming more engaged in the learning process and allowing them to be more enthusiastic, which enhances their patience and dedication (Ghazal et al., 2018). Blended Learning uses a blend of online and traditional face-to-face (F2F) learning to assist professors in achieving educational goals in higher education students, to build efficient and productive logical knowledge, help improve educational aspects, and establish social discipline (Subramaniam & Muniandy, 2019). Keeping in mind student and lecturer perspectives, blended learning works to establish a peaceful, coherent equilibrium, prosperous, and healthy combination among online information availability and traditional learning in higher education (Bervell & Umar, 2018). A previous research also mentioned that blended learning comprises a combination of several activities, which is achieved by integrating 70% online learning and 30% face-to-face engagement (Anthony et al., 2019). Students' enthusiasm in their learning path grows as a result of blended learning (Chang-Tik, 2018), allows students to learn at their own pace, and prepares students for the future by giving real-world knowledge and skills (Ustunel & Tokel, 2018), which let students promptly use their academic capabilities, self-learning skills, and obviously, computerized know-how in the workplace (Yeou, 2016). The authors also stated that blended learning positively affects socialization in higher education, increases student intellectual ability and self-reliance capacity, improves student learning quality, improves their ability to think critically, and combines advanced technologies as an operational tool to demonstrate course curriculum to students (Al-shami et al., 2018). However, prior research authors are mostly recommending blended learning as an active education in higher education.

## 10 Financial Profitability and Complexity Among Learners

Mobile and electronic learning processes have been introduced among learners to eradicate education barriers. It is undeniable that m-learning reduces the cost of learners and may bring the whole process to fruition. According to the recent theoretical developments, this medium of learning has enabled accessibility for learners. The introduction of mobile learning among people has ensured the sustainability of education. In addition to the benefits of m-learning, it has some definite intricacies. This report will find the financial profitability and complexity with which learners may deal while obtaining this medium.

### *10.1 Financial Profitability*

The advancement of technology is quickly becoming more efficient and faster. M-learning has been facilitated by technology that helps enhance the collaboration between the student and the teacher. Changing the approaches to learning is not only the motive of this process, but also makes education more affordable for the learners. In the context of availability, the lectures, tools, and other materials of learning are available on the required application or website. This helps learners practice anytime. Mobile technologies have helped to adopt the new learning process that improves the traditional learning method.

Indisputably, online learning helps save money and also allows users to access any content. The books may not be affordable to some people who are from remote areas of developing or underdeveloped countries. Online courses and classes have been financially convenient for students. The process improves the educational system while being financially beneficial to learners. Through online learning, learners can get financial profits such as saving them money, accommodating in a comfortable place, commuting costs, expenses of buying materials, and so on. There is no other alternative way than accessing all the content through m-learning which diminishes the cost of buying books and other accessories.

The profitability demonstrates that students can learn sustainably. Learners may collaborate with teachers while connecting online, and it ensures cost-effectiveness. The cost includes proctoring of exams, which may help invigilators as well as students save the money of transportation. Online learning such as m-learning and e-learning does not require learners or trainees to purchase books as all the materials and PDFs are already uploaded online. Mobile technologies have established mobile education to enable learning to be affordable and accessible.

Learners can attend classes or courses through mobile learning technologies that help them learn virtually instead of spending transportation or any other accommodation cost. Because of mobile phone education, people don’t need to leave their city or areas for training, college, or any other institution. Some people may work while learning online or reading content by mobile phone which would not hamper the job. This approach has been influential because learners can save operational costs and also printing costs. Previously, they needed to buy printing copies and also print the documents and files with the expense. Online courses helped them reduce the cost of these tools. Online education does not only provide financial benefits with education, but also helps lessen additional costs including meal plans, room-and-board.

## *10.2 Financial Complexity*

Students who live in remote areas cannot afford the Internet and high-end devices. The cost of mobile devices is a challenging financial issue for learners, and the impact of mobile education from the financial perspective may affect the learning process. Sometimes, online learning requires a high-configured computer with available tools that become difficult to obtain.

## 11 Conclusion

Learning through mobile education is a progressive way of learning and practicing. Mobile education technology is a mainstream medium that is helping students with content, pdfs, and saving time. Analysis of the past decades has shown that online learning has integrated distance-educated students around the world. The emerging technology of mobile education has increased education in a great way. The results have shown that the learners have positive attitudes toward mobile learning and online education with respect to the current phenomena. Educational technology is emerging in its learning process. Academic resources are available and accessible to all learners, making education more flexible. Problems related to the needs of learners are usually overcome by evaluating their attitudes. Online learning is easy to adopt and appropriate for exchanging information with faculty and students, working from anywhere, and also learning new technological features. However, some learners and teachers have reported issues while working online and using technology. The technology advancement made the inferiority complex among learners, and also minimized the social interaction, which makes people antisocial. Universities, colleges, and other institutions should analyze the effectiveness and provide proper knowledge in the research and learning process. Mobile education technologies should be manufactured in a way that students can afford them. Academicians should formulate a proper policy on the use and operation of mobile phones in education to avoid misuse and bullying. It is undoubtedly true that interactivity is the key element of learning and online learning ensures giving prompt feedback on their performances. Implementing online learning in higher education is a huge initiative for the future, and this makes education more creative and feasible.

## References

Abedin, M. Z., Guotai, C., Moula, F.-E., Azad, A. S. M. S., & Khan, M. S. U. (2019). Topological applications of multilayer perceptions and support vector machines in financial decision support systems. *International Journal of Finance & Economics, 24*(1), 474–507. https://doi.org/10.1002/ijfe.1675

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2021). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research, 2021*, 1. https://doi.org/10.1007/s10479-021-04420-6

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class imbalanced prediction of small business credit risk. *Complex & Intelligent Systems.* https://doi.org/10.1007/s40747-021-00614-4

Ahmed, A. A. A., Hassan, I., Pallathadka, H., Keezhatta, M. S., Noorman Haryadi, R., Al Mashhadani, Z. I., Attwan, L. Y., & Rohi, A. (2022). MALL and EFL learners' speaking: Impacts of Duolingo and WhatsApp applications on speaking accuracy and fluency. *Education Research International, 2022*, 1–10. https://doi.org/10.1155/2022/6716474

Ali, W. (2020). Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic. *Higher Education Studies, 10*(3), 16. https://doi.org/10.5539/hes.v10n3p16

Al-shami, S. A., Aziz, H., & Rashid, N. (2018). The adoption of MOOC utilization among undergraduate students in Universiti Teknikal Malaysia Melaka (UTEM). *Journal of Fundamental and Applied Sciences, 10*(6), 2634–2654. Retrieved from https://www.ajol.info/index.php/jfas/article/view/172198

Anthony, B., Kamaludin, A., Romli, A., Raffei, A. F., Nincarean, A., Eh Phon, L., Abdullah, D., Ming, G. L., Shukor, N. A., Nordin, M. S., & Baba, S. (2019). Exploring the role of blended learning for teaching and learning effectiveness in institutions of higher learning: An empirical investigation. *Education and Information Technologies, 24*(6), 3433–3466. https://doi.org/10.1007/s10639-019-09941-z

Arnold, D., & Sangrà, A. (2018). Dawn or dusk of the 5th age of research in educational technology? A literature review on (e-) leadership for technology-enhanced learning in higher education (2013-2017). *International Journal of Educational Technology in Higher Education, 15*(1), 24.

Awaji, B., Solaiman, E., & Albshri, A. (2020). Blockchain-based applications in higher education. In *Proceedings of the 5th International Conference on Information and Education Innovations*. https://doi.org/10.1145/3411681.3411688

Azzalis, L., Sato, S., De Mattos, M., Fonseca, F., & Giavarotti, L. (2009). Active learning versus traditional teaching. *Revista de Ensino de Bioquímica, 7*(2), 2. https://doi.org/10.16923/reb.v7i2.186

Baldwin, R. (2021). *Technology in education*. Retrieved from https://education.stateuniversity.com/pages/2496/Technology-in-Education-HIGHER- EDUCATION.html

Bao, W. (2020). COVID -19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies, 2*(2), 113–115. https://doi.org/10.1002/hbe2.191

Bervell, B., & Umar, I. N. (2018). Blended learning or face-to-face? Does tutor anxiety prevent the adoption of learning management systems for distance education in Ghana? *Open Learning: The Journal of Open, Distance and e-Learning, 35*(2), 159–177. https://doi.org/10.1080/02680513.2018.1548964

Chang-Tik, C. (2018). Impact of learning styles on the community of inquiry presences in multidisciplinary blended learning environments. *Interactive Learning Environments, 26*(6), 827–838. https://doi.org/10.1080/10494820.2017.1419495

Chen, G., Xu, B., Lu, M., & Chen, N. (2018). Exploring blockchain technology and its potential applications for education. *Smart Learning Environments, 5*(1). https://doi.org/10.1186/s40561-017-0050-x

Chen, H., Wu, S., & Shi, J. (2020). Design of online and offline blending teaching mode. In 2020 *IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI)*. https://doi.org/10.1109/csei50228.2020.9142519

Coulouris, G. F., Dollimore, J., & Kindberg, T. (2012). *Distributed systems: Concepts and design*. Springer.

Criollo-C, S., Lujan-Mora, S., & Jaramillo-Alcazar, A. (2018). Advantages and disadvantages of M-learning in current education. In *2018 IEEE World Engineering Education Conference (EDUNINE)*. https://doi.org/10.1109/edunine.2018.8450979

Dumford, A. D., & Miller, A. L. (2018). Online learning in higher education: Exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education, 30*(3), 452–465. https://doi.org/10.1007/s12528-018-9179-z

Edward, C. N., Asirvatham, D., & Johar, M. G. (2018). Effect of blended learning and learners' characteristics on students' competence: An empirical evidence in learning oriental music. *Education and Information Technologies, 23*(6), 2587–2606. https://doi.org/10.1007/s10639-018-9732-4

Elsisi, M., Tran, M., Mahmoud, K., Mansour, D.-E. A., Lehtonen, M., & Darwish, M. M. F. (2022). Effective IoT-based deep learning platform for online fault diagnosis of power transformers against cyberattacks and data uncertainties. *Measurement, 190*, 110686. https://doi.org/10.1016/j.measurement.2021.110686

Elsobeihi, M., & Abu Naser, S. (2022). *Effects of mobile technology on human relationships*. Retrieved 13 January 2022, from http://hdl.handle.net/10419/167579

Fabian, K., Topping, K. J., & Barron, I. G. (2015). Mobile technology and mathematics: Effects on students' attitudes, engagement, and achievement. *Journal of Computers in Education, 3*(1), 77–104. https://doi.org/10.1007/s40692-015-0048-8

Gao, P., & Lu, M. (2021). Research on closed-loop online and offline hybrid teaching from the first perspective of students. In *2021 16th International Conference on Computer Science & Education (ICCSE)*. doi:https://doi.org/10.1109/iccse51940.2021.9569389

Ghazal, S., Al-Samarraie, H., & Aldowah, H. (2018). "I am still learning": Modeling LMS critical success factors for promoting students' experience and satisfaction in a blended learning environment. *IEEE Access, 6*, 77179–77201. https://doi.org/10.1109/access.2018.2879677

Guangul, F. M., Suhail, A. H., Khalit, M. I., & Khidhir, B. A. (2020). Challenges of remote assessment in higher education in the context of COVID-19: A case study of Middle East college. *Educational Assessment, Evaluation and Accountability, 32*(4), 519–535. https://doi.org/10.1007/s11092-020-09340-w

Guernsey, L., Ishmael, K., & Prescott, S. (2020). *Online learning in the wake of COVID-19 tips and resources for PreK-12 with equity in mind, New America*. Retrieved from https://www.newamerica.org/education-policy/edcentral/online-learning-wakecovid-19/

Hadad, S., Meishar-Tal, H., & Blau, I. (2020). The parents' tale: Why parents resist the educational use of smartphones at schools? *Computers & Education, 157*, 103984. https://doi.org/10.1016/j.compedu.2020.103984

Han, S. (2022). Impact of smartphones on students: How age at first use and duration of usage affect learning and academic progress. *Technology in Society, 70*, 102002. https://doi.org/10.1016/j.techsoc.2022.102002

Han, I., & Shin, W. S. (2016). The use of a mobile learning management system and academic achievement of online students. *Computers & Education, 102*, 79–89. https://doi.org/10.1016/j.compedu.2016.07.003

Hasan, M., Noor, T., Gao, J., Usman, M., & Abedin, M. Z. (2022). Rural consumers' financial literacy and access to FinTech services. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-00936-9

Hashim, H. (2018). Application of technology in the digital era education. *International Journal of Research in Counseling and Education, 1*(2), 1. https://doi.org/10.24036/002za0002

Heflin, H., Shewmaker, J., & Nguyen, J. (2017). Impact of mobile technology on student attitudes, engagement, and learning. *Computers & Education, 107*, 91–99. https://doi.org/10.1016/j.compedu.2017.01.006

Hossain, S. F. A., Shan, X., & Nurunnabi, M. (2019). Is M-learning a challenge?: Students attitudes toward the sustainable learning and performance. *International Journal of e-Collaboration, 15*(1), 21–37. https://doi.org/10.4018/IJeC.2019010102

Hossain, S. F. A., Nurunnabi, M., & Hussain, K. (2021). Continuous mobile devices usage tendency in the TPACK-based classroom and academic performance of university students. *Technology, Pedagogy and Education, 30*(4), 589–607. https://doi.org/10.1080/1475939X.2021.1933160

Hossain, S. F. A., Xi, Z., Nurunnabi, M., & Anwar, B. (2022). Sustainable academic performance in higher education: A mixed method approach. *Interactive Learning Environments, 30*(4), 707–720. https://doi.org/10.1080/10494820.2019.1680392

Hyun, J., Ediger, R., & Lee, D. (2017). Students' satisfaction on their learning process in active learning and traditional classrooms. *International Journal of Teaching and Learning in Higher Education, 29*(1), 108–118. https://doi.org/10.1186/s41239-017-0041-6

Jones, A., Fouladi-Nashta, S., Eady, N., Bedeau, S., & Hafezi, H. (2022). "Just say no" (or at least ask why) STOMP medication reviews in tower hamlets community learning disability service. *BJPsych Open, 8*(S1), S182–S182. https://doi.org/10.1192/bjo.2022.504

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68.

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. *MERLOT Journal of Online Learning and Teaching, 8*(3).

Keengwe, J., & Bhargava, M. (2013). Mobile learning and integration of mobile technologies in education. *Education and Information Technologies, 19*(4), 737–746. https://doi.org/10.1007/s10639-013-9250-3

Kekwaletswe, R. (2007). Social presence awareness for knowledge transformation in a mobile learning environment. *International Journal of Education and Development using ICT, 3*(4), 102–109.

Krokos, E., Plaisant, C., & Varshney, A. (2019). Correction to: Virtual memory palaces: Immersion aids recall. *Virtual Reality, 23*(1), 17–17. https://doi.org/10.1007/s10055-018-0360-5

Lai, J. W. M., De Nobile, J., Bower, M., & Breyer, Y. (2022). Comprehensive evaluation of the use of technology in education – Validation with a cohort of global open online learners. *Education and Information Technologies., 27*, 9877. https://doi.org/10.1007/s10639-022-10986-w

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2010). The sequence read archive. *Nucleic Acids Research, 39*(suppl_1), D19–D21.

Liguori, E., & Winkler, C. (2020). From offline to online: Challenges and opportunities for entrepreneurship education following the COVID-19 pandemic. *Entrepreneurship Education and Pedagogy, 3*(4), 346–351. https://doi.org/10.1177/2515127420916738

Liu, Z., & Han, Z. (2020). Exploring trends of potential user experience of online classroom on virtual platform for higher education during COVID-19 epidemic: A case in China. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. https://doi.org/10.1109/tale48869.2020.9368349

Long, H. M., & Bouck, E. C. (2022). Calculators and online games: Supporting students with learning disabilities in mathematics. *Intervention in School and Clinic, 2022*, 105345122210937. https://doi.org/10.1177/10534512221093787

Massey, C. C., Dogan, S., Muhammad, E., Hogan, E., Head, C., & Kim, J. (2022). The effect of flipped instruction on special education preservice teachers' perceptions. *International Journal for the Scholarship of Teaching and Learning, 16*(2), 10.20429/ijsotl.2022.160213.

Mindster. (2020, December 25). *10 best free educational apps for students & kids learning*. Retrieved from https://mindster.com/free-educational-appsstudents/#Best_10_Free_Educational_Apps_for_Students

Moslehi, P., Rilling, J., & Adams, B. (2022). A user survey on the adoption of crowd-based software engineering instructional screencasts by the new generation of software developers. *Journal of Systems and Software, 185*, 111144. https://doi.org/10.1016/j.jss.2021.111144

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., Niyato, D., & Poor, H. V. (2022). 6G Internet of Things: A comprehensive survey. *IEEE Internet of Things Journal, 9*(1), 359–383. https://doi.org/10.1109/JIOT.2021.3103320

Paul, J., & Jefferson, F. (2019). *A comparative analysis of student performance in an online vs. face- to-face environmental science course from 2009 to 2016* (p. 1). Frontiers in Computer Science. https://doi.org/10.3389/fcomp.2019.00007

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education, 147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Raja, R., & Nagasubramani, P. C. (2018). Impact of modern technology in education. *Journal of Applied and Advanced Research*, S33–S35. https://doi.org/10.21839/jaar.2018.v3is1.165

Razeeth, M. S. S., Kariapper, R. K. A. R., Pirapuraj, P., Nafrees, A. C. M., Rishan, U. M., & Nusrath Ali, S. (2019). E-learning at home vs traditional learning among higher education students: A survey- based analysis. In *9th International Symposium 2019*. Retrieved from http://ir.lib.seu.ac.lk/handle/123456789/3909

Richardson, J. C., & Swan, K. (2003). Examining social presence in online courses in relation to student's perceived learning and satisfaction. *Journal of Asynchronous Learning Networks, 7*, 68–88.

Rmenardi. (2012). *More smart phones than desktop computers? It's time*. Retrieved from http://torsionmobile.com/2012/01/26/more-smart-phones-than-desktop-computers-its-time/

Rogers, S. (2019). *Virtual reality: The learning aid of the 21st century*. Retrieved from https://www.forbes.com/sites/solrogers/2019/03/15/virtual-reality-the-learning-aid-of-the-21st-century/?sh=39d4bda0139b

Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. *Online Journal of Distance Learning Administration, 7*(2), 1–10.

Sattarov, A., & Khaitova, N. (2019). Mobile learning as new forms and methods of increasing the effectiveness of education. *European Journal of Research and Reflection in Educational Sciences, 7*(12), 1169–1175.

Schindler, L. A., Burkholder, G. J., Morad, O. A., & Marsh, C. (2017). Computer-based technology and student engagement: A critical review of the literature. *International Journal of Educational Technology in Higher Education, 14*(1), 1–28. https://doi.org/10.1186/s41239-017-0063-0

Scholz, K. W., Komornicka, J. N., & Moore, A. (2021). Gamifying history: Designing and implementing a game-based learning course design framework. *Teaching & Learning Inquiry, 9*(1), 99–116. https://doi.org/10.20343/teachlearninqu.9.1.9

Senior, J. (2022). Vocabulary taught via Mobile application gamification: Receptive, productive and Long-term usability of words taught using Quizlet and Quizlet live. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICBATS54253.2022.9759019

Shahzad, A., Hassan, R., Aremu, A. Y., Hussain, A., & Lodhi, R. N. (2020). Effects of COVID-19 in E-learning on higher education institution students: The group comparison between male and female. *Quality & Quantity, 55*(3), 805–826. https://doi.org/10.1007/s11135-020-01028-z

Sharples, M. (2007). *Big issues in Mobile learning: Report of a workshop by the kaleidoscope network of excellence mobile learning initiative*. Learning Sciences Research Institute, University of Nottingham.

Shi, Y., & Lin, X. (2021). Exploring the characteristics of adults' online learning activities: A case study of EdX online institute. *Research in Learning Technology, 29*, 1–13. https://doi.org/10.25304/rlt.v29.2622

Shyshkanova, G., Zaytseva, T., & Frydman, O. (2017). Mobile technologies make education a part of everyday life. *Information and Learning Science, 118*(11/12), 570–582. https://doi.org/10.1108/ils-03-2017-0019

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. *Computers in Human Behavior, 87*, 192–206.

Subramaniam, S. R., & Muniandy, B. (2019). The effect of flipped classroom on students' engagement. *Technology, Knowledge and Learning, 24*(3), 355–372. https://doi.org/10.1007/s10758-017-9343-y

Sun, Z., & Qiu, X. (2017). Developing a blended learning model in an EFL class. *International Journal of Continuing Engineering Education and Life-Long Learning, 27*(1/2), 4. https://doi.org/10.1504/ijceell.2017.080998

Tarteer, S., Badah, A., & Khlaif, Z. N. (2021). Employing Google classroom to teach female students during the COVID-19 pandemic. *Computers in the Schools, 38*(4), 300–321. https://doi.org/10.1080/07380569.2021.1988318

Traxler, J. (2005). *Defining mobile learning. IADIS International Conference Mobile Learning*. Retrieved from http://iadis.net/dl/final_uploads/200506C018.pdf

Troussas, C., Krouska, A., & Sgouropoulou, C. (2020). Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education. *Computers & Education, 144*, 103698. https://doi.org/10.1016/j.compedu.2019.103698

Turkanovic, M., Holbl, M., Kosic, K., Hericko, M., & Kamisalic, A. (2018). EduCTX: A blockchain-based higher education credit platform. *IEEE Access, 6*, 5112–5127. https://doi.org/10.1109/access.2018.2789929

Ustunel, H. H., & Tokel, S. T. (2018). Distributed scaffolding: Synergy in technology-enhanced learning environments. *Technology, Knowledge and Learning, 23*(1), 129–160. https://doi.org/10.1007/s10758-017-9299-y

Van Laer, S., & Elen, J. (2020). Adults' self-regulatory behaviour profiles in blended learning environments and their implications for design. *Technology, Knowledge and Learning, 25*(3), 509–539. https://doi.org/10.1007/s10758-017-9351-y

Vijayakumar, S. (2022). Foreign language learning then, now and after COVID-19: An exploration of digital tools to augment the receptive and productive skills of language learners. In A. Hamdan, A. E. Hassanien, T. Mescon, & B. Alareeni (Eds.), *Technologies, artificial intelligence and the future of learning post-COVID-19* (pp. 283–302). Springer. https://doi.org/10.1007/978-3-030-93921-2_17

Vujić, A., & Szabo, A. (2022). Hedonic use, stress, and life satisfaction as predictors of smartphone addiction. *Addictive Behaviors Reports, 15*, 100411. https://doi.org/10.1016/j.abrep.2022.100411

Yang, F., Qiao, Y., Abedin, M. Z., & Huang, C. (2022). Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Transactions on Industrial Informatics*, 1–1. https://doi.org/10.1109/TII.2022.3151917

Yeou, M. (2016). An investigation of students' acceptance of Moodle in a blended learning setting using technology acceptance model. *Journal of Educational Technology Systems, 44*(3), 300–318. https://doi.org/10.1177/0047239515618464

Yip, K. H., Lo, P., Ho, K. K., & Chiu, D. K. (2020). Adoption of mobile library apps as learning tools in higher education: A tale between Hong Kong and Japan. *Online Information Review, 45*(2), 389–405. https://doi.org/10.1108/oir-07-2020-0287

# Knowledge Mining from Health Data: Application of Feature Selection Approaches

**Md. Rabiul Auwul, Md. Ajijul Hakim, Fahmida Tasnim Dhonno, Nusrat Afrin Shilpa, and Mohammad Zoynul Abedin**

**Abstract** This paper aims to measure the performance of feature selection approaches for mining knowledge from health datasets. We compare seven popular knowledge mining approaches, namely relaxed Lasso, random forest, ReliefF, OneR, information gain, *T*-test, and Chi-squared test. The support vector machine (SVM) classifier applies to determine the accuracy and area under the curve (AUC) values of the knowledge miners. We use six popular Affymetrix and cDNA datasets. The results reveal that the relaxed lasso works well with Affymetrix, and the relaxed Lasso with random forest approaches perform well with the cDNA datasets. This paper will enrich the existing literature and assist to identify the best feature for knowledge mining in the health informatics domain.

**Keywords** Knowledge mining · Feature selection · Classification · Cancer data · SVM · Affymetrix · cDNA datasets

M. R. Auwul
Department of Mathematics, Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh

M. A. Hakim
Foreign Exchange and Remittance Department, Travelex Qatar, Golbex Business Center, Doha, Qatar

F. T. Dhonno · N. A. Shilpa
Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

M. Z. Abedin (✉)
Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley, UK
e-mail: m.abedin@tees.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_13

# 1 Introduction

Knowledge mining (data mining) is the investigation and exploration of a large amount of data to find significant patterns and directions. Extracting knowledge from the existing information is the principal job of knowledge mining. Knowledge mining depends on two parameters; one of them is the association rule that produced by scrutinizing data for regular patterns, and then discovering the most significant associations within the data by using the support and confidence criteria. The second parameter includes Classification, Sequence or Path Analysis, Clustering, and Forecasting. In the modern era, the rapid growth of data in every field is quickly mounting with time, as is the importance of knowledge mining (Lu et al., 2022).

The health sector is one of the fast developing and challenging sections in data mining, and it is becoming popular day by day. All the parties involved in health section are greatly benefited through data mining like the healthcare insurers identify fraud and misuse; healthcare groups create client relationship management decisions, and patients get improved and more affordable healthcare facilities (Koh &Tan, 2005). The high-dimensional microarray data used in health research are mostly constructed from two vital chips: Affymetrix and cDNA that are too complex and big to be handled and investigated by classical approaches. Knowledge mining offers the procedure and equipment to make over these large volumes of data into beneficial material for decision making.

These microarray data formed as an array with relevant and redundant features and samples. Feature selection is an important part to select a subset of relevant features to build effective prediction models, especially for classification purposes. To select the significant features from high-dimensional data, there exist several works in literature. But no one can exactly show the application of feature selection methods differentially in the fields of Affymetrix and cDNA microarray data (Rahman et al., 2021).

Feature selection is more essential for high-dimensional data to improve the performance of prediction models by eliminating irrelevant and redundant features in cancer research mostly in gene expressed data that are provided mostly from DNA microarray technology. The problems come in front of researchers when these high-dimensional gene expressed data contain huge number of genes but a few number of samples. The importance of feature selection helps to remove irrelevant genes in high-dimensional data with small sample to prevent declining the classification accuracy from the influence of redundant genes. In this paper, we explore the seven most important feature selection approaches, namely, Relaxed Lasso, Random Forest, ReliefF, OneR, Information Gain, *T*-test, and Chi-squared test on six popular Affymetrix and cDNA cancer gene expressed datasets.

The results obtained from the analysis of seven popular feature selection approaches reveal that Relaxed Lasso works well with Affymetrix and Relaxed Lasso and Random Forest methods work well with cDNA datasets compared to other methods. This study provides outlines of applied assessment to access the results of feature selection in gene expression cancer datasets.

In the healthcare section through the findings of our paper, feature selection methods will be more effective in areas such as predictive medicine, recognition of fraud and misapplication, consumer relationship administration, controlling of healthcare and measuring the effectiveness of definite treatments, as well as used to reduce costs by growing efficiencies, progress patient superiority of life, and possibly, utmost notably, protect the lives of more patients. For academia, researchers can easily find the best feature selection approaches for Affymetrix and cDNA data when they work with knowledge mining such as clustering, classification, etc., and will contribute to the health section.

The rest of this paper is organized as follows. Section 2 delivers a brief review of the literature. Section 3 presents a brief description of the methods and feature selection approaches used in this paper. Section 4 describes the results and discussion. Finally, Sect. 5 concludes the paper.

## 2 Related Works

To reduce the dimensionality and select relevant genes, various features/gene selection approaches existed. Filters, wrappers, and embedded methods are three categories of feature selection approaches. The important features are selected by measuring the correlation between individual features and output class labels, without involving any learning algorithm through filter methods such as ReliefF (Kira & Rendell, 1992), Information Gain (Dagliyan et al., 2011), *T*-test (Abedin et al., 2018), and Chi-squared test (Guotai et al., 2017). Through wrapper methods, a subset of features evaluated by machine learning algorithm employs a search approach to look through the space of potential feature subsets, evaluating each subset based on the quality of the performance of a given algorithm. The sequential feature selection method such as forward/backward selection is an example of wrapper method that is also known as greedy method for its searching strategy. Wrapper methods are more complex and expensive than simpler filter methods. Through embedded methods a penalty term is added against complexity to reduce the degree of over fitting or variance of a model by adding more bias such as L1 (or Lasso) regression for generalized linear regression (Tibshirani, 1996), relaxed lasso (Abedin et al., 2019). The embedded methods are usually faster than the wrapper methods and able to provide a suitable feature subset for the learning algorithm.

The correlation-based feature selection approach is used by Harb and Desuky to develop the classification of health datasets (Harb & Desuky, 2014). Jović et al. (2015) reviewed several filter, wrapper, and embedded feature selection methods with their application. They showed the best for text mining, image processing, computer vision, and industrial application. The Lasso feature selection approach with information gain has been compared by Kamkar et al. to build clinical prediction models (Guo et al., 2015). Lasso and ridge regression are being compared by Fonti and Belitser to implement feature selection on high-dimensional datasets (Fonti & Belitser, 2017). B. Remeseiro and V. Bolon-Canedo (2019) reviewed six state-of-the-art algorithms: CFS, INTERACT, InfoGain, CFS, ReliefF, and SVM-RFE for medical application in terms of four classification algorithms, namely: Naive Bayes, SVM, C4.5, and K-NN. They showed that the classifier performance improved with significant selected features. Chuanze Kang et al. (2019) showed the effect of feature gene selection ReliefF, Relaxed Lasso, Information gain, and Kruskal–Wallis rank sum test for eight microarray data with several classifiers. Relaxed Lasso gave better results for all microarray datasets. ShrutiKaushik et al. compared the traditional feature selection approaches on a healthcare dataset for classification purposes involving several attributes (Kaushik et al., 2019).

The above literature has shown the application of feature selection approaches on healthcare data, but no one has analyzed them on Affymetrix and cDNA microarray data. In this paper, we compare seven popular feature selection approaches, namely: Relaxed Lasso, Random Forest, ReliefF, OneR, Information Gain, *T*-test, and Chi-squared test on six popular Affymetrix and cDNA datasets.

# 3 Material and Methods

## *3.1 Datasets*

In this paper, we used three Affymetrix and three cDNA datasets to evaluate the performance of feature selection approaches. These datasets have been used in many other research papers, among which we will mention only a few examples. Datasets of CNS, Lung DLBCL have been used to analyze the impact of selecting significant features on the classification performance by Chuanze Kang et al. (2019). A comparative study of clustering algorithms for several cancer gene expression data like Shipp, Alizada, Bittner, and Chen datasets is used by Marcilio CP de Souto et al. (2008) (Table 1).

**Table 1** Affymetrix and cDNA datasets used in this paper

| Dataset | Chip | #Sample | Dist. Classes | #Genes |
|---|---|---|---|---|
| CNS (Pomeroy et al., 2002) | Affy | 60 | 21,39 | 7129 |
| Lung (Beer et al., 2002) | Affy | 86 | 62,24 | 7129 |
| DLBCL/Shipp (Shipp et al., 2002) | Affy | 77 | 58,19 | 7129 |
| Alizadeh-V1 (Alizadeh et al., 2000) | cDNA | 42 | 21,21 | 4022 |
| Bittner (Bittner et al., 2000) | cDNA | 38 | 19,19 | 8067 |
| Chen (Chen et al., 2002) | cDNA | 180 | 104,76 | 22,699 |

## 3.2 Feature Selection Approaches

**Relaxed Lasso**

A generalization method proposed by Meinshausen (2006) as of soft-thresholding and hard-thresholding known as relaxed Lasso is defined as:

$$\hat{\beta}^{\lambda,\varphi} = \underset{\beta}{\operatorname{argmin}}\, n^{-1} \sum \left(X_i - Y_i^T\{\beta.1\rho_\lambda\}\right)^2 + \varphi\lambda\|\beta\|_1, \tag{1}$$

for $\lambda \in [0, \infty)$ and $\varphi \in (0, 1]$. The indicator functions on the set of variables $\rho_\lambda \subseteq \{1, \ldots, p\}$ noted as $1\rho_\lambda$, $\forall k \in \{1, \ldots, p\}$:

$$\beta.1\rho_\lambda = \begin{cases} 0, \; k \notin \rho_\lambda \\ \beta_k, \; k \in \rho_\lambda \end{cases}. \tag{2}$$

The predictor variables in the set $\rho_\lambda$ are measured for the relaxed Lasso estimator. For the variable selection part, the parameter $\lambda$ controls in ordinary Lasso estimation. The shrinkage of the coefficients is controlled by the relaxation parameter $\varphi$. For example, for $\varphi = 1$, the relaxed Lasso estimators tend to Lasso estimators. For $\varphi < 1$, the shrinkage of relaxed Lasso is reduced parallel to ordinary Lasso estimation. The above definition would produce a decadent solution in the case of $\varphi = 0$. Accordingly, it minizes the limitation of the relaxed Lasso for $\varphi = 0$ of the above definition for $\varphi \to 0$. All the coefficients in the model $\rho_\lambda$ are estimated by the OLS-solution.

*Step 1*: Compute all ordinary Lasso solutions, e.g., with the Lars-algorithm in Efron et al. (2004) under the Lasso modification. Let $\rho_1, \ldots, \rho_c$ be the resulting set of s models. Let $\lambda_1 > \ldots > \lambda_c = 0$ be a sequence of penalty terms so that $\rho_\lambda = \rho_k$ iff, $\lambda \in (\lambda_k, \lambda_{k\text{-}1}]$.

*Step 2*: Let $g(k) = \left(\hat{\beta}^{\lambda_k} - \hat{\beta}^{\lambda_{k-1}}\right)/(\lambda_{k-1} - \lambda_k)$ for each $k = 1, \ldots, c$. Through this direction, ordinary Lasso solutions can be estimated. Let $\hat{\beta} = \hat{\beta}^{\lambda_k} + \lambda_k g(k)$. If there is at least one component $l$ so that $sign\left(\hat{\beta}_l\right) \neq sign\left(\hat{\beta}_l^{\lambda_k}\right)$, then relaxed Lasso solutions for $\lambda \in \Lambda_k$ have to be computed as in Step 2 of the simple algorithm. Otherwise, all relaxed Lasso solutions for $\lambda \in \Lambda_k$ and $\varphi \in [0, 1]$ are given by linear interpolation between $\hat{\beta}^{\lambda_{k-1}}$.

Let $Y \sim N(0, \Sigma)$, then the response variable can be written by the following linear combination:

$$X = Y^T\beta + \varepsilon, \tag{3}$$

where $\varepsilon \sim N(0, \sigma^2)$, the loss function of relaxed Lasso under parameter $\lambda$ and $\varphi$ is defined as:

$$L(\lambda, \varphi) = E\left(X - Y^T \hat{\beta}^{\lambda,\varphi}\right)^2 - \sigma^2. \quad (4)$$

For sporadic high-dimensional data, a relaxed Lasso is more appropriate.

**Random Forest**

Random forest (RF) is an embedded feature selection approach proposed by Breiman (2001) that generates numerous decision trees based on averaging random selection of response variables of training set. The importance of a variable in a data set $Z_n = \{(a_j, b_j)\}, j = 1, 2, , \ldots n$ is measured by fitting a random forest to the data and the error for each data point is calculated and averaged over the forest. The importance score for the $j^{\text{th}}$ feature is computed by averaging the difference in error before and after the permutation for all the trees. Select those features that produce larger values for this score.

**ReliefF**

An extension version of Relief (Kira & Rendell, 1992) that randomly procures a sample S each time from training samples is known as ReliefF (Robnik-Sikonja & Kononenko, 2003). The weight values are computed and updated by findings k nearest neighbor samples from samples of the same class as $S$ and samples of different class from $S$, respectively, as follows:

$$W_Z = W_Z - \sum\nolimits_{i=1}^{k} diff(Z, S, H)/nk$$
$$+ \sum\nolimits_{B \notin class(S)} \left[ {}^{P(B)}\!/\!{}_{P(class(S)} \times \sum\nolimits_{i=1}^{k} diff(Z, S, N_i(B))\right]/nk. \quad (5)$$

The $i$th nearest neighbor sample in class $B$ is denoted as $N_i(B)$ and diff($g$, $t_1$, $t_2$) denotes the difference between sample $t_1$ and sample $t_2$ in the feature $g$. The formula for diff($g$, $t_1$, $t_2$) if $g$ is discrete is the following:

$$\text{diff}(g, t_1, t_2) = \begin{cases} 0, & t_1[g] = t_2[g] \\ 1, & t_1[g] \neq t_2[g] \end{cases}. \quad (6)$$

The formula for diff($g$, $t_1$, $t_2$) if $g$ is continuous is:

$$\text{diff}(g, t_1, t_2) = {}^{|t_1[g] - t_2[g]|}\!/\!{}_{\max(g) - \min(g)}. \quad (7)$$

The feature with high correlation with the class gives the highest weight, and the features are selected according to the orderly weights (Kang et al., 2019).

**Information Gain**

An entropy-based feature selection method computes the mutual information for each attribute and class and then yields an ordered ranking of all of the features known by information gain (IG). If $X$ and $Y$ are the features and $p(x)$ is the marginal probability density function, then the entropy of given dataset is equated as:

$$H(X) = -\sum_{x \in X} p(x) \, log_2[p(x)]. \tag{8}$$

The conditional entropy of $X$ is given that $Y$ is observed before with the conditional probability $p(x|y)$,

$$H(X|Y) = -\sum_{x \in X} p(x) \sum_{x \in X} p(x|y) \, log_2[p(x|y)]. \tag{9}$$

Finally, the information gain metric is:

$$\mathrm{IG} = H(X) - H(X|Y). \tag{10}$$

Features are ranked according to the IG value. Whose IG value is greater are more important features than others (Dagliyan et al., 2011).

**OneR**

Rule-based embedded feature selection methods construct one rule in training data for each attribute and select rule with smallest error and so that the accuracy could be optimized (Holte, 1993). The features are selected according to the ordered accuracy to the corresponding rules. It follows a decision tree approach. For example, if $R = (x, y)$ is a classification rule with precondition $x$ that executes a sequence of tests that can be estimated as true or false and $y$ is a class that can be suitable to occurrences enclosed by rule $R$. For OneR, a one-level decision tree constructs and tests an individual attribute at a time and branches for every value of that attribute.

**T-Test**

To test the independence of two features, the $T$-test proposed by Gosset is used to quantify the significance of each single feature by determining the following $t$-statistic with respect to the class:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{2/n}}, \tag{11}$$

where $s_p = \sqrt{\left(s_{y_1}^2 + s_{y_2}^2\right)/2}$ for $n = n_1 + n_2$, $s_{y_1}^2$ and $s_{y_2}^2$ are the unbiased estimators of the variances of the two samples. The $p$-value based on these $t$ scores then computed, and based on these $p$-values (the smaller the $p$-value, the more important the feature), the important features are selected.

**Chi-Squared Test**

To test the independence of two features, Chi-squared ($\chi^2$) is used that quantifies the significance of each single feature by determining the following Chi-squared statistic with respect to the class:

$$\chi_d^2 = \sum \frac{(\mathrm{Obs} - \mathrm{Exp})^2}{\mathrm{Exp}}, \tag{12}$$

where Obs are the observed values, Exp are the expected values, and $d$ are the degrees of freedom. The aim of every feature selection method is to select those features that are highly dependent on the response. The larger the Chi-squared value means that the observed values are close to the expected values, the higher the importance of that feature. This method gives misleading results for small frequencies (especially $<5$).

**Classifier Application**

Classification is a popular data mining process for classifying test data based on training data. For finding the accuracy of feature selection methods, we applied an SVM classifier (Boser et al. 1992) with ten-fold cross-validation. The standardization of each feature was also used, which reduces the learning time and equalizes the impact of each predictor on the target variable. SVM is used to find the hyperplane that separates two different sets of features with the maximum distance of the hyperplane to the nearest feature from both sets.

The linear SVM formula is as follows:

$$S = \bar{w}.\bar{y} - b. \tag{13}$$

Here, for the hyperplane, $y$ is the input vector and $w$ is the normal vector with the following distance:

$$d = {}^{1}\!/_{\|w\|_2}. \tag{14}$$

If $y_j$ is the $j$th training sample and $z_j$ is the correct output of the SVM for the $j$th training sample, then the maximum distance $d$ can be expressed as:

$$\min_{w,b} \left[\frac{1}{2}\|w\|^2\right] \text{subject to } z_j\left(\overline{w} \cdot \overline{y}_j - b\right) \geq 1. \tag{15}$$

For the positive and negative samples, $z_j$ is +1 and $-1$, respectively.

**Performance Analysis**

In order to assess the performance of different feature selection methods, we calculate the area under the receiving operating characteristics curve (AUC) and accuracy of each method, Accuracy = (TP + TN)/(TP + FP + TN + FN), where TP, TN, FP, and FN denote the number of true positive, number of true negative, number of false positive, and number of false negative, respectively. Based on these two parameters, we declare a method as a good performer if it produces larger values of Accuracy and AUC values (Fig. 1).

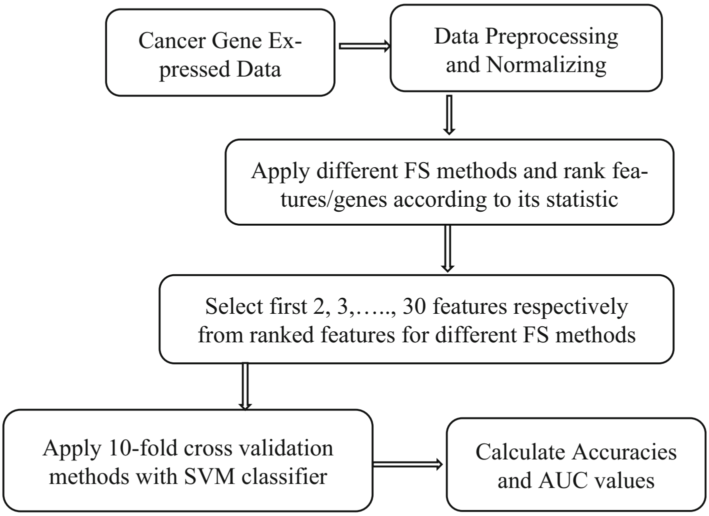



**Fig. 1** Flowchart of the methodology of this paper

## 4 Results and Discussion

Chuanze Kang et al. (2019) specified that the classification accuracy reaches the highest value when the number of feature genes is below 30 for all datasets. Hence, the performance of feature selection methods compared in this paper within the domain of 2 to 30 numbers of feature.

Figure 2 shows the association between the number of feature genes (NF) in the range of 2–30 and the classification accuracy (ACC) for three Affymetrix data. Figure 2a shows the performance of CNS data, Fig. 2b for Lung data, and Fig. 2c for Shipp data. For the Shipp dataset and the CNS dataset, the accuracies touch almost 100%, whereas other approaches have more variations in accuracies for 2–30 features, and there is no development with the increasing of NF. Figure 2b shows that the Relaxed Lasso has the highest accuracy for Lung datasets. When NF is larger than 13, the features selected by the other three methods do not hold the resultant in variations of accuracy, except the *T*-test and random forest with the classification. The other methods probably select redundant genes foremost to decrease the accuracy with increasing the NF. This figure shows that the Relaxed Lasso has the highest accuracies. Hence, we may conclude that Relaxed Lasso achieves better and is more appropriate for feature selection of high-dimensional and small-sample Affymetrix data.

Figure 3 shows the association between the number of feature genes (NF) in the range of 2–30 and the classification accuracy (ACC) for three cDNA data. Figure 3a shows the performance for Bittner data, Fig. 3b is for Alizada data, and Fig. 3c is for Chen data. Figure 3 shows that the Relaxed Lasso has the highest accuracy. For the Bittner dataset and the Alizada dataset, the ACC is almost 100% for Relaxed Lasso.

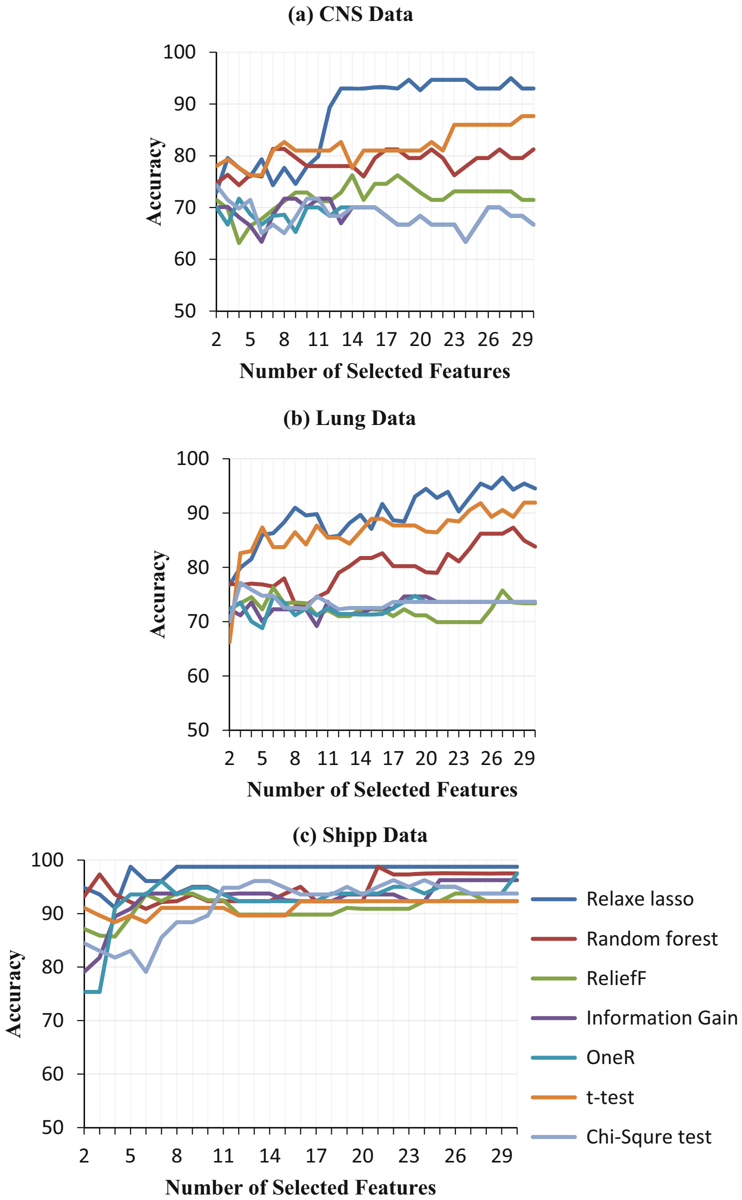



**Fig. 2** Accuracy plot for Affymetrix data

Whereas other methods have more variations in accuracies, there has been no development with increasing of NF. For Chen datasets, Relaxed Lasso has the highest accuracy for more than 13 NF and for less than 13 NF; Random Forest gives a better accuracy than Relaxed Lasso and the others. The feature selected by

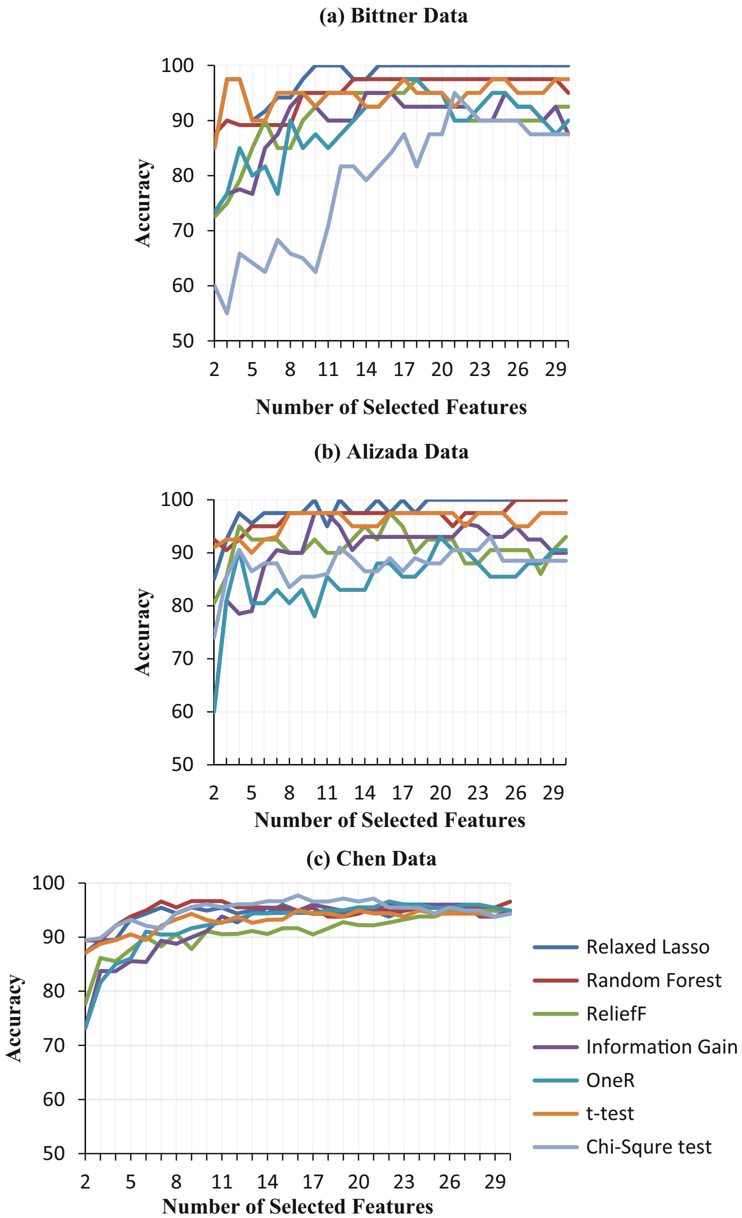



**Fig. 3** Accuracy plot for three cDNA data

methods does not uphold the resultant in variations of accuracy except for Relaxed Lasso and Random Forest with the classification. The other methods probably select redundant genes foremost to decrease accuracy with the increase in NF. Hence,

**Table 2** Performance evaluation of the average AUC values of feature selection methods

| Feature Selection Methods | Affymetrix Datasets | | | cDNA Datasets | | |
|---|---|---|---|---|---|---|
| | CNS | Lung | Shipp | Alizada-V1 | Bittner | Chen |
| Relaxed Lasso | **0.844** | **0.845** | **0.962** | **0.981** | **0.978** | 0.939 |
| Random Forest | 0.733 | 0.703 | 0.929 | 0.970 | 0.951 | **0.945** |
| ReliefF | 0.625 | 0.531 | 0.874 | 0.910 | 0.916 | 0.904 |
| Information gain | 0.596 | 0.527 | 0.890 | 0.903 | 0.900 | 0.919 |
| OneR | 0.590 | 0.532 | 0.878 | 0.851 | 0.894 | 0.923 |
| T-test | 0.795 | 0.847 | 0.923 | 0.958 | 0.945 | 0.928 |
| Chi-Square test | 0.586 | 0.538 | 0.870 | 0.879 | 0.812 | 0.944 |

Bold values indicate the maximum AUC across the datasets and feature selection methods

**Table 3** Efficiency measurement of the feature selection approaches by the Mann-Whitney U test

| Datasets | Random Forest | ReliefF | IG | OneR | T-test | Chi-squared test |
|---|---|---|---|---|---|---|
| | 629 | 817.5 | 841 | 841 | 578.5 | 839.5 |
| CNS | 0.00115 | 6.03E-10 | 5.40E-11 | 4.90E-11 | 0.014 | 6.28E-11 |
| | Yes | Yes | Yes | Yes | Yes | Yes |
| | 777 | 841 | 841 | 841 | 592.5 | 840 |
| Lung | 3.06E-08 | 6.20E-11 | 5.69E-11 | 5.33E-11 | 0.008 | 5.26E-11 |
| | Yes | Yes | Yes | Yes | Yes | Yes |
| | 753 | 823 | 779.5 | 786.5 | 823 | 786.5 |
| Shipp | 6.53E-08 | 8.16E-11 | 6.81E-09 | 3.38E-09 | 5.51E-11 | 3.56E-09 |
| | Yes | Yes | Yes | Yes | Yes | Yes |
| | 577.5 | 795.5 | 783 | 822 | 673 | 813 |
| Alizada-V1 | 0.00875 | 3.19E-09 | 1.03E-08 | 2.52E-10 | 3.17E-05 | 6.42E-10 |
| | Yes | Yes | Yes | Yes | Yes | Yes |
| | 681 | 763 | 771.5 | 772.5 | 698 | 820 |
| Bittner | 2.23E-05 | 5.09E-08 | 2.44E-08 | 2.41E-08 | 8.72E-06 | 2.49E-10 |
| | Yes | Yes | Yes | Yes | Yes | Yes |
| | 297 | 700 | 464.5 | 446 | 647.5 | 285.5 |
| Chen | 0.05517 | 1.40E-05 | 0.498 | 0.697 | 4E-04 | 0.03629 |
| | No | Yes | No | No | Yes | Yes |

N.B: The first, second, and third rows of each dataset are the Mann-Whitney U test score, *p*-Values, and the statement on the average efficiency of Relaxed Lasso greater (Yes) or not (No)

Relaxed Lasso and Random Forest achieve better results and are more appropriate for feature selection of high-dimensional and small-sample cDNA data. Table 2 shows the average AUC values of the feature selection methods. The relaxed lasso gives the maximum values of AUC on an average: for CNS, Lung, Ship, Alizada-V1 andspiepr and Fig. 3(b) is for Alizada data, and Fig. 3c is for Chen data. Table 3 shows the Mann-Whitney U test score and their corresponding *p*-values for six datasets. The results indicate that the Relaxed Lasso feature selection methods are more efficient than the other six algorithms for the CNS, Lung, Shipp, alizada-V1, and Bittner datasets and for the Chen dataset Relaxed Lasso performed better than ReliefF and *T*-test. The efficiency of the Random Forest methods is comparatively higher than that of the others except for Relaxed Lasso for the first five datasets, and for Chen datasets its performance is better than Relaxed Lasso.

## 5 Conclusion

To investigate and explore a large amount of existing information, knowledge mining plays a significant role in the health sector. The findings indicate that knowledge mining is an important and prerequisite part for the stakeholders such as cancer biomarker, genetic pattern for infectious diseases, medicine analytics, and so on.

The superior nature of microarray data is the huge number of genes but small number of samples that generates the prerequisite for important gene selection. To classify large volumes of data, feature selection is a vital issue. There are abundant studies on feature selection to identify cancer classification using microarray gene expression data. But none of these papers include the performance of feature selection approaches in different sections for Affymetrix and cDNA microarray datasets.

This paper has reviewed and analyzed seven popular feature selection approaches, namely: Relaxed Lasso, Random Forest, ReliefF, OneR, Information Gain, *T*-test, and Chi-squared test for cancer classification. A widespread analysis has been conducted and compared these feature selection approaches separately across six Affymetrix and cDNA datasets. The performance evaluation is conducted by finding their accuracy and AUC values with SVM classifier. From our investigation we found that Relaxed Lasso works well with Affymetrix, and Relaxed Lasso and Random Forest approaches work well with cDNA datasets comparatively with other approaches.

Through the findings of our paper in healthcare sector, feature selection approaches will be more effective in areas such as finding biomarker cancer gene, predictive medicine for infectious diseases such as COVID-19, reduction of medical costs by increasing the efficiency of methods, progressing patient superiority of life, and possibly most importantly, protecting the lives of more patients by using clustering, classification, pattern recognition, and other knowledge mining approaches. In the academia sector, researchers can easily find the best feature selection approaches for Affymetrix and cDNA data when they work with knowledge mining approaches and will contribute to the health section.

Regarding future research, we will explore the performance of these feature selection approaches with big data in deep learning. This will be more reliable, informative, and enrich the existing literature.

## References


Abedin, M. Z., Guotai, C., Colombage, S., & Moula, F. E. (2018). Credit default prediction by using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27.

Abedin, M. Z., Guotai, C., Moula, F. E., Zhang, T., & Hassan, M. K. (2019). An optimized support vector machine intelligent technique using optimized feature selection methods: Evidence from Chinese credit approval data. *Journal of Risk Model Validation, 13*(2), 1–46.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., & Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine, 8*, 816–824.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature, 406*(6795), 536–540.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classes. In *Proceedings of the 5th annual workshop on computational learning theory* (pp. 144–152). Springer.

Breiman, L. (2001). Random forest. *Machine Learning, 45*, 5–32.

Chen, X., Cheung, S., So, S., Fan, S., Barry, C., Higgins, J., et al. (2002). Gene expression patterns in human liver cancers. *Molecular Biology of the Cell, 13*(6), 1929–1939.

Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I. H., & Turkay, M. (2011). Optimization based tumor classification from microarray gene expression data. *PLoS One, 6*(2), e14579.

de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics, 9*, 1–14. https://doi.org/10.1186/1471-2105-9-497

Efron, B., Hastie, T., Johnstone, I. M., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*, 407–499.

Fonti, V., & Belitser, E. (2017). *Feature selection using Lasso*. VU Amsterdam Research Paper in Business Analytics.

Guo, X., Minai, A. A., & Lu, L. J. (2015). Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *Journal of Biomedical Informatics, 27*, 277–290.

Guotai, C., Abedin, M. Z., & Moula, F. E. (2017). Modeling credit approval data with neural networks: An experimental investigation and optimization. *Journal of Business Economics and Management, 18*(2), 224–240.

Harb, H. M., & Desuky, A. S. (2014). Feature selection on classification of medical datasets based on particle swarm optimization. *International Journal of Computer Applications, 104*(5), 14–17.

Holte, R. C. (1993). *Very simple classification rules perform well on most commonly used datasets*. Springer.

Jovi, A., Brki, K., & Bogunovi, N. (2015). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 25–29).

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of Theoretical Biology, 463*, 77–91. https://doi.org/10.1016/j.jtbi.2018.12.010

Kaushik, S., Choudhury, A., & Jatav, A. (2019). Comparative analysis of features selection techniques for classification in healthcare. *MLDM*, 488–472.

Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *AAAI-92 Proceedings* (pp. 129–134).

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management, 19*(2), 64–72.

Lu, Y., Yang, L., Shi, B., Li, J., & Abedin, M. Z. (2022). A novel framework of credit risk feature selection for SMEs during industry 4.0. *Annals of Operations Research*.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research, 7*, 983–999.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovizky, G., & Louis. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature, 415*(436), 436.

Rahman, S., Shahriare, S., Khan, M. I., & Abedin, M. Z. (2021). Risk prediction with machine learning in cesarean section: Optimizing healthcare operational decisions. In *Signal processing techniques for computational health informatics, intelligent systems reference library* (Vol. 192). Springer.

Remeseiro, B., & Bolon-canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine, 112*, 103375.

Robnik-Sikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning, 53*, 23–69.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine, 8*, 68–74.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 58*(1), 267–288.