Alexandria Journal of Accounting Research
Issued by the Accounting Department, Third Issue, September, 2024, Vol. 8

Detecting the Probability of Fraud in Interim Financial Statements Using Machine Learning Models: Do Correlation-Based Analysis and Principal Component Analysis for Dimensionality Reduction Matter?

Dr. Hosam Mohamed Ragab Moubarak
Accounting Department, Faculty of Business,
Alexandria University, Alexandria, Egypt
Accounting and Information Technology Department,
Faculty of International Business and Humanities,
Egypt-Japan University of Science and Technology (EJUST),
New Borg Al-Arab, Alexandria, Egypt
Hosam.moubarak@alexu.edu.eg

Abstract

This study compares different machine learning (ML) models, datasets, and dimensionality reduction techniques to determine their effectiveness in detecting the probability of interim financial statements fraud (FSF). Using the design science research (DSR) approach, the study adopts a quantitative approach with a set of secondary data from the financial reports published by non-financial firms listed on the Egyptian Stock Exchange from 2015 to 2022. The research used a set of financial features compromising ratios reflecting the firm's leverage, profitability, liquidity, and efficiency. Indicators of fraud are based on the Beneish M-score model that demonstrates the possibility of reporting earning manipulations. The findings reveal that the Random Forest classifier outperforms other classifiers, especially with the oversampling dataset after preprocessing using the correlation-based dimensionality reduction method. This study aims to benefit investors, stakeholders, auditors, regulatory bodies, fraud examiners, and academics who pay precise attention to creating new, better methods to detect the probability of FSF. This study introduces novel ML models and dimensionality reduction techniques that have not been previously applied to detect the probability of FSF in an emerging context. The research provides unique insights and evidence on the most effective dimensionality reduction techniques for achieving the best detection results. Additionally, the study introduces innovative solutions to the data imbalance problem. Therefore, the results can enable regulatory bodies and practitioners to detect managerial opportunistic behaviors more accurately in a timely manner and provide a foundation for further academic research in the field.

Key Words: Financial statement fraud, Fraud detection, Machine learning classification algorithms, Design science research, Comparative study, Dimensionality reduction, Imbalanced datasets

Received on 25/6/2024, accepted on 21/7/2024.

كشف احتمالية الغش في القوائم المالية المرحلية باستخدام نماذج التعلم الآلة: هل معامل تحليل الارتباط وتحليل العنصر الرئيسي مهم للحد من الأبعاد؟

ملخص البحث

تهدف تلك الدراسة إلى مقارنة مختلف نماذج التعلم الآلة، مجموعات البيانات وتقنيات الحد من الأبعاد لتحديد فعاليتها في كشف احتمالية وجود الغش في القوائم المالية المرحلية. باستخدام منهجية البحث في علم التصميم design science research، تبنت تلك الدراسة نهجًا كميًا مبني على مجموعة من البيانات الثانوية التي تم تجميها من التقارير المالية المنشورة من قِبل للشركات غير المالية المقيدة في البورصة المصرية من 2015 حتى 2022. استخدم البحث مجموعة من السمات المالية التي تعكس نسب الرفع المالي، الربحية، السيولة و الكفاءة للشركات. كما استندت على نموذج بينيش إم Beneish M-score كمؤشر للغش والذي يعكس بدوره احتمالية وجود تلاعب في التقرير عن أرباح الشركة. أوضحت النتائج أن نموذج الغابات العشوائية Random Forest يتفوق في الأداء على باقي النماذج، خاصةً مع مجموعة الإفراط في أخذ العينات oversampling بعد أن تم معالجته باستخدام معامل تحليل الارتباط correlation-based analysis كطريقة الحد من الأبعاد dimensionality reduction. تهدف هذه الدراسة إلى إفادة المستثمرين، أصحاب المصالح، مراقبي الحسابات، الهيئات التنظيمية، مختبري الغش و الأكاديميين الذين يولون اهتمامًا دقيقًا لإنشاء طرق جديدة أفضل للكشف عن احتمالية الغش في القوائم المالية. تلك الدراسة هي فريدة من نوعها في اقتراحها لنماذج التعلم الآلة التي لم يتم تطبيقها مسبقًا لكشف احتمالية وجود الغش في القوائم المالية في سياق الدول الناشئة. كما تقدم أدلة جديدة عن أكثر تقنيات الحد من الأبعاد فعالية لتحقيق أفضل النتائج. إضافةً إلى ذلك، تبرز الدراسة حلولًا مبتكرة لمشكلة اختلال توازن البيانات data imbalance problem. وبالتالي قد تُمكن تلك النتائج الهيئات التنظيمية والمهنيين للكشف عن السلوكيات الانتهازية للإدارة بشكل أكثر دقة في الوقت المحدد، كما توفر أسسًا لمزيد من الأبحاث الأكاديمية في هذا المجال.

الكلمات المفتاحية: الغش في القوائم المالية، كشف الغش، نماذج التصنيف للتعلم الآلة، البحث في علم التصميم، دراسة مقارنة، الحد من الأبعاد، مشكلة اختلال توازن البيانات.

1- Introduction

Financial statements are essential for communicating valuable qualitative and quantitative information to estimate the firm's value and evaluate its stock prices (El-Diftar & Elkalla, 2019). To ensure effective decision-making, financial statements must be free from errors, relevant, and faithfully represented. However, in light of the exponential growth in the global business markets and fierce competition over the last decade, managers find refuge in intentionally manipulating financial reports in different ways, such as overstating revenue, profit, and assets or understating expenses, losses, and liabilities to meet the analysts' and public expectations or to fulfill personal objectives. Knowing the various internal and external fraud categorizations denoted by the Association of Certified Fraud Examiners (ACFE) (2020) and PwC (2020), this study emphasizes internal committed fraud, more specifically 'occupational fraud', with a focus on financial statements fraud (FSF) regardless of other occupational fraud schemes, including asset misappropriation and corruption.

The National Commission on Fraudulent Financial Reporting first defined FSF, or fraudulent financial reporting (FFR), in 1987 as the intentional misrepresentation or omission that leads to materially misleading financial statements. FSF implies that the firm shows falsified information to conceal its unhealthy status and poor performance before receiving public supervisors' forewarning notices (Chung et al., 2014). However, as a matter of fact, FSF causes significant losses to all stakeholders (Mandal & S, 2023), as investors find no optimal returns on their investments, creditors suffer getting their payments, employees lose their jobs, the accounting profession loses its credibility, and the public loses confidence in the financial statements, thereby leaving the firm with heavy financial losses and litigation costs that may lead to bankruptcy declaration (Rezaee, 2002). According to the ACFE estimations in the 2024 Fraud Report to Nations, 1,921 occupational fraud cases investigated between January 2022 and September 2023 in 138 worldwide countries caused losses of more than US$3.1 billion, which counts about five percent of the firms' annual revenue, in which FSF shows the greatest median loss per case (ACFE, 2024).

Moreover, occupational fraud is complex and often difficult to identify until it is too late (Vousinas, 2019). Unlike other crimes, financial fraud is hard to prove in a court of law using scientific evidence such as fingerprints or DNA (Omar et al., 2017). ACFE (2012) reported that financial fraud detection usually takes three to six years; by then, any related evidence may have been tampered with or destroyed. Knowing these devastating consequences, FSF has become a primary concern that haunts each stakeholder's mind, especially after the enormous financial scandals provoked by the commission of FSF by Enron, WorldCom, Lehman Brothers, etc. As such, researchers studied the motivations, causes, effects, and factors related to FSF to assist auditors, as the first line of defense, in timely detecting and preventing the spread of fraud activities. Various models, including financial statements analysis, financial ratios, trend analysis, the Beneish M-score model, and the Altman Z-score, have been used to detect the probability of FSF (Kukreja et al., 2020).

However, in the era of the rapid change in business markets and the extensive increase in multiple-sourced complex data, manual FSF detection seems nearly impossible. Therefore, to overcome these challenges, researchers have lately suggested using data mining techniques to detect or even predict the probability of FSF early. Data mining is a process tool that extracts knowledge and patterns from massive data using sophisticated search capabilities and advanced statistical algorithms (Witten et al., 2017). Algorithms, such as Logistic Regression, Decision Trees, Bayesian Belief Networks, and Artificial Neural Networks, have proved their efficiency in accurately detecting FSF (Omar et al., 2017; Riskiyadi, 2023). However, prior studies have produced varying results regarding the best data mining algorithm for detecting FSF. These discrepancies are probably due to the influence of specific cultural characteristics (Darsono et al., 2021), and corruption levels (Lakshmi et al., 2021) in the countries where the research samples are selected, demonstrating the ongoing desire for improved FSF intelligent detection methods.

Accordingly, this study holds particular significance as it extends the application of data mining algorithms to detect the probability of FSF in Egyptian non-financial firms listed on the Egyptian Stock Exchange (EGX). This research aims to contribute to the fraud literature in the emerging Egyptian context, leading in the region of the Middle East and North Africa (MENA) by constructing three machine learning (ML) models, including first Logistic Regression (LR), second, Decision Tree (DT), and finally, Random Forest (RF).

Egypt's unique institutional environment is characterized by a concentrated ownership structure, limited investor control, and low market scrutiny (Abozaid et al., 2020), which leaves room for greater opportunities to commit fraud. Besides, according to the Transparency International (2023) corruption perceptions index, Egypt scored a corruption of 35 degrees on a scale that ranges from zero (highly corrupted) to 100 (very clean). In other words, Egypt is ranked among the most corrupt countries worldwide, averaging a score of 32.31 degrees from 1996 until 2023, recording its highest score of 37 degrees in 2014 and lowest record of 28 degrees in 2008 (Trading Economics, 2024). Hence, Egypt marks an ideal setting for such an FSF research scope.

Moreover, this study breaks new ground by comparing the performance of ML classification models in accurately detecting FSF probability in interim financial statements. This is a departure from most previous studies that focused on annual financial statements. Additionally, this research contributes to the application of ML in accounting and auditing disciplines by comparing the performance of ML classifiers with four different data preprocessing techniques: first, without using any dimensionality reduction; second, after applying correlation-based analysis (CBA); third, after applying principal component analysis (PCA) algorithm, and finally, after using both CBA and PCA algorithm as a combined dimensionality reduction technique. The objective behind the proposed research framework is to provide a subset of financial features that ensure highly accurate FSF classification in the original imbalanced and balanced oversampling and undersampling datasets (Haixiang et al., 2017).

The remainder of this research paper is structured in the following manner: Section 2 reviews the relevant FSF prior literature. Section 3 represents the research methodology, including the dataset, financial ratio features, the different data preprocessing techniques, and the classification algorithms used to build the FSF detection models. Section 4 reports the findings and discussions, and finally, Section 5 provides the paper's conclusion with a summary, contributions, limitations, and further research opportunities.

2-Literature Review and Problem Statement

In recent academic literature, there has been a growing recognition of the limitations associated with traditional methods of detecting FSF. Consequently, researchers have significantly shifted their focus toward leveraging ML models to enhance the accuracy of detecting potentially fraudulent reporting within a firm's financial statements. In this regard, this section aims to offer a comprehensive overview of occupational fraud, encompassing its definition and classifications. Furthermore, the literature review will outline and summarize previous literature that has employed traditional detection methods for FSF specifically. Through this exploration, the essential findings and limitations of these conventional approaches will be highlighted, establishing the need for more robust and sophisticated detection methods. Moreover, the study will thoroughly examine the most recent literature that investigates the application of a diverse range of ML models in detecting the probability of FSF. This review will encompass various contexts, shedding light on the effectiveness of utilizing ML models for fraud detection across different economies.

2-1 Occupational Fraud and FSF: An Overview

In light of the International Standard on Auditing (ISA) 240 issued by the International Auditing and Assurance Standards Board (IAASB), fraud is the deliberate deception by one or more managers, those charged with governance, employees, or third parties to obtain an illegal advantage. The ACFE (2024, p.104) referred to the occupational fraud as:

"The use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets."

Occupational fraud encompasses three key schemes: asset misappropriation, corruption, and FSF (ACFE, 2020). FSF, commonly known as 'cooking the books,' involves intentionally manipulating or misrepresenting financial statements to achieve a specific objective. Although FSF is the least common fraud scheme, accounting for only five percent of cases covered in the ACFE 2024 report, it is considered the costliest scheme, with an average loss of US$766,000 per case. This is significantly higher than the median loss of US$200,000 per case of corruption schemes and US$120,000 for the most pervasive asset misappropriation schemes (ACFE, 2024). Disturbingly, FSF has marked a 29 percent increase in median loss per case from 2022 until 2024, according to the ACFE report. This upward trend underscores the persistent and significant threat that the FSF poses, necessitating constant vigilance from all stakeholders.

FSF may be accomplished by any means of unfaithful representation of a firm's performance and financial position in financial statements, including financial misstatements, restatements, disclosure delays, disclosure cancelation, or any other potential unknowns (An & Suh, 2020). First, financial misstatements represent intentional material misstating or omitting reports to deceive stakeholders (Rezaee, 2005). Second, based on the General Accounting Office (GAO) (2002) in the United States, financial restatements are the revision of details disclosed in the previously issued financial statements. Third, disclosure delay is the action of postponing the issuance of financial statements over the due date set by the regulatory authorities (Adams et al., 1995). Finally, disclosure cancellation implies denying or canceling submitted financial statements to avoid discovering falsified information (Chung et al., 2014). To provide a targeted and efficient approach to FSF investigation, this study focuses on the potential of ML models to detect misstatements as a critical category of FSF.

2-2 Detecting the Probability of FSF Using Traditional Techniques

Recognizing the lasting, profound impact of FSF, researchers have diligently explored decision-aided tools to assist auditors, investors, and stakeholders in assessing the likelihood of FSF. Among these, financial ratio analysis stands out as a pivotal technique. This method, which evaluates a firm's performance by analyzing the relationships between financial statements' accounts, has proved its efficiency over time in assessing the likelihood of FSF (Drake & Fabozzi, 2012). The pioneering work of Altman (1968), who developed the 'Z-score' model, is a testament to the dedication and foresight of researchers in the field. The Z-score model is based on multiple discriminant analysis and conditional probability techniques using a set of five financial indicators to identify firms' financial health. Prior research suggested that distressed firms with poor financial conditions are more likely to commit FSF. Thus, these indicators have been extensively associated with detecting FFR, along with the prediction of bankruptcy.

Subsequently, using a probit analysis on a sample of US firms, Beneish (1999) devised the 'M-score' model that evaluates the probability of financial statements being prone to earnings manipulation from one period to another. Beneish consists of a weighted blend of eight financial ratios derived from information reported in financial statements; if the M-score is greater than -2.22, then a disposition to fraud in financial statements is indicated. Walking in the footsteps of Beneish (1999), Dechow et al. (2011) have recently developed the "F-score" fraud risk assessment tool, which determines the probability of misstating the financial figures reported in the financial statements based on ratio analysis and a set of 28 financial and non-financial proxies. If the F-score is more than one, the financial statements reflect the probability of manipulation. Dechow et al. (2011) suggested that off-balance sheet financing engagements, level of accruals, percentage of soft assets, stock performance, and raising finance or issuing additional stocks at the time of misstatement are the characteristics that distinguish fraud-committed firms from other non-fraud committed firms. Table

1 summarizes some of the wide streams of empirical research that applied these financial ratio analysis models in detecting or predicting FSF over several years. Thus, it is organized in terms of research references, data, traditional detection model used, and key findings of each study.

Table 1: Prior literature on the probability of FSF detection using traditional techniques

Researcher(s)	Data (country, period)	Traditional model employed	Key findings
Marais et al. (2023)	South Africa, 2006 - 2018	Beneish M-score and Dechow F-score	Both models failed to show high sensitivity and precision
Aviantara (2023)	Indonesia, 2007 - 2018	Beneish M-score and Dechow F-score	Dechow scores fraud indication by nine times, while Beneish scores fraud indication by eight times
Saleh et al. (2021)	Jordan, 2015 -2019	Altman Z-score and Dechow F-score	The models confirmed the validity and specificity of detecting fraud
Kukreja et al. (2020)	USA, 2012 - 2018	Altman Z-score and Beneish M-score	Altman's Z-score is found to be more predictable in fraud detection compared to Beneish's M-score
MacCarthy (2017)	USA, 1996 - 2000	Altman Z-score and Beneish M-score	The study recommended that both models should be used simultaneously to detect FSF better
Mehta & Bhavani (2017)	Japan, 2008 - 2014	Altman Z-score, Beneish M-score and Benford’s law	Altman Z-score revealed the most accurate results in detecting fraud in published financial statements of Toshiba corporation
Bhavani& Amponsah (2017)	Japan, 2008 - 2014	Altman Z-score and Beneish M-score	Beneish's M-score failed to provide any indication of detection of fraud, while Altman's Z-score provided some indication of manipulation in Toshiba’s published financial statements
Helbig (2016)	Spain, 2009 - 2013	Altman Z-score and Beneish M-score	Both models are recommended applying both models to detect manipulation of financial statements
Anh & Linh (2016)	Vietnam, 2013- 2014	Beneish M-score	Beneish M-score is suggested as one of the most useful techniques in detecting earnings manipulation
Dalnial et al. (2014)	Malaysia, 2000 - 2011	Altman Z-score	The Altman Z-score successfully detected FSF.

Source: Author’s own creation based on the prior literature mentioned

2-3 Detecting the Probability of FSF Using Data Mining Techniques

In recent times, auditors, as the responsible party for detecting FSF, have been grappling with significant challenges arising from the cut-throat competition that emerged due to the drastic changes in the business market during the last decade. The ACFE 2024 report revealed that only a limited number of cases are identified by internal and external auditors, with rates of only 13 percent and 3 percent, respectively. This data underscores the fact that the traditional FSF detection models, under the current circumstances, are not just impractical but also imprecise, time-consuming, and costly (Al-Hashedi & Magalingam, 2021). Hence, like other disciplines, accounting and auditing researchers have shown interest in exploring the potential of data mining and machine learning (ML) in the intelligent detection of the probability of FSF. This interest stems from the belief that these advanced methods could revolutionize auditing by making it more effective and efficient in detecting financial fraud.

Data mining is discovering database knowledge and finding hidden information patterns in the data set. Data mining is based on multiple disciplines, including statistics, artificial intelligence, and ML (Gorunescu, 2011). ML is a part of artificial intelligence, defined as a system that can learn independently and identify patterns from pre-existing data, adapt to new inputs, and create automatic actions without explicit intervention from the user to make appropriate decisions (Riskiyadi, 2023; Witten et al., 2017). ML is classified into three types commonly known as supervised, semi-supervised, and unsupervised (Brownlee, 2020; Zaki & Jr, 2020). In line with the FSF intelligent detection literature summarized in Table 2, this study emphasizes the supervised ML models as it aims to compare multiple ML models' performance in detecting FSF using a dataset of fraudulent and non-fraudulent financial statements.

Table 2 discusses a list of empirical studies that developed and utilized ML models to detect the probability of FSF; it is structured in terms of research references, data, data mining models, and model performance evaluation criteria upon which the best performers are identified. As provided in Table 2, the studies have focused on comparing several ML models and selecting the best performer in detecting or predicting FSF in various contexts by choosing a research sample of fraudulent and non-fraudulent financial reports from a particular country over several periods. The ML models ranged from simple to complex ensemble models. Simple ML models, such as LR, Support Vector Machine, DT, K-Nearest Neighbors, Neural Networks, Bayesian Belief Networks, and Naïve Bayesian, ensure the application of fast and easy computations. However, they fail to solve larger complex data, thereby leading to an immense need for building better models, such as ensemble models or a combination of various classifiers (Riskiyadi, 2023). Ensemble models include Random Forest (RF), RUSBoost, CUSBoost, AdaBoost, XGBoost, Gradient Tree Boosting, Extremely Randomized Trees, and Long Short-Term Memory.

The performance indicators upon which the best ML model performer has been identified among other models are widely diversified over previous studies. Thus, the researchers used multiple indicators at once to identify more effective and efficient measurement indicators for identifying fraudulent reporting (Riskiyadi, 2023). Based on Table 2, among these various indicators, the area under the receiver operating characteristic (ROC) curve (AUC) is one of the most used performance indicators in evaluating the classifiers detecting FFR. AUC is a critical 'single number' performance discriminator widely recognized for its high ability to evaluate the classifiers' performance (Fawcett, 2006), in which a higher AUC indicates a better performance of the ML classification model. Additionally, prior studies used accuracy and error rates and misclassification costs to obtain a complete performance measurement of the models developed. Using these performance indicators, the studies concluded varying results about the best ML classification performer.

Based on the table provided below, there is a notable scarcity of data mining FSF detection research in the MENA region, specifically in Egypt, despite being ranked as a highly corrupt country with an average score of 32.31 degrees over the last 28 years, according to the Transparency International corruption rankings. This study, therefore, makes a unique contribution by addressing this critical research gap in comparing the performance of simple ML models, including LR and DT, with the RF ensemble model in detecting the probability of financial misstatement issues.

Table 2: Prior literature on the probability of FSF detection using data mining techniques

Researcher(s)	Data (country, period, fraud/non-fraud)	Models	Input Features	Model performance evaluation criteria	Best models performers
Rahman & Zhu (2024)	China, 2003 – 2017, 1,921/13,633 (for family firms) and 1,428/16,596 (for non-family firms)	LR, Bagging, RF, RUSBoost, CUSBoost	28 raw financial data	AUC, AUPR and NDCG @k	CUSBoost
Duan et al. (2024)	China, 2007 – 2018, 395/22,976	BRF, RUSBoost, XGBoost, SVM, LSTM and LR	20 financial variables, 13 non-financial variables and 8 textual variables	AUC, sensitivity, precision and F-measure	BRF
Rickyard (2023)	Indonesia, 2010 – 2019, 1,058/2,412	SGD, SVM, KNN, DT, RF, ERT, AdaBoosts, GTB and NN	27 financial variables	Accuracy, specificity, sensitivity, precision, G-mean, F-measure, FNR, FPR, CM and AUC	ERT
Wang et al. (2023)	China, 2014 – 2018, 404/1,666	LR, SVM, Bagging, RF, ANN, ANN-LSTM, RCMA	Financial variables and textual variables	AUC, type I error rate and type II error rate	RCMA
Ali et al. (2023)	MENA region, 2012 – 2019, 102/1798	SVM, DT, LR, RF, AdaBoost, XGBoost	26 financial variables	Accuracy, precision, recall and F-measure	XGB
Xu et al. (2023)	China, 2009 – 2018, 4,440//35,922	RF, GBDT, RUSBoost, LR, SVM and ANN	14 financial variables and 30 non-financial variables	AUC, precision, recall, F-measure, NGCG@k, precision@k and recall@k	RF
Xiuguo & Shengyong (2022)	China, 2016 – 2020, 244/4886	LR, RF, SVM, XGBoost, ANN, CNN, LSTM, GRU and Transformer	58 financial variables, 16 non-financial variables, and textual variables	AUC, sensitivity, specificity, F-measure, and accuracy	LSTM
Jan (2021)	Taiwan, 2001 – 2019, 51/102	RNN, LSTM	14 financial variables and 4 non-financial variables	Accuracy, precision, sensitivity, specificity, F-measure, type I error rate, type II error rate, AUC	LSTM
An & Suh (2020)	Korea, 1996 – 2003, 1,591/31,628	CART, RF, Bagging of DTs, Boosting of DTs, LR, SVM, ANN and Modified RF	23 financial variables	Accuracy, recall, precision, and F-measure	Modified RF
Papík & Papíková (2020)	USA, not identifiable, 47/316	LR and LDA	8 financial ratios from the Beneish model and 28 financial raw data	Accuracy, sensitivity, specificity	LR

Mohammadi et al. (2020)	Iran, 2011 – 2016, 165/165	LR, DA, SVM, ANN and BN	17 financial variables	Accuracy, type I error rate, and type II error rate	ANN
Jan (2018)	Taiwan, 2004 – 2014, 40/120	FS: ANN and SVM with CART, CHAID, C5.0 and Quest	19 financial variables and 3 non-financial variables	Accuracy, type I error rate and type II error rate	ANN+CART
Tang et al. (2018)	USA, 1998 – 2016, 130/130	Knowledge-based system based on C4.5	18 financial variables	Accuracy, recall and F-measure	Knowledge-based system based on C4.5
Hajek & Henriques (2017)	USA, 2005 – 2015, 311/311	LR, NB, BN, DTNB, SVM, JRip, C4.5, CART, LMT, MLP, VP, Bagging, RF, AB	32 financial variables	Accuracy, TP rate, TN rate, F-measure, AUC and MCC	BNN
Dutta et al. (2017)	USA, 2001 – 2014, 3,513/60,720	DT, ANN, NB, SVM and BN	116 financial variables	Sensitivity, FPR, accuracy, precision, F-measure, and AUC	ANN
Kim et al. (2016)	USA, 788/2,156	MLogit, SVM, BN	40 financial variables and 9 non-financial variables	Accuracy, G-mean, % of mis-statements detected, and misclassification costs	MLogit
Lin et al. (2015)	Taiwan, 1998 – 2010, 129/447	LR, ANN, and CART	32 fraud factors	Accuracy, type I error rate, type II error rate, and misclassification costs	ANN
Liu et al. (2015)	China, 1998 – 2014, 138/160	RF, LR, CART, SVM and KNN	29 financial variables	Accuracy, type I error rate, and type II error rate	RF
Song et al. (2014)	China, 2008 – 2012, 110/440	LR, BPNN, C5.0 DT, SVM, and proposed ensemble classifier (voting)	23 financial variables	Accuracy, type I error rate, type II error rate, and AUC	proposed ensemble classifier (voting)
Huang et al. (2014)	Taiwan, 1992 – 2006, 72/72	Dual GHSOM, KNN, BPNN, SVM, SOM+LDA and GHSOM+LDA	24 financial variables	Type I error rate and type II error rate	Dual GHSOM
Chen et al. (2014)	Taiwan, 1998 – 2008, 47/47	RST, C5.0 and BPNN	21 financial variables, and 11 non-financial variables	Accuracy, type 1 error rate, and type II error rate	RST

Note(s): LR: Logistic Regression; RF: Random Forest; BRF: Balanced Random Forest; SVM: Support Vector Machine; LSTM: Long Short Term Memory; SGD: Stochastic Gradient Descent; KNN: K-Nearest Neighbors; DT: Decision Tree; ERT: Extremely Randomized Trees; GTB: Gradient Tree Boosting; NN: Neural Network; ANN: Artificial Neural Network; RCMA: Ratio-aware, Chapter-aware, and Modality-aware Attention Mechanisms; GBDT: Gradient-Boosted Decision Tree; CNN: Convolution Neural Network; GRU: Gated Recurrent Unit; RNN: Recurrent Neural Network; CART: Classification and Regression Tree; LDA: Linear Discriminant Analysis; DA: Discriminant Analysis; BN: Bayesian Network; CHAID: Chi-Square Automatic Interaction Detector; QUEST: Quick Unbiased Efficient Statistical Tree; NB: Naïve Bayes; DTNB: Decision Table/Naïve Bayes; LMT: Logistic Model Trees; MLP: Multilayer Perceptron; VP: Voted Perceptron; MLogit: Multinominal Logistic Regression; BPNN: Back Propagation Neural Network; GHSOM: Growing Hierarchical Self-Organizing Map; SOM: Self-Organizing Maps; RST: Rough Set Theory; AUC: Area Under Receiver Operating Characteristic (ROC) Curve; AUPR: Area Under the Precision-Recall Curve; NDCG@k: Normalized Discounted Cumulative Gain at k; FNR: False Negative Rate; FPR: False Positive Rate; CM: Cost Minimization

Source: Author's own creation based on the prior literature mentioned

3- Research Methodology

The primary objective of this research study is to empirically assess the effectiveness of simple machine learning models as well as ensemble machine learning models in detecting the probability of FSF. The emphasis is placed on the application of three specific models: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). In order to accomplish this research goal, the paper utilizes the design science research (DSR) approach, which has been widely acknowledged and established in the information systems discipline as evidenced by existing literature. DSR is a problem-solving research paradigm that focuses on creating new artifacts to address real-world problems and enhance organizations' knowledge and innovation capabilities (Hevner et al., 2004). It is employed to advance the understanding of information systems and introduces knowledge relevant to constructing database systems, model processing, aligning information systems with business strategy, and employing data analytics for effective decision-making. DSR has been extensively utilized in the information systems literature to contribute to developing knowledge and solutions in these areas (Becker et al., 2015).

Following the groundbreaking research conducted by Mousa et al. (2022), Hevner & Chatterjee (2010), and Horváth (2007), this paper utilizes the robust four-step DSR approach across five distinct stages over four stages, as illustrated in Figure 1. Step one revolves around identifying the problem and setting the solution objectives. This paper effectively addresses the research problem in the introduction, followed by a comprehensive review of pertinent prior studies in the literature review section. The primary focus of this paper revolves around critically examining the efficacy of ML models in accurately detecting the probability of FSF in the non-financial sector within the Egyptian context. In the second step of the process, it focuses on designing the models. This will involve three stages: identifying the research sample, data collection and the input features, determining the label (dependent variable) for the machine learning models, and preprocessing the data. Moving on to step three, or stage four, the development of the ML models. This will include creating LR, DT, and RF models. Finally, step four, conducted in stage five, involves evaluating the machine learning models based on various performance criteria to identify the most effective classifier and apply it across other applications. The following sections will delve into the specifics of each of the five stages of applying DSR.

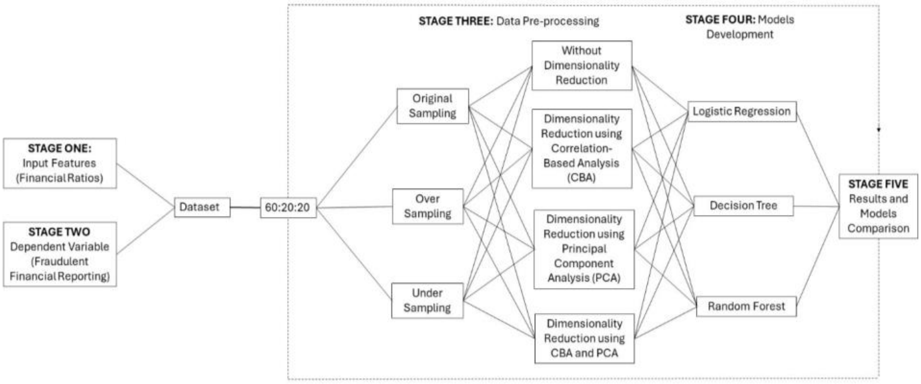

STAGE THREE: Data Pre-processing
STAGE FOUR: Models Development
STAGE ONE: Input Features (Financial Ratios)
STAGE TWO Dependent Variable (Fraudulent Financial Reporting)
Dataset
60:20:20
Original Sampling
Over Sampling
Under Sampling
Without Dimensionality Reduction
Dimensionality Reduction using Correlation-Based Analysis (CBA)
Dimensionality Reduction using Principal Component Analysis (PCA)
Dimensionality Reduction using CBA and PCA
Logistic Regression
Decision Tree
Random Forest
STAGE FIVE Results and Models Comparison

Figure 1: Proposed Research Flow Following the Design Science Research (DSR) Approach

Source: Author’s own creation

3-1 Stage 1: Research Sample, Data and Input Features

The population of this study consists of all firms listed on the Egyptian Stock Exchange (EGX) between 2015 and 2022. The paper used a purposive sampling technique with the criteria of non-financial listed firms on EGX during the research period that issued interim financial reports containing the required variables. The research sample process is demonstrated in Table 3. Years before 2015 were disregarded in the research period to avoid the impact of the turbulent conditions witnessed in Egypt during the June 2013 Egyptian protests, the 2014 official constitutional referendum, and the presidential elections. This study did not include financial institutions and banks listed on the EGX, as they adhere to specific auditing standards. The data used in this research is secondary data from interim financial statements obtained from the Refinitiv Datastream (previously known as the Thomson Reuters Datastream), firms’ official websites, and the Mubasher.info database.

Table 3: Research Sample Process

Description	Firms	Firm-Quarter Year Observations
Firms listed on EGX during the research period	223	7,136
Financial institutions listed on EGX	(47)	(1,504)
Non-financial firms with irrelevant interim financial reports	(67)	(2,144)
Extreme or missing firm-quarter year observations	-	(581)
The final non-financial research sample	109	2,907

Source: Author’s own creation

3-1-1 Input Features

Financial ratios irregularities are regarded as a red flag for FFR (Elsayed 2017). Therefore, this study uses a set of 24 financial ratios that indicate the firm’s leverage, profitability, liquidity, and efficiency levels, which may signal the presence of FFR based on the previous research. Table 4 lists this study's set of financial ratios, along with their calculation formulas and corresponding references.

Table 4: Input Financial Features List

Category/ Symbol	Financial Ratio	Formula	Source
Leverage Ratios			
X1	TL/TA	Total liabilities/total assets	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X2	TL/TE	Total liabilities/total equity	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X3	LTD/TA	Long-term debt/total assets	(Dutta et al. 2017; Riskiyadi 2023)
Profitability Ratios			
X4	NI/ATA	Net income/average total assets	(An & Suh 2020; Jan 2018; Riskiyadi 2023)
X5	RE/TA	Retained earnings/total assets	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)
X6	EBIT/ATA	Earnings before interest and taxes /average total assets	(Dutta et al. 2017; Riskiyadi 2023)
X7	NI/Sales	Net income/net sales	(An & Suh 2020; Riskiyadi 2023)
X8	GP/Sales (Gross_Margin)	Gross profit/net sales	(Jan 2018; Riskiyadi 2023)

Liquidity Ratios			
X9	WC/TA	Working capital/total assets	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)
X10	CA/TA	Current assets/total assets	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)
X11	CA/CL	Current assets/current liabilities	(An & Suh 2020; Dutta et al. 2017; Jan 2018; Riskiyadi 2023)
X12	OCF/NI	Operating cash flows/net income	(Riskiyadi 2023)
Efficiency Ratios			
X13	REC/Sales	Receivables/net sales	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X14	REC/TA	Receivables/total assets	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X15	INV/Sales	Inventory/net sales	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X16	INV/CA	Inventory/current assets	(Jan 2018; Riskiyadi 2023)
X17	INV/COGS	Inventory/cost of goods sold	(Jan 2018; Riskiyadi 2023)
X18	Sales/ATA	Net sales/average total assets	(Dutta et al. 2017; Riskiyadi 2023)
X19	Sales/ATE	Net sales/average total equity	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)
X20	COGS/Sales	Cost of goods sold/net sales	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)
X21	FA/TA	Fixed assets/total assets	(An & Suh 2020; Dutta et al. 2017; Riskiyadi 2023)
X22	IE/TL	Interest expense/total liabilities	(Chen et al. 2014; Riskiyadi 2023)
X23	OE/Sales	Operating expenses/ net sales	(Jan 2018; Riskiyadi 2023)
X24	EBIT/Sales	Earnings before interest and taxes/net sales	(Kanapickienė & Grundienė 2015; Riskiyadi 2023)

Source: Author's own creation

3-2 Stage 2: Dependent Variable

In terms of focusing on the supervised ML approach, this study labels fraudulent and non-fraudulent interim financial statements based on Beneish's (1999) M-score model. The reason why this study specifically selected Beneish M's score to assess the probability of FSF is because of its well-defined indicators that examine and describe the overall accruals of the firm, surpassing other detection tools. Additionally, it is known for its predictive solid capability in identifying companies that have indeed manipulated and misrepresented their reported earnings, providing reassurance in its superiority (Lehenchuk et al., 2021). Beneish's

(1999) model comprises eight financial ratios (as shown in Table 5) that determine the likelihood of earnings manipulation in financial statements. If the sum of the eight financial ratios is greater than -2.22, then the firm observation is categorized as committing FSF (*value* = 1). Conversely, the firm observation is regarded as non-fraudulent if the M-score is less than-2.22 (*value* = 0). The M-score (1999) model with eight variables is expressed as follows:

$$\text{M-score} = -4.84 + 0.920\ \text{DSRI} + 0.528\ \text{GMI} + 0.404\ \text{AQI} + 0.892\ \text{SGI} + 0.115\ \text{DEPI} - 0.172\ \text{SGAI} - 0.327\ \text{LVGI} + 4.697\ \text{TATA}$$

Applying the M-score model to identify fraudulent and non-fraudulent financial reports for the 2,907 quarter-year observations for firms listed on EGX from 2015 to 2022 revealed that 1,213 interim financial statements are considered fraudulent, representing approximately 41.73 percent. The classification of fraudulent financial reports by year is presented in Table 6.

Table 5: Beneish M-score model financial ratios

Financial Ratios	Formula
Days' sales in receivable index (DSRI)	$\frac{\text{Account Receivable}_t/\ \text{Sales}_t}{\text{Account Receivable}_{t-1}/\ \text{Sales}_{t-1}}$
Gross Margin Index (GMI)	$\frac{\text{Sales}_{t-1} - \text{Costs of goods sold}_{t-1}/\ \text{Sales}_{t-1}}{\text{Sales}_t - \text{Costs of goods sold}_t/\ \text{Sales}_t}$
Asset Quality Index (AQI)	$\frac{\text{Total Assets}_t - (\text{Current Assets}_t + \text{Property,Plant and Equipment}_t)/\ \text{Total Assets}_t}{\text{Total Assets}_{t-1} - (\text{Current Assets}_{t-1} + \text{Property,Plant and Equipment}_{t-1})/\ \text{Total Assets}_{t-1}}$
Sales Growth Index (SGI)	$\frac{\text{Sales}_t}{\text{Sales}_{t-1}}$
Depreciation Index (DEPI)	$\frac{\text{Depreciation}_{t-1}\ /(\text{Property,Plant and Equipment}_{t-1} + \text{Depreciation}_{t-1})}{\text{Depreciation}_t\ /(\text{Property,Plant and Equipment}_t + \text{Depreciation}_t)}$
Sales, General and Administrative Expense Index (SGAI)	$\frac{\text{SGandA Expense}_t/\ \text{Sales}_t}{\text{SGandA}_{t-1}/\ \text{Sales}_{t-1}}$
Leverage Index (LVGI)	$\frac{(\text{Current Liabilities}_t + \text{Long} - \text{Term Debt}_t)/\ \text{Total Assets}_t}{(\text{Current Liabilities}_{t-1} + \text{Long} - \text{Term Debt}_{t-1})/\ \text{Total Assets}_{t-1}}$
Total Accruals to Total Assets (TATA)	$\frac{\text{Income from Continuing Operations}_t - \text{Operational Cash Flows}_t}{\text{Total Assets}_t}$

Table 6: Classification of Fraudulent Interim Financial Statements in Years

Year	Number of fraudulent interim financial statements	Percentage
2015	160	5.50
2016	166	5.71
2017	161	5.54
2018	163	5.61
2019	161	5.54
2020	128	4.40
2021	135	4.64
2022	139	4.78
Total	*1,213*	*41.73*

Source: Author's own creation

3-3 Stage 3: Data Pre-processing

By this stage, the dataset derived from the financial statements of non-financial firms listed on EGX during the research period initiates the research flow. This dataset is then divided into training, validation, and testing subsets, following the 60:20:20 ratio recommended by previous studies. The training, in addition to the validation datasets, are used to train and fit the models, with the former being used to build the model and the latter to determine the most appropriate hyperparameter values (An & Suh, 2020). Finally, the testing dataset is used to assess the classification performance of the ML models generated. However, to boost the performance of ML classifiers, it is essential to preprocess the data to ensure a clean dataset before the algorithms' actual use. Before executing the data into the model, data preprocessing addresses real-life data challenges such as noise, errors, inconsistencies, and missing values. Therefore, the following section discusses the data pre-processing conducted in this study across two main steps: using different sampling techniques and performing dimensionality reduction.

3-3-1 Sampling Methods

The dataset used in this study presents a class imbalance, with the majority of observations being non-fraudulent and the minority being fraudulent. The class imbalance problem poses a challenge as it would result in less reliable performance for the ML models if used without data preprocessing. Thereby, in line with prior literature (Chen et al., 2014; Dutta et al., 2017; Hajek & Henriques, 2017; Liu et al., 2015; Riskiyadi, 2023), this research takes an innovative approach to resolving the class imbalance problem by using oversampling, which adds new synthetic data for the minority class using the Synthetic Minority Oversampling Technique (SMOTE). It also considers undersampling implemented using the Equal Size Sampling (ESS).

3-3-1-1 Oversampling

SMOTE was first introduced by Chawla et al. (2002) as an oversampling method that artificially generates new random synthetic samples of the minority class by interpolating between the adjacent minority class's original instances instead of just replicating the existing instances (Luo, 2019). As such, from given sample x, the minority class k with the nearest neighbors x^N characterized by the smallest Euclidean distance for quantitative features and the smallest value distance metric for qualitative features are identified and selected. The minority class k is randomly chosen depending on the amount of the required oversampling in the study, in which the new SMOTE sample x_{new} is defined as follows:

$$x_{\text{new}} = x + u\,(x^{N} - x) \qquad (1)$$

Where the difference between the minority class sample and its nearest neighbor $(x^N - x)$ is multiplied by a random number u that lies between zero and one and then added to the sample x. This resampling technique guarantees that the new SMOTE sample x_{new} is lying on the line segment between the two original samples used in generating it (Chawla et al., 2002). SMOTE is the most efficient technique for obtaining a more balanced training dataset.

3-3-1-2 Undersampling

In contrast to oversampling, undersampling works on removing random instances of the majority class to ensure equal data distribution (Ayad et al., 2023). According to Konstanz Information Miner (KNIME), the ESS undersampling technique implies that the node will randomly drop rows belonging to the majority class and return the entire minority class records so that the dataset sent back by the node would be of equal size of both classes (Shang et al., 2021). However, randomization may be challenging, as the node may remove the clean instance that would originally enhance the model performance and keep the noisy instance that would deteriorate the model's performance (Hasanin et al., 2019). For this reason, the best resampling technique has been a subject of debate. To address this, the research has comprehensively evaluated the effectiveness of oversampling and undersampling techniques carried out on the research sample dataset of listed Egyptian non-financial firms.

3-3-2 Dimensionality Reduction

In the era of big, high-dimensional data, ML models' performance may deteriorate, and computational complexity may rise with the increasing number of input features, especially since not all features are always essential. Therefore, dimensionality reduction aims to mitigate the high-dimensionality problem by reducing the number of features while retaining the most representative features from the dataset. This research has applied dimensionality reduction using correlation-based analysis (CBA) and Principal Component Analysis (PCA).

3-3-2-1 Dimensionality Reduction Using CBA

To address the endogeneity problem, CBA has been implemented between all input features used in this study before being executed in the model training. Pearson's Product Moment Correlation Coefficient has been used to analyze the correlations between input features. The heatmap of Figure 2 plots the correlation matrix of all input features in the dataset, where positive correlations are shown in orange, and negative correlations are in blue. From the CBA of the entire dataset, the results indicated a significant correlation of *TL/TA* with *RE/TA; thus, RE/TA* has been dropped from the input features dataset. Similarly, a high correlation has been concluded between *EBIT/ATA* and *NI/ATA,* leading to the further choice of the highly significant *NI/ATA* to reflect information about the firm's profitability. Moreover, the *WC/TA* has been removed from the dataset consequent to its significant correlation with *TL/TA*. Focusing on the features indicating the efficiency of the firm, the *INV/CA, INV/COGS, Sales/ATE, COGS/Sales, FA/TA,* and *EBIT/Sales* have been dropped as they resulted in a significant correlation with other features in the dataset. To sum up, ten of the highly correlated insignificant input features have been excluded from the original dataset to ensure a higher quality of results. Thus, a total of 14 input features are concerned with further processing. Figure 3 shows the heatmap of correlations between the 14 selected features.

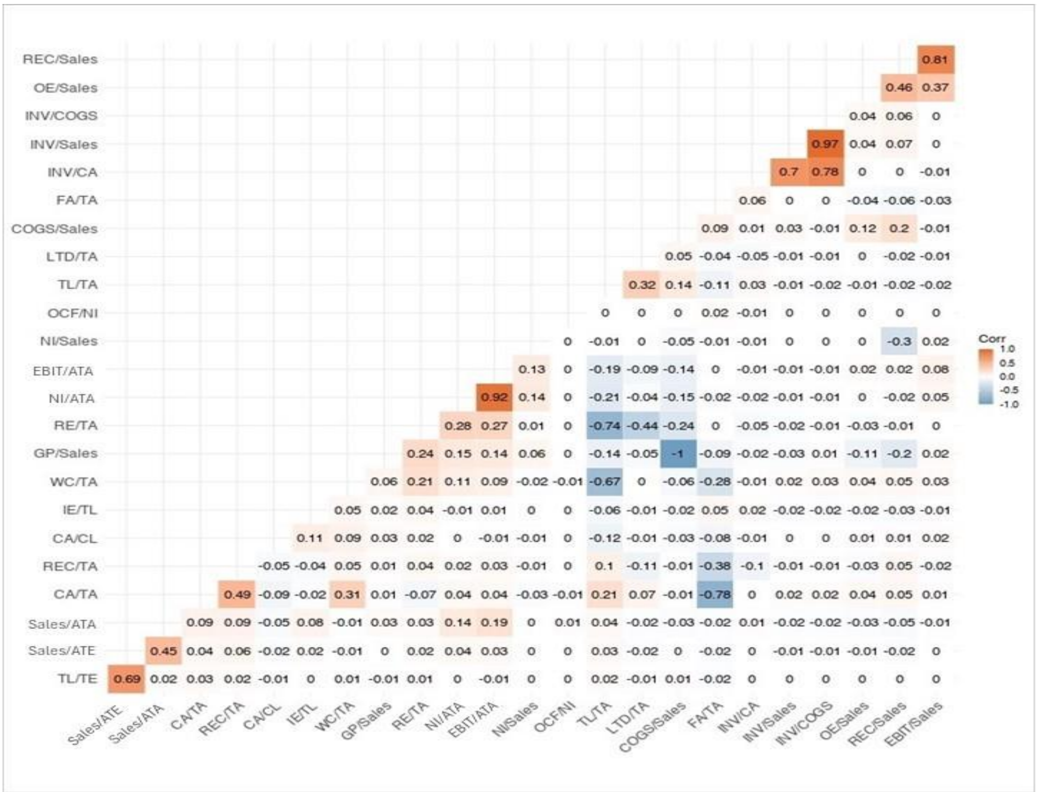


Figure 2: Heat map plot showing the correlations between all 24 input features

Source: Author’s own creation

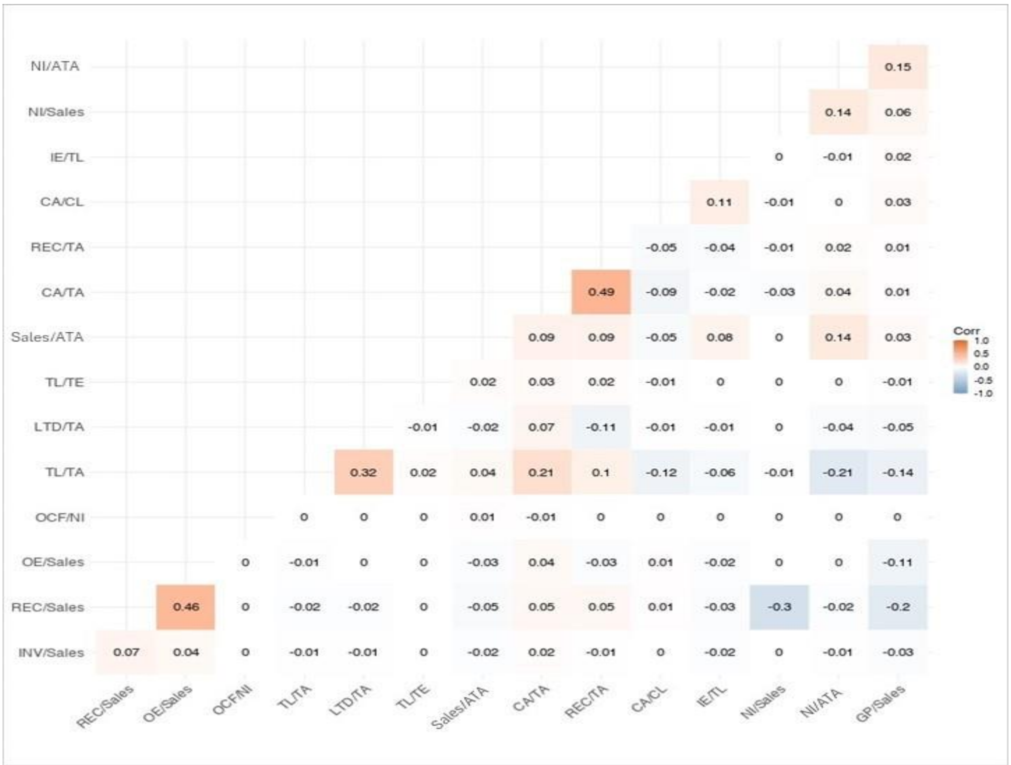


Figure 3: Heat map plot showing the correlations between selected 14 input features

Source: Author's own creation

3-3-2-2 Dimensionality Reduction Using PCA

For further analysis of the input features based on their variances, this study also used the best-known PCA algorithm to reduce the number of features while facing a limited loss of information in the original dataset (Karamizadeh et al., 2013). In the accounting discipline, PCA offers a way of reducing the number of ratios already noted in the literature to be used in the analysis by selecting the most statistically important ratios to be processed in further analysis with a minimum bias (Mbona & Yusheng, 2019). The first principal component combines the *X-features* of maximum variance among all combinations, in which most of the data variations are taken by this first component. The following component takes the remaining maximum variance in the data while keeping the condition of zero correlation between the first and the second components. This process is to be repeated until the *i*th component counts to the last maximum variation missed by other components while maintaining the fulfillment of the zero-correlation condition. The zero-correlation condition between components creates the independence between features used in the dataset. Using the eigenvector $\hat{e}$ as the coefficient, the PCA comes up with the following equations:

$$Y_1 = \hat{e}_{11}\, ZX_1 + \hat{e}_{12}\, ZX_2 + \hat{e}_{13}\, ZX_3 + \dots + \hat{e}_{1i}\, ZX_i, \qquad (2)$$

$$Y_2 = \hat{e}_{21}\, ZX_1 + \hat{e}_{22}\, ZX_2 + \hat{e}_{23}\, ZX_3 + \dots + \hat{e}_{2i}\, ZX_i, \qquad (3)$$

$$Y_i = \hat{e}_{i1}\, ZX_1 + \hat{e}_{i2}\, ZX_2 + \hat{e}_{i3}\, ZX_3 + \dots + \hat{e}_{ii}\, ZX_i, \qquad (4)$$

Whereas Y is the principal component and *ZX* is the ratios' standardized values.

3-4 Stage 4: Developing Machine Learning Models

This research has employed three models to detect the presence of FSF in the Egyptian context. These models were selected based on a thorough review of the literature and their suitability with the research objective. Logistic Regression was chosen for its simplicity and interpretability, Decision Tree for its ability to handle large amounts of data, and Random Forest (RF) for its ensemble approach that combines the strengths of multiple decision trees. These models were implemented in Python using the user-friendly Scikit-Learn library that provides the basic constructions for the ML algorithms.

3-4-1 Logistic Regression (LR)

LR was first introduced by Ohlson (1980) who revolutionized financial studies after predicting corporate failure, evidenced by the bankruptcy declaration. Since then, LR has been considered one of the most popular classical statistical methods used for classification types of problems; as such, it has gained the wide attention of many studies intending to the probability of a specific event using a set of input features. Liou (2008) defined LR as the non-linear method for modeling a dichotomous variable of interest (i.e., a dependent variable) of more than one value, such as true/false, zero/one, etc., with a set of independent variables. LR employs a logistic function that maps a linear combination between a binary dependent variable and input features, along with transforming labels by converting the log odds to a probability range of zero, and one (Li & Wu, 2022). The following formula illustrates the process of constructing a LR model:

$$\text{Output} = Y \approx P(X) = \frac{1}{1+e^{-(\beta_0+\beta X)}} \qquad (5)$$

$$\text{Input} = X = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})^T \qquad (6)$$

$$\beta = (\beta_1, \beta_2, \ldots, \beta_n) \qquad (7)$$

X is the vector of influencing factors, P is the model probability, and β is the parameter. Using the approach of maximum likelihood estimation to estimate the model parameters, the study owns an input parameter X and a binary output variable Y for each influential factor. In this research context, if label $Y = 1$ is satisfied (i.e., there is a FFR), then the probability of the output variable is $p(x_i)$, or if $Y = 0$, then there is no FFR, and the probability of output variable is $1-p(x_i)$. Thus, the likelihood of the presence of FFR can be defined in the following function:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1-p(x_i))^{1-y_i} \qquad (8)$$

3-4-2 Decision Tree (DT)

Turning to tree-based algorithms, DT has emerged as a popular modeling approach in modern data analytical studies. Its appeal lies in its ability to handle numerical, categorical, and even missing data with ease of understanding and interpretation. Moreover, DTs are known for their efficiency, making them ideal for large datasets. Lin et al. (2015) defined DT as the predictive model characterized by having a hierarchical or tree structure. Halteh et al. (2018) defined DTs as models that construct tree-based classification rules that break down a dataset into smaller subsets. As such, DT in the FSF context aims to divide observations into mutually exclusive fraudulent/non-fraudulent subgroups by selecting the features that can best classify them (Zhou & Kapoor, 2011). There are various DT algorithms, but the Chi-squared Automatic Interaction Detector (CHAID), C4.5, C5.0, Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), and the Quick, Unbiased, Efficient Statistical Tree (QUEST) are the most commonly known ones.

This study utilizes Supervised Learning in QUEST, commonly known as SLIQ, where QUEST is the binary-split DT algorithm used for classification and data mining (Loh & Shih, 1997). SLIQ was established by M. Mehta et al. (1996) as the first fast, scalable decision tree classifier that can handle large numeric and categorical datasets using a pre-sorting technique during the tree-building procedure instead of having the data recursively sorted at each node of the DT, as found in CART and C4.5, this presorting technique is integrated by a breadth-first tree-growing strategy that enables the one-time sorting of disk-resident datasets without being resident in the main memory. Thus, it reduces the evaluation costs of the numeric attributes. With the use of a data *class* list that stores the class labels for each single record, the Gini impurity index is used as the quality criteria for evaluating all possible splits, and the ones with the least impurity are selected. To avoid overfitting, SLIQ uses a Minimum Description Length (MDL) strategy to prude trees built during the growth phase. MDL is the strategy used for evaluating the node's accuracy, which is generalized based on the code length at each DT node, as it states that the *best* tree is the one encoded with the least number of bits.

3-4-3 Random Forest (RF)

RF is an ensemble decision tree model that has a decision tree as the basic unit. Randomly selected variables and observations extracted from the database construct each tree N in RF, where each tree is a classifier that will have a classification result for input cases (Li & Wu, 2022), and the final output represents the result provided by the majority of the trees (Breiman, 2001). Similar to DT, RF can be applied to classification and regression tasks of binary and continuous outputs, respectively. Assuming that a dataset D stated as follows:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \qquad (9)$$

Whereas D is composed of an output (i.e., dependent variable) y_i and a set of inputs x_i whose features are denoted by $(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T$. Forests are built based on a number of decision trees expected to be generated by the algorithm. Trees are generated based on random execution in the following iterative way: On each iteration, a random subsample of the features extracted from dataset D is selected through Bootstrap to form a subsample D_i. Next, each D_i generates a single tree T_i using the CART algorithm. Therefore, each tree has only a limited number of features m less than the total number of features in the dataset D. Bagging methods are utilized after building the random trees to forecast the final output.

3-5 Stage 5: Models Evaluation

A set of evaluation criteria was used in this study to compare the models and identify the best FSF classifier. The evaluation criteria were used to assess the accuracy and performance power of the models built using the confusion matrix illustrated in Table 7. Based on the research work of Riskiyadi (2023), West & Bhattacharya (2015), and other studies shown in Table 2, this research uses the following criteria in evaluating the ML models:

3-5-1 Cohen’s Kappa

Cohen’s kappa, a statistical indicator of interrater reliability, plays an essential role in understanding the appropriate degree of variable demonstration for the model training. It measures the degree of agreement between the classifier and the actual case, providing practical insights into the model performance. Values of kappa closer to one indicate a strong demonstration of variables for model training, while values closer to zero indicate uncertainty. A negative value of kappa indicates that the demonstration of variables is less than that expected by chance.

3-5-2 Accuracy

Accuracy assesses the model's ability to correctly distinguish the fraudulent and non-fraudulent firm quarter-year observations. Accuracy is calculated by dividing the number of true positive and true negative cases identified by the algorithm by the total number of cases, as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (10)$$

3-5-3 Type I error

Type I error, or False Positive Rate (FPR), is an indicator of the number of actual non-fraudulent financial statements classified by the ML model as fraudulent financial statements, compared to the total actual non-fraudulent cases. The Type I error rate is calculated by taking the number of false positive cases and dividing it by the total number of actual negative cases., as follows:

$$Type\ I\ error\ or\ FPR = \frac{FP}{TN+FP} \qquad (11)$$

3-5-4 Type II error

Type II error, or False Negative Rate (FNR), is conversely an indicator of the number of actual fraudulent financial statements classified by the ML model as non-fraudulent financial statements, compared to the total actual FFR observations. Type II error is calculated by scaling the number of false negative cases by the total actual positive cases, as shown below:

$$Type\ II\ error\ or\ FNR = \frac{FN}{TP+FN} \qquad (12)$$

3-5-5 Sensitivity

Sensitivity, commonly known as recall or true positive rate (TPR), shows the ability of the algorithm to identify the observations having FFR. Thus, it is calculated by scaling the number of true positive cases classified by the model to the actual positive cases, as shown below:

$$Sensitivity\ or\ Recall\ or\ TPR = \frac{TP}{TP+FN} \qquad (13)$$

3-5-6 Specificity

The specificity, commonly known as the true negative rate (TNR), is the ability of the algorithm to determine the non-fraudulent cases correctly. Specificity is calculated as the number of classified true negative cases compared to the actual negative cases, as follows:

$$\textit{Specificity or TNR} = \frac{TN}{TN+FP} \qquad (14)$$

3-5-7 Precision

Precision is the number of true positive cases classified by the ML model compared to all positive cases classified by the model; as such, the precision is determined using the following equation:

$$\textit{Precision} = \frac{TP}{TP+FP} \qquad (15)$$

3-5-8 F-measure

F-measure is the metric used to evaluate the performance of the ML model by integrating the precision and sensitivity into a single score, as follows:

$$\textit{F-measure} = \frac{2*precision*sensitivity}{precision+sensitivity} \qquad (16)$$

Table 7: Confusion Matrix

Classification detected by ML classifier	**Actual situation**	
	Fraudulent Financial Reporting (Positive, value =1)	*Non-Fraudulent Financial Reporting (Negative, value = 0)*
Fraudulent Financial Reporting (Positive, value =1)	True Positive (TP)	False Positive (FP) (Type I error)
Non-Fraudulent Financial Reporting (Negative, value = 0)	False Negative (FN) (Type II error)	True Negative (TN)

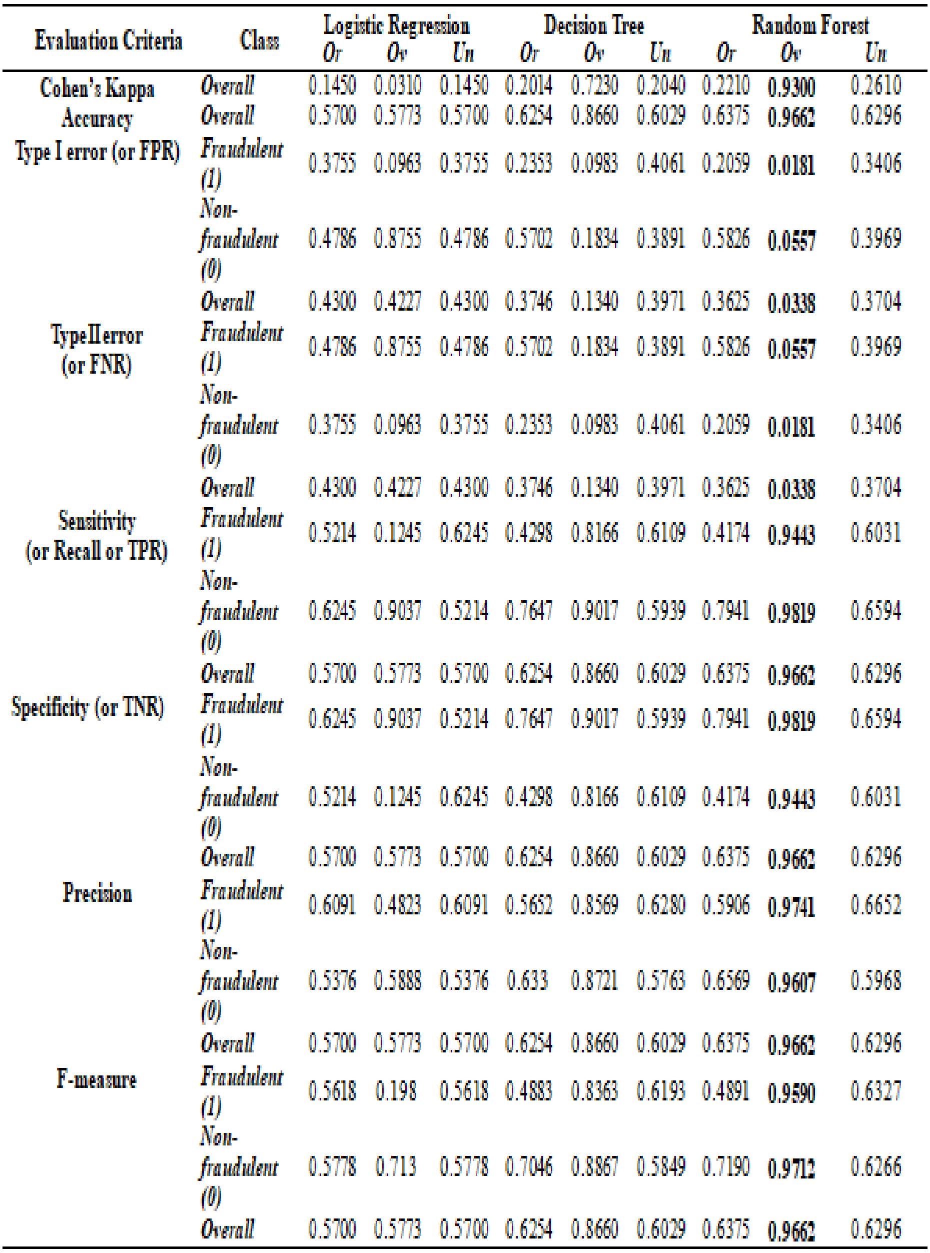
Note(s): TP is the correct positive classification for an actual positive case; TN is the correct negative classification for an actual negative case; FP is the false positive classification for an actual negative case; FN is the false negative classification for an actual positive case.

Source: Author's own creation

4- Findings and Discussion

Tables 8-11 summarize the research results, showing the best classifier results in bold. Additionally, Figures 4-7 plot the corresponding confusion matrices. Table 8 compares the classification performance of LR, DT, and ensemble classifier RF, in which ML models were developed using the original imbalanced dataset and balanced oversampling and undersampling datasets. Figure 4 illustrates the respective matrices for each ML model with the different data sampling techniques. Each of these models has been developed without using any dimensionality reduction technique. The results indicate unreliable and inaccurate performance of ML models, especially for LR, which resulted in an extremely low sensitivity, precision, and F-measure. Based on the oversampling dataset, DT and RF showed better FSF classification performance than LR, especially RF, which resulted in the highest Cohen's kappa and accuracy rate of 93 and 96.62 percent, respectively.

Table 8: Performance Evaluation of Classification Models Without Dimensionality Reduction



Evaluation Criteria	Class	Logistic Regression			Decision Tree			Random Forest		
		Or	*Ov*	*Un*	*Or*	*Ov*	*Un*	*Or*	*Ov*	*Un*
Cohen's Kappa	*Overall*	0.1450	0.0310	0.1450	0.2014	0.7230	0.2040	0.2210	**0.9300**	0.2610
Accuracy	*Overall*	0.5700	0.5773	0.5700	0.6254	0.8660	0.6029	0.6375	**0.9662**	0.6296
Type I error (or FPR)	*Fraudulent (1)*	0.3755	0.0963	0.3755	0.2353	0.0983	0.4061	0.2059	**0.0181**	0.3406
	Non-fraudulent (0)	0.4786	0.8755	0.4786	0.5702	0.1834	0.3891	0.5826	**0.0557**	0.3969
	Overall	0.4300	0.4227	0.4300	0.3746	0.1340	0.3971	0.3625	**0.0338**	0.3704
Type II error (or FNR)	*Fraudulent (1)*	0.4786	0.8755	0.4786	0.5702	0.1834	0.3891	0.5826	**0.0557**	0.3969
	Non-fraudulent (0)	0.3755	0.0963	0.3755	0.2353	0.0983	0.4061	0.2059	**0.0181**	0.3406
	Overall	0.4300	0.4227	0.4300	0.3746	0.1340	0.3971	0.3625	**0.0338**	0.3704
Sensitivity (or Recall or TPR)	*Fraudulent (1)*	0.5214	0.1245	0.6245	0.4298	0.8166	0.6109	0.4174	**0.9443**	0.6031
	Non-fraudulent (0)	0.6245	0.9037	0.5214	0.7647	0.9017	0.5939	0.7941	**0.9819**	0.6594
	Overall	0.5700	0.5773	0.5700	0.6254	0.8660	0.6029	0.6375	**0.9662**	0.6296
Specificity (or TNR)	*Fraudulent (1)*	0.6245	0.9037	0.5214	0.7647	0.9017	0.5939	0.7941	**0.9819**	0.6594
	Non-fraudulent (0)	0.5214	0.1245	0.6245	0.4298	0.8166	0.6109	0.4174	**0.9443**	0.6031
	Overall	0.5700	0.5773	0.5700	0.6254	0.8660	0.6029	0.6375	**0.9662**	0.6296
Precision	*Fraudulent (1)*	0.6091	0.4823	0.6091	0.5652	0.8569	0.6280	0.5906	**0.9741**	0.6652
	Non-fraudulent (0)	0.5376	0.5888	0.5376	0.633	0.8721	0.5763	0.6569	**0.9607**	0.5968
	Overall	0.5700	0.5773	0.5700	0.6254	0.8660	0.6029	0.6375	**0.9662**	0.6296
F-measure	*Fraudulent (1)*	0.5618	0.198	0.5618	0.4883	0.8363	0.6193	0.4891	**0.9590**	0.6327
	Non-fraudulent (0)	0.5778	0.713	0.5778	0.7046	0.8867	0.5849	0.7190	**0.9712**	0.6266
	Overall	0.5700	0.5773	0.5700	0.6254	0.8660	0.6029	0.6375	**0.9662**	0.6296

Source: Author's own creation

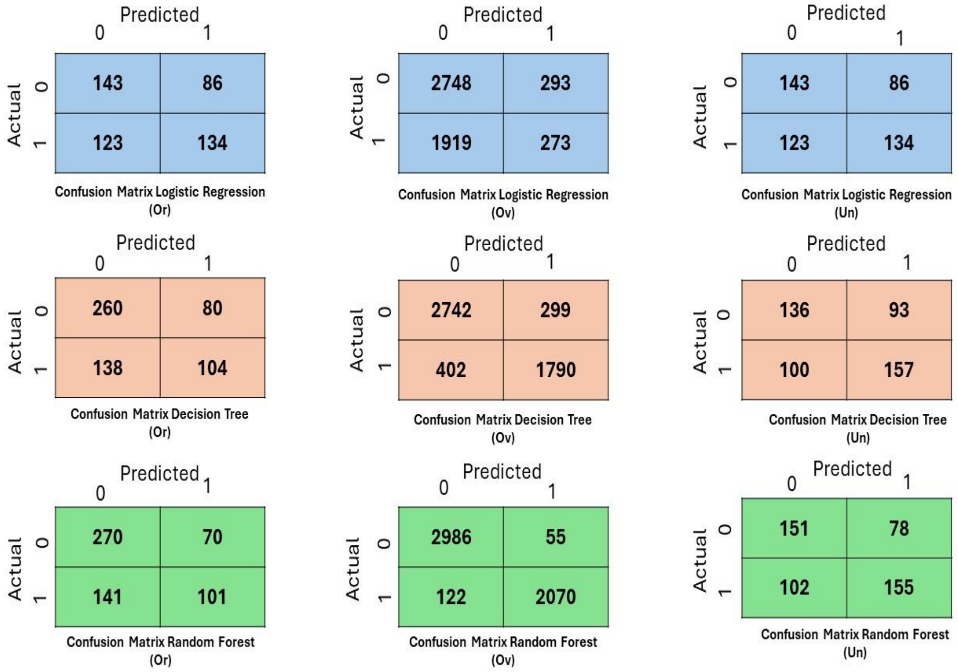


Figure 4: Confusion Matrices for Machine Learning Models without Dimensionality Reduction

Source: Author’s own creation

After using the CBA for dimensionality reduction, better results were achieved by the different ML models, as shown in Table 9. The confusion matrices plotted in Figure 5 reflect that RF outperformed itself without the dimensionality reduction, especially with the oversampling dataset. Although oversampling significantly improved the performance of RF as it resulted in a Cohen’s kappa closer to one, a high accuracy rate of 98.85 percent, and a high overall sensitivity of 98.55 percent, the results shown in Table 9 indicate that the undersampling conversely does not affect RF. The findings are compatible with other studies, such as Mohammed et al. (2020) who suggested that undersampling results in poorer performance than oversampling, as it may drop valuable data essential for classifiers.

Table 9: Performance Evaluation of Classification Models With Dimensionality Reduction Using Correlation-Based Analysis

Evaluation Criteria	**Class**	**Logistic Regression**			**Decision Tree**			**Random Forest**		
		Or	***Ov***	***Un***	***Or***	***Ov***	***Un***	***Or***	***Ov***	***Un***
Cohen's Kappa	***Overall***	-0.010	0.0270	0.0870	0.2020	0.8300	0.1960	0.2880	**0.9700**	0.2880
Accuracy	***Overall***	0.5636	0.5840	0.5391	0.6203	0.9173	0.5947	0.6440	**0.9855**	0.6440
Type I error (or FPR)	***Fraudulent (1)***	0.1000	0.0424	0.3712	0.2735	0.0677	0.3231	0.3537	**0.0056**	0.3537
	Non-fraudulent (0)	0.9091	0.9343	0.5409	0.5289	0.1036	0.4786	0.3580	**0.0269**	0.3580
	Overall	0.4364	0.4160	0.4609	0.3797	0.0827	0.4053	0.3560	**0.0145**	0.3560
Type II error (or FNR)	***Fraudulent (1)***	0.9091	0.9343	0.5409	0.5289	0.1036	0.4786	0.3580	**0.0269**	0.3580
	Non-fraudulent (0)	0.1000	0.0424	0.3712	0.2735	0.0677	0.3231	0.3537	**0.0056**	0.3537
	Overall	0.4364	0.4160	0.4609	0.3797	0.0827	0.4053	0.3560	**0.0145**	0.3560
Sensitivity (or Recall or TPR)	***Fraudulent (1)***	0.0909	0.0657	0.4591	0.4711	0.8964	0.5214	0.6420	**0.9731**	0.6420
	Non-fraudulent (0)	0.9000	0.9576	0.6288	0.7265	0.9323	0.6769	0.6463	**0.9944**	0.6463
	Overall	0.5636	0.5840	0.5391	0.6203	0.9173	0.5947	0.6440	**0.9855**	0.6440
Specificity (or TNR)	***Fraudulent (1)***	0.9000	0.9576	0.6288	0.7265	0.9323	0.6769	0.6463	**0.9944**	0.6463
	Non-fraudulent (0)	0.0909	0.0657	0.4591	0.4711	0.8964	0.5214	0.6420	**0.9731**	0.6420
	Overall	0.5636	0.5840	0.5391	0.6203	0.9173	0.5947	0.6440	**0.9855**	0.6440
Precision	***Fraudulent (1)***	0.3929	0.5275	0.5813	0.5507	0.9051	0.6442	0.6707	**0.9921**	0.6707
	Non-fraudulent (0)	0.5817	0.5871	0.5088	0.6587	0.9259	0.5576	0.6167	**0.9809**	0.6167
	Overall	0.5636	0.5840	0.5391	0.6203	0.9173	0.5947	0.6440	**0.9855**	0.6440
F-measure	***Fraudulent (1)***	0.1477	0.1168	0.5130	0.5078	0.9008	0.5763	0.6561	**0.9825**	0.6561
	Non-fraudulent (0)	0.7067	0.7279	0.5625	0.6909	0.9291	0.6114	0.6311	**0.9876**	0.6311
	Overall	0.5636	0.5840	0.5391	0.6203	0.9173	0.5947	0.6440	**0.9855**	0.6440

Source: Author's own creation

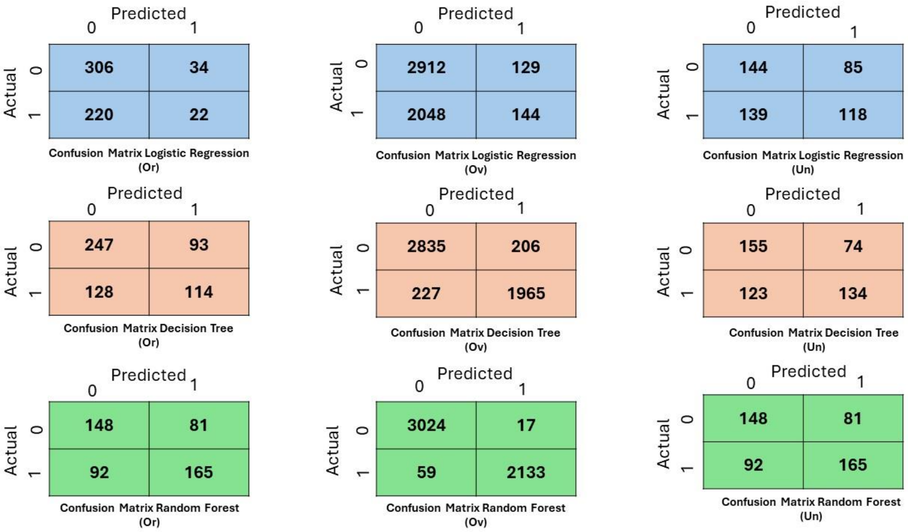


Figure 5: Confusion Matrices for Machine Learning Models with Dimensionality Reduction Using Correlation-Based Analysis

Source: Author's own creation

The study also employed the PCA algorithm on the entire dataset to compare the effect of different dimensionality reduction techniques. The models' confusion matrices and performance evaluation results are provided in Figure 6 and Table 10, respectively. The results show that the DT algorithm marked the superior performer with the original and undersampling datasets with accuracy rates of 63.92 and 62.14 percent, respectively. However, RF still outperforms other classifiers with the oversampling dataset. Moreover, the results support the findings of LR's worst performance with the three datasets, as suggested when using the CBA.

Table 10: Performance Evaluation of Classification Models With Dimensionality Reduction Using PCA

Evaluation Criteria	**Class**	**Logistic Regression**			**Decision Tree**			**Random Forest**		
		Or	***Ov***	***Un***	***Or***	***Ov***	***Un***	***Or***	***Ov***	***Un***
Cohen's Kappa	***Overall***	0.0000	-0.0030	0.052	0.2456	0.7013	0.2420	0.1800	**0.8680**	0.1800
Accuracy	***Overall***	0.5842	0.5788	0.5062	0.6392	0.8552	0.6214	0.6117	**0.9364**	0.6117
Type I error (or FPR)	***Fraudulent (1)***	0.0000	0.0082	0.1092	0.2706	0.1144	0.3755	0.2706	**0.0372**	0.2706
	Non-fraudulent (0)	1.0000	0.9941	0.8366	0.4876	0.1870	0.3813	0.5537	**0.1004**	0.5537
	Overall	0.4158	0.4212	0.4938	0.3608	0.1448	0.3786	0.3883	**0.0636**	0.3883
Type II error (or FNR)	***Fraudulent (1)***	1.0000	0.9941	0.8366	0.4876	0.1870	0.3813	0.5537	**0.1004**	0.5537
	Non-fraudulent (0)	0.0000	0.0082	0.1092	0.2706	0.1144	0.3755	0.2706	**0.0372**	0.2706
	Overall	0.4158	0.4212	0.4938	0.3608	0.1448	0.3786	0.3883	**0.0636**	0.3883
Sensitivity (or Recall or TPR)	***Fraudulent (1)***	0.0000	0.9918	0.1634	0.5124	0.8130	0.6187	0.4463	**0.8996**	0.4463
	Non-fraudulent (0)	1.000	0.0059	0.8908	0.7294	0.8856	0.6245	0.7294	**0.9628**	0.7294
	Overall	0.5842	0.5788	0.5062	0.6392	0.8552	0.6214	0.6117	**0.9364**	0.6117
Specificity (or TNR)	***Fraudulent (1)***	1.0000	0.9918	0.8908	0.7294	0.8856	0.6245	0.7294	**0.9628**	0.7294
	Non-fraudulent (0)	0.0000	0.0059	0.1634	0.5124	0.8130	0.6187	0.4463	**0.8996**	0.4463
	Overall	0.5842	0.5788	0.5062	0.6392	0.8552	0.6214	0.6117	**0.9364**	0.6117
Precision	***Fraudulent (1)***	0.0000	0.3421	0.6269	0.5741	0.8366	0.6490	0.5400	**0.9458**	0.5400
	Non-fraudulent (0)	0.5842	0.5806	0.4869	0.6776	0.8679	0.5934	0.6492	**0.9301**	0.6492
	Overall	0.5842	0.5788	0.5062	0.6392	0.8552	0.6214	0.6117	**0.9364**	0.6117
F-measure	***Fraudulent (1)***	0.0000	0.0117	0.2593	0.5415	0.8246	0.6335	0.4887	**0.9221**	0.4887
	Non-fraudulent (0)	0.7375	0.7324	0.6296	0.7025	0.8766	0.6085	0.6870	**0.9462**	0.6870
	Overall	0.5842	0.5788	0.5062	0.6392	0.8552	0.6214	0.6117	**0.9364**	0.6117

Source: Author's own creation

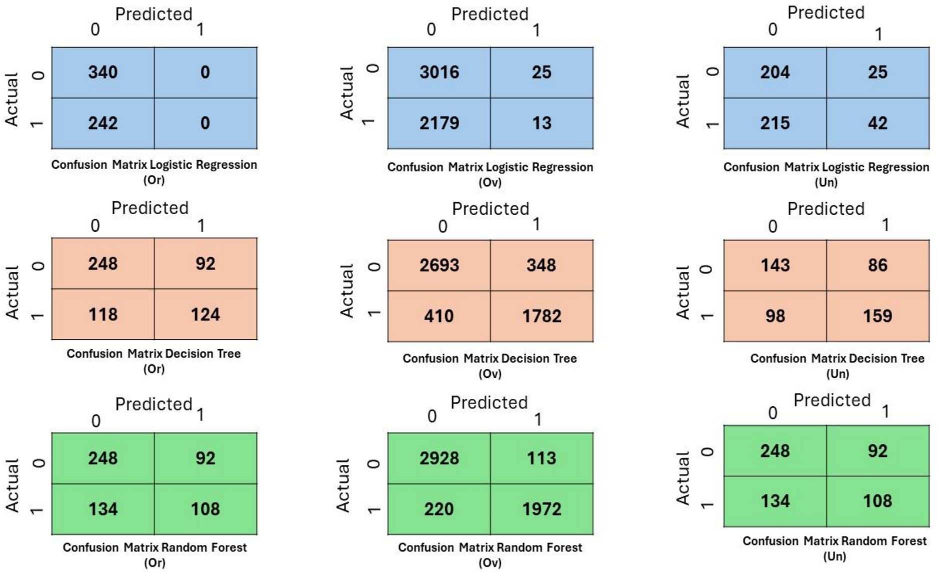


Figure 6: Confusion Matrices for Machine Learning Models with Dimensionality Reduction Using Principal Component Analysis

Source: Author's own creation

Based on results provided in Table 11 and confusion matrices in Figure 7 showing the models' performance after employing a combined technique of CBA and PCA, it can be concluded that the RF once again performs best with each dataset treatment over fraudulent, non-fraudulent, and overall classes.

Table 11: Performance Evaluation of Classification Models With Dimensionality Reduction Using Correlation-Based Analysis and PCA

Performance Evaluation Criteria	**Class**	**Logistic Regression**			**Decision Tree**			**Random Forest**		
		Or	***Ov***	***Un***	***Or***	***Ov***	***Un***	***Or***	***Ov***	***Un***
Cohen's Kappa	***Overall***	0.0010	-0.0020	0.0170	0.1010	0.7380	0.1870	0.2100	**0.9210**	0.2110
Accuracy	***Overall***	0.5842	0.5790	0.4877	0.5739	0.8737	0.5905	0.6271	**0.9618**	0.6049
Type I error (or FPR)	***Fraudulent (1)***	0.0029	0.0085	0.1223	0.3059	0.0881	0.3406	0.2500	**0.0161**	0.3755
	Non-fraudulent (0)	0.9959	0.9932	0.8599	0.5950	0.1793	0.4708	0.5455	**0.0689**	0.4125
	Overall	0.4158	0.4210	0.5123	0.4261	0.1263	0.4095	0.3729	**0.0382**	0.3951
Type II error (or FNR)	***Fraudulent (1)***	0.9959	0.0085	0.8599	0.5950	0.1793	0.4708	0.5455	**0.0689**	0.4125
	Non-fraudulent (0)	0.0029	0.9932	0.1223	0.3059	0.0881	0.3406	0.2500	**0.0161**	0.3755
	Overall	0.4158	0.4210	0.5123	0.4261	0.1263	0.4095	0.3729	**0.0382**	0.3951
Sensitivity (or Recall or TPR)	***Fraudulent (1)***	0.0041	0.0068	0.1401	0.4050	0.8207	0.5292	0.4545	**0.9311**	0.5875
	Non-fraudulent (0)	0.9971	0.9915	0.8777	0.6941	0.9119	0.6594	0.7500	**0.9839**	0.6245
	Overall	0.5842	0.5790	0.4877	0.5739	0.8737	0.5905	0.6271	**0.9618**	0.6049
Specificity (or TNR)	***Fraudulent (1)***	0.0041	0.9915	0.8777	0.6941	0.9119	0.6594	0.7500	**0.9839**	0.6245
	Non-fraudulent (0)	0.9971	0.0068	0.1401	0.4050	0.8207	0.5292	0.4545	**0.9311**	0.5875
	Overall	0.5842	0.5790	0.4877	0.5739	0.8737	0.5905	0.6271	**0.9618**	0.6049
Precision	***Fraudulent (1)***	0.5000	0.3659	0.5625	0.4851	0.8703	0.6355	0.5641	**0.9766**	0.6371
	Non-fraudulent (0)	0.5845	0.5807	0.4763	0.6211	0.8759	0.5551	0.6589	**0.9520**	0.5743
	Overall	0.5842	0.5790	0.4877	0.5739	0.8737	0.5905	0.6271	**0.9618**	0.6049
F-measure	***Fraudulent (1)***	0.7370	0.0134	0.2243	0.4414	0.8448	0.5775	0.5034	**0.9533**	0.6113
	Non-fraudulent (0)	0.0082	0.7374	0.6175	0.6556	0.8935	0.6028	0.7015	**0.9677**	0.5983
	Overall	0.5842	0.5790	0.4877	0.5739	0.8737	0.5905	0.6271	**0.9618**	0.6049

Source: Author's own creation

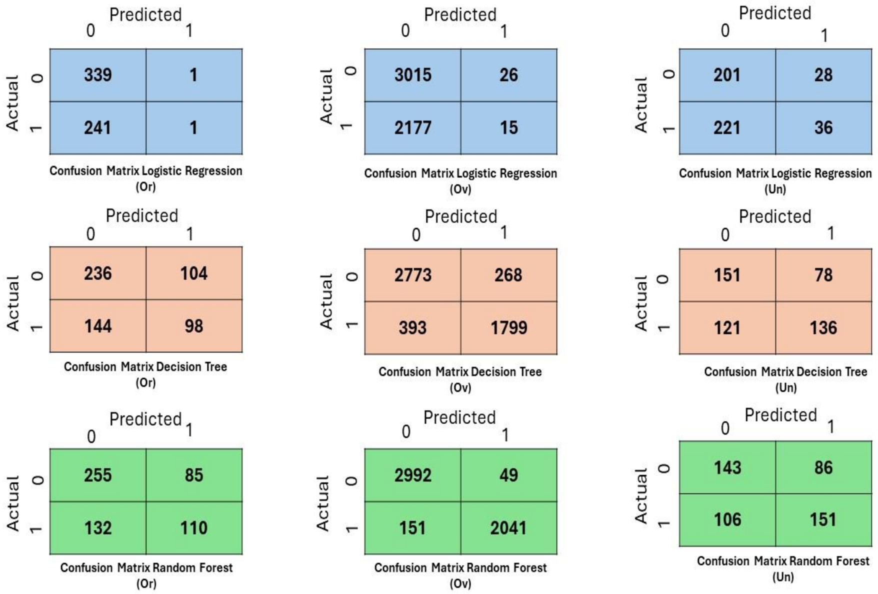


Figure 7: Confusion Matrices for Machine Learning Models with Dimensionality Reduction Using Correlation-Based Analysis and Principal Component Analysis

Source: Author's own creation

In summary, the overall results indicate that in most cases, the ensemble classifier RF always shows superior performance among other ML models with three different dataset treatments, especially RF with CBA dimensionality reduction, which achieved the highest accuracy rate of 98.55 percent, as well as the highest sensitivity, specificity, precision, and F-measure. In other words, for the different dimensionality reduction techniques used, the CBA works best with the ensemble classifier compared to PCA and combining both dimensionality reduction techniques. However, DT outperforms RF when using PCA dimensionality reduction with original sampling and undersampling datasets, with a difference of 2.75 and 0.97 percent in the overall accuracy rate, respectively. Furthermore, the best dataset treatment is oversampling compared to original sampling and undersampling. SMOTE oversampling employed in this study solves the class imbalance problem by not just duplicating the existing dataset but adding new synthetic observations with values close to the minority class. On the contrary, the original dataset permits the imbalanced biased training of ML models, and undersampling may cause a significant reduction of valuable datasets, leading to the deterioration of ML models' performance.

5- Conclusions, Recommendations, and Avenues for Future Research

Fraudulent financial reporting (FFR) is a haunting concern that imposes significant losses on different stakeholders and threatens the growth of business and the achievement of the capital market's sustainable development goals. Meanwhile, timely, accurate detection is fairly challenging for auditors, as the perpetrators occasionally shield themselves through the unidentifiable, hard, and late detection nature of occupational fraud. As such, this study started with the aim of developing a machine learning (ML) model of high classification performance for financial statements fraud (FSF). The study compared the performance of simple and ensemble ML models, namely Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), with imbalanced original datasets, including different financial ratios and labels based on Beneish M-score, as well as balanced oversampling and undersampling datasets using SMOTE and Equal Size Sampling (ESS), respectively. The performance of ML classifiers was evaluated in the following stage after using different data preprocessing techniques, including correlation-based dimensionality reduction, principal component analysis, or both. Various performance evaluation criteria have been reported for each classifier, including Cohen’s kappa, which is generally thought to be a more robust measure for inter-rater agreement for categorical items than simple percent agreement calculation. Other evaluation criteria include accuracy, error rates, sensitivity, specificity, precision, and F-measure for fraudulent, non-fraudulent, and overall classes.

The results are not just significant; they are game-changing for the auditing discipline as they show that the ensemble classifier RF has demonstrated superior performance compared to other simple ML classification models. Moreover, the reported findings provide conclusive evidence of the outperformance of the SMOTE oversampling dataset among original and undersampling datasets. The analysis of different dimensionality reduction techniques employed by this study in the data preprocessing stage has led to a groundbreaking conclusion-correlation-based analysis that works best for the superior FFR classifier. In other words, using correlation-based analysis for dimensionality reduction, the RF classification model with the oversampling dataset is the most appropriate for detecting FFR.

This study has multiple contributions. First, the study has created different ML models that significantly perform better than traditional statistical models developed early in past decades. This is a significant step towards facilitating the probability of FSF detection and developing a true classification analysis rather than traditional causal inference. Second, the study has proposed innovative solutions for the dataset imbalance problem faced when building the models using SMOTE for oversampling and ESS for undersampling. Third, as best acknowledged, this study represents a pioneer attempt to compare the classifiers' performance in the FSF context with different correlation-based and PCA dimensionality reduction techniques. This ensures the identification of the most important variables that play a vital role in detecting the probability FFR by the ML classifiers.

After thorough research and analysis, this study culminates with several detailed and specific practical recommendations that can be implemented for real-world applications. The findings presented in this paper underscore the critical need for fraud examiners and regulators worldwide, including organizations like the Financial Supervisory Authority in Egypt, to overhaul their approaches to detecting fraud. It is imperative for them to incorporate advanced technologies into their fraud detection methods to enhance their effectiveness and keep pace with evolving fraudulent activities. Consequently, auditors also must undergo training in new technological techniques to enhance their ability to detect and report material misstatements, including instances of fraud. This technological training is essential for auditors to adapt to the evolving landscape of financial reporting and maintain the integrity of their audits.

While this study has several limitations, it also opens up a world for future research possibilities. The research only used financial ratios gathered from the interim financial statements and emphasized the FSF using specific ML models. However, this research can be extended in a few ways. Future research could use other financial and non-financial data for FSF detection analysis. Additionally, different FSF types other than misstatements could be used to give further insights, such as restatements and delayed or canceled disclosures. Examining other types of occupational fraud, such as asset misappropriation and corruption, could also lead to significant advancements in the field. Moreover, further research can compare the performance of different simple and complex ML models with the same research procedure, paving the way for more accurate and efficient FSF detection methods.

References

Abozaid, E. M., Elshaabany, M. M., & Diab, A. A. (2020). The impact of audit quality on narrative disclosure: Evidence from Egypt. *Academy of Accounting and Financial Studies Journal*, *24*(1), 1–14.

ACFE. (2012). *Report to the Nations on Occupational Fraud and Abuse - 2012 Global Fraud Study*.

ACFE. (2020). *What is Fraud?* Association of Certified Fraud Examiners (ACFE).

ACFE. (2024). *Occupational Fraud 2024: A Report to the Nations*.

Adams, C. A., Coutts, A., & Harte, G. (1995). Corporate equal opportunities (non-) disclosure. *The British Accounting Review*, *27*(2), 87–108. https://doi.org/10.1006/bare.1994.0005

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402

Ali, A. Al, Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences*, *13*(4), 2272. https://doi.org/10.3390/app13042272

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589. https://doi.org/10.2307/2978933

An, B., & Suh, Y. (2020). Identifying financial statement fraud with decision rules obtained from Modified Random Forest. *Data Technologies and Applications*, *54*(2), 235–255. https://doi.org/10.1108/DTA-11-2019-0208

Anh, N. H., & Linh, N. H. (2016). Using the M-score model in detecting earnings management: evidence from Non-Financial Vietnamese listed companies. *VNU Journal of Science: Economics and Business*, *32*(2), 14–23.

Aviantara, R. (2023). Scoring the financial distress and the financial statement fraud of Garuda Indonesia with «DDCC» as the financial solutions. *Journal of Modelling in Management*, *18*(1), 1–16. https://doi.org/10.1108 /JM2-01-2020-0017

Ayad, O. M., Hegazy, A.-El. F., & Dahroug, A. (2023). A Proposed Model for Loan Approval Prediction Using Explainable Artificial Intelligence. *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, 166–173. https://doi.org/10.1109 /ICICIS58388 .2023.10391163

Becker, J., vom Brocke, J., Heddier, M., & Seidel, S. (2015). In Search of Information Systems (Grand) Challenges. *Business & Information Systems Engineering*, *57*(6), 377–390. https://doi.org/10.1007/s12599-015-0394-0

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), 24–36.

Bhavani, G., & Amponsah, C. T. (2017). M score and Z score for detection of accounting fraud. *Accountancy Business and the Public Interest*, *1*(1), 68–86.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, F. H., Chi, D.-J., & Zhu, J.-Y. (2014). *Application of Random Forest, Rough Set Theory, Decision Tree and Neural Network to Detect Financial Statement Fraud – Taking Corporate Governance into Consideration* (pp. 221–234). https://doi.org/10.1007/978-3-319-09333-8_24

Chung, C. Y., Lee, J., & Park, J. (2014). Are Individual Investors Uninformed? Evidence from Trading Behaviors by Heterogeneous Investors around Unfaithful Corporate Disclosure. *Asia-Pacific Journal of Financial Studies*, *43*(2), 157–182. https://doi.org/10.1111/ajfs.12043

Clarke, B., Fokoue, E., & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning* (1st ed.). Springer New York. https://doi.org/10.1007/978-0-387-98135-2

Dalnial, H., Kamaluddin, A., Sanusi, Z. M., & Khairuddin, K. S. (2014). Accountability in Financial Reporting: Detecting Fraudulent Firms. *Procedia - Social and Behavioral Sciences*, *145*, 61–69. https://doi.org/10.1016/j.sbspro.2014.06.011

Darsono, S. N. A. C., Wong, W.-K., Ha, N. T. T., Jati, H. F., & Dewanti, D. S. (2021). Cultural Dimensions and Sustainable Stock Exchanges Returns in the Asian Region. *Journal of Accounting and Investment*, *22*(1), 133–149. https://doi.org/10.18196/jai.v22i1.10318

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements*. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Drake, P. P., & Fabozzi, F. J. (2012). *Analysis of Financial Statements* (3rd ed.). John Wiley & Sons.

Duan, W., Hu, N., & Xue, F. (2024). The information content of financial statement fraud risk: An ensemble learning approach. *Decision Support Systems*, *182*, 114231. https://doi.org/10.1016/j.dss.2024.114231

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, *90*, 374–393. https://doi.org/10.1016/j.eswa.2017.08.030

El-Diftar, D., & Elkalla, T. (2019). The value relevance of accounting information in the MENA region. *Journal of Financial Reporting and Accounting*, *17*(3), 519–536. https://doi.org/10.1108/JFRA-09-2018-0079

Elsayed, A. A. (2017). Indicators of the Financial Statement Fraud (Red Flags). *SSRN Electronic Journal*, 1–20. https://doi.org/10.2139/ssrn.3074187

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

GAO. (2002). *Financial statement restatements: trend, market impacts, regulatory responses, and remaining challenges.*

Gorunescu, F. (2011). *Data Mining* (Vol. 12). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19721-5

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Halteh, K., Kumar, K., & Gepp, A. (2018). Financial distress prediction of Islamic banks using tree-based stochastic techniques. *Managerial Finance*, *44*(6), 759–773. https://doi.org/10.1108/MF-12-2016-0372

Hasanin, T., Khoshgoftaar, T. M., Leevy, J., & Seliya, N. (2019). Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data. *2019 IEEE Fifth International Conference on Big Data Computing*

Service and Applications (BigDataService), 346–356. https://doi.org/10.1109/BigDataService.2019.00063

Helbig, E. (2016). *Detecting accounting fraud – the case of let's Gowex SA*. LAP LAMBERT Academic Publishing.

Hevner, A., & Chatterjee, S. (2010). *Design Science Research in Information Systems: Theory and Practice* (R. Sharda & S. Vob, Eds.; Vol. 22). Springer Science and Business Media.

Hevner, A., March, S. T., & Park, J. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Horváth, I. (2007). Comparison of three methodological approaches of design research. *The 16th International Conference on Engineering Design. Proceedings of ICED 2007*, 1–11.

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 1–3.

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

Jan, C. (2018). An Effective Financial Statements Fraud Detection Model for the Sustainable Development of Financial Markets: Evidence from Taiwan. *Sustainability*, *10*(2), 513. https://doi.org/10.3390/su10020513

Jan, C.-L. (2021). Detection of Financial Statement Fraud Using Deep Learning for Sustainable Development of Capital Markets under Information Asymmetry. *Sustainability*, *13*(17), 9879. https://doi.org/10.3390/su13179879

Kanapickienė, R., & Grundienė, Ž. (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Sciences*, *213*, 321–327. https://doi.org/10.1016/j.sbspro.2015.11.545

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An Overview of Principal Component Analysis. *Journal of Signal and Information Processing*, *04*(03), 173–175. https://doi.org/10. 4236/jsip.2013.43B031

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, *62*, 32–43. https://doi.org/10.1016/j.eswa.2016.06.016

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kukreja, G., Gupta, S. M., Sarea, A. M., & Kumaraswamy, S. (2020). Beneish M-score and Altman Z-score as a catalyst for corporate fraud detection. *Journal of Investment Compliance*, *21*(4), 231–241. https://doi.org /10.1108/JOIC-09-2020-0022

Lakshmi, G., Saha, S., & Bhattarai, K. (2021). Does corruption matter for stock markets? The role of heterogeneous institutions. *Economic Modelling*, *94*, 386–400. https://doi.org/10.1016/j.econmod.2020.10.011

Lehenchuk, S., Mostenska, T., Tarasiuk, H., Polishchuk, I., & Gorodysky, M. (2021). Financial Statement Fraud Detection of Ukrainian Corporations on the Basis of Beneish Model. In B. Alareeni, A. Hamdan, & I. Elgedawy (Eds.), *The Importance of New Technologies and Entrepreneurship in Business Development: In The Context of Economic Diversity in Developing Countries* (Vol. 194, pp. 1341–1356). Springer, Cham. https://doi.org/10.1007/978-3-030-69221-6_100

Li, L., & Wu, D. (2022). Forecasting the risk at infractions: an ensemble comparison of machine learning approach. *Industrial Management & Data Systems*, *122*(1), 1–19. https://doi.org/10.1108/IMDS-10-2020-0603

Lin, C.-C., Chiu, A.-A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*, 459–470. https://doi.org/10.1016/j.knosys.2015.08.011

Liou, F. (2008). Fraudulent financial reporting detection and business failure prediction models: a comparison. *Managerial Auditing Journal*, *23*(7), 650–662. https://doi.org/10.1108/02686900810890625

Liu, C., Chan, Y., Kazmi, S. H. A., & Fu, H. (2015). Financial Fraud Detection Model: Based on Random Forest. *International Journal of Economics and Finance*, *7*(7), 178–188.

Loh, W.-Y., & Shih, Y.-S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*, *7*(4), 815–840.

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. https://doi.org/10.1016/j.ins.2013.07.007

Luo, C. (2019). A comprehensive decision support approach for credit scoring. *Industrial Management & Data Systems*, *120*(2), 280–290. https://doi.org/10.1108/IMDS-03-2019-0182

MacCarthy, J. (2017). Using Altman Z-score and Beneish M-score Models to Detect Financial Fraud and Corporate Failure: A Case Study of Enron Corporation. *International Journal of Finance and Accounting*, *6*(6), 159–166.

Mandal, A., & S, A. (2023). Preventing financial statement fraud in the corporate sector: insights from auditors. *Journal of Financial Reporting and Accounting*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JFRA-02-2023-0101

Marais, A., Vermaak, C., & Shewell, P. (2023). Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models. *Cogent Economics & Finance*, *11*(1). https://doi.org/10.1080/23322039.2023.2190215

Mbona, R. M., & Yusheng, K. (2019). Financial statement analysis. *Asian Journal of Accounting Research*, *4*(2), 233–245. https://doi.org/10.1108/AJAR-05-2019-0037

Mehta, A., & Bhavani, G. (2017). Application of forensic tools to detect fraud: the case of Toshiba. *Journal of Forensic and Investigative Accounting*, *9*(1), 692–710.

Mehta, M., Agrawal, R., & Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In P. Apers, M. Bouzeghoub, & G. Gardarin (Eds.), *Advances in Database Technology - EDBT'96* (1st ed., pp. 18–32). Springer. https://doi.org/10.1007/BFb0014141

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

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Mousa, G. A., Elamir, E. A. H., & Hussainey, K. (2022). Using machine learning methods to predict financial performance: Does disclosure tone matter? *International Journal of Disclosure and Governance*, *19*(1), 93–112. https://doi.org/10.1057/s41310-021-00129-x

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

Omar, N., Johari, Z. 'Amirah, & Smith, M. (2017). Predicting fraudulent financial reporting using artificial neural network. *Journal of Financial Crime*, *24*(2), 362–387. https://doi.org/10.1108/JFC-11-2015-0061

Papík, M., & Papíková, L. (2020). Detection models for unintentional financial restatements. *Journal of Business Economics and Management*, *21*(1), 64–86. https://doi.org/10.3846/jbem.2019.10179

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice & Theory*, *30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

PwC. (2020). *Fighting Fraud: A Never-Ending Battle PwC's Global Economic Crime and Fraud Survey.*

Rahman, M. J., & Zhu, H. (2024). Detecting accounting fraud in family firms: Evidence from machine learning approaches. *Advances in Accounting*, *64*, 100722. https://doi.org/10.1016/j.adiac.2023.100722

Ravisankar, P., Ravi, V., Rao, R. G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500. https://doi.org/10.1016/j.dss.2010.11.006

Rezaee, Z. (2002). *Financial Statement Fraud: Prevention and Detection* (J. T. Wells, Ed.). John Wiley and Sons.

Rezaee, Z. (2005). Causes, consequences, and deterence of financial statement fraud. *Critical Perspectives on Accounting*, *16*(3), 277–298. https://doi.org/10.1016/S1045-2354(03)00072-8

Riskiyadi, Moh. (2023). Detecting future financial statement fraud using a machine learning model in Indonesia: a comparative study. *Asian Review of Accounting*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/ARA-02-2023-0062

Saleh, M. M. A., Aladwan, M., Alsinglawi, O., & Almari, M. O. S. (2021). Predicting fraudulent financial statements using fraud detection models. *Academy of Strategic Management*, *20*(3), 1–17.

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making*, *21*(S2), 57. https://doi.org/10.1186 /s12911-021-01423-y

Song, X., Hu, Z., Du, J., & Sheng, Z. (2014). Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China. *Journal of Forecasting*, *33*(8), 611–626. https://doi.org/10.1002/for.2294

Tang, X.-B., Liu, G.-C., Yang, J., & Wei, W. (2018). Knowledge-based Financial Statement Fraud Detection System: Based on an Ontology and a Decision Tree. *Knowledge Organization*, *45*(3), 205–219.

Trading Economics. (2024). *Egypt Corruption Index*. Trading Eonomics.

Transparency International. (2023). *Corruption Perceptions Index 2023*.

Vousinas, G. L. (2019). Advancing theory of fraud: the S.C.O.R.E. model. *Journal of Financial Crime*, *26*(1), 372–381. https://doi.org/10. 1108/JFC-12-2017-0128

Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, *167*, 113913. https://doi.org/10.1016/j.dss. 2022.113913

West, J., & Bhattacharya, M. (2015). Mining financial statement fraud: An analysis of some experimental issues. *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, 461–466. https://doi.org/10.1109/ICIEA.2015.7334157

Witten, I. H., Frank, E., Hall, M. H., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufman.

Xiuguo, W., & Shengyong, D. (2022). An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. *IEEE Access*, *10*, 22516–22532. https://doi.org/10.1109/ACCESS .2022.3153478

Xu, X., Xiong, F., & An, Z. (2023). Using Machine Learning to Predict Corporate Fraud: Evidence Based on the GONE Framework. *Journal of Business Ethics*, *186*(1), 137–158. https://doi.org/10.1007/s10551-022-05120-2

Zaki, M. J., & Jr, W. M. (2020). *Data mining and Machine Learning: Fundamental Concepts and Algorithms* (2nd ed.). Cambridge University Press.

Zhang, A. (2012). An Examination of the Effects of Corruption on Financial Market Volatility. *Journal of Emerging Market Finance*, *11*(3), 301–322. https://doi.org/10.1177/0972652712466501

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575. https://doi.org /10.1016/j.dss.2010.08.007