IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

Financial Statement Fraud Detection by Machine Learning and Data Mining

Mrs. Yaazhini[1], Boopathiraja P[2], Manohar P[3], Vijay Rs[4], Viknesh R[5]

Assistant Professor, Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India[1]

Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India[2,3,4,5]

ABSTRACT: Financial fraud, considered as deceptive tactics for gaining financial benefits, has recently become a widespread menace in companies and organizations. Conventional techniques such as manual verifications and inspections are imprecise, costly, and time consuming for identifying such fraudulent activities. With the advent of artificial intelligence, machine-learning-based approaches can be used intelligently to detect fraudulent transactions by analyzing a large number of financial data. Therefore, this paper attempts to present a systematic literature review (SLR) that systematically reviews and synthesizes the existing literature on machine learning (ML)-based fraud detection. Particularly, the review employed the Kitchenham approach, which uses well-defined protocols to extract and synthesize the relevant articles; it then report the obtained results. Based on the specified search strategies from popular electronic database libraries, several studies have been gathered. After inclusion/exclusion criteria, 93 articles were chosen, synthesized, and analyzed. The review summarizes popular ML techniques used for fraud detection, the most popular fraud type, and evaluation metrics. The reviewed articles showed that support vector machine (SVM) and artificial neural network (ANN) are popular ML algorithms used for fraud detection, and credit card fraud is the most popular fraud type addressed using ML techniques. The paper finally presents main issues, gaps, and limitations in financial fraud detection areas and suggests possible areas for future research.

KEYWORDS: financial fraud; fraud detection; machine learning; data mining; systematic literature review; Kitchenham approach

I. INTRODUCTION

Financial fraud is the act of gaining financial benefits by using illegal and fraudulent methods [1,2]. Financial fraud can be committed in different areas, such as insurance, banking, taxation, and corporate sectors [3]. Recently, financial transaction fraud [4], money laundering, and other types of financial fraud [5] have become an increasing challenge among companies and industries [4]. Despite several efforts to reduce financial fraudulent activities, its persistence affects the economy and society adversely, as large amounts of money are lost to fraud every day [6]. Several fraud detection approaches were introduced many years ago [1]. Most traditional methods are manual, and this is not only time consuming, costly, and imprecise but also impractical [7]. More studies are conducted to reduce losses resulting from fraudulent activities, but they are not efficient [5]. With the advancement of the artificial intelligence (AI) approach, machine learning and data mining have been utilized to detect fraudulent activities in the financial sector [8,9]. Both unsupervised and supervised methods were employed to predict fraud activities [4,10]. Classification methods have been the most popular method for detecting financial fraudulent transactions. In this scenario, the first stage of model training uses a dataset with class labels and feature vectors. The trained model is then used to classify test samples in the next step.

Contribution: This study aims to establish a superior model for financial fraud prediction using a powerful ensemble technique, the XGBoost (eXtreme Gradient Boosting) algorithm, to help identify fraud on a set of sample companies drawn from the MENA region by spotting the early signs. This approach can help to mitigate losses to investors, auditors, and all stakeholders in the financial market. This research enables researchers and practitioners to gain a deeper understanding and make informed decisions based on a financial statement fraud detection model. We focus on utilizing data from publicly available financial statements of firms in the Middle East and North Africa (MENA) region. We selected a powerful ML ensemble model, namely, the XGBoost algorithm, to model our proposed method

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

for a number of reasons. Ensemble algorithms have been successfully used in many fields of research [12], though they have been less utilized in financial fraud studies.

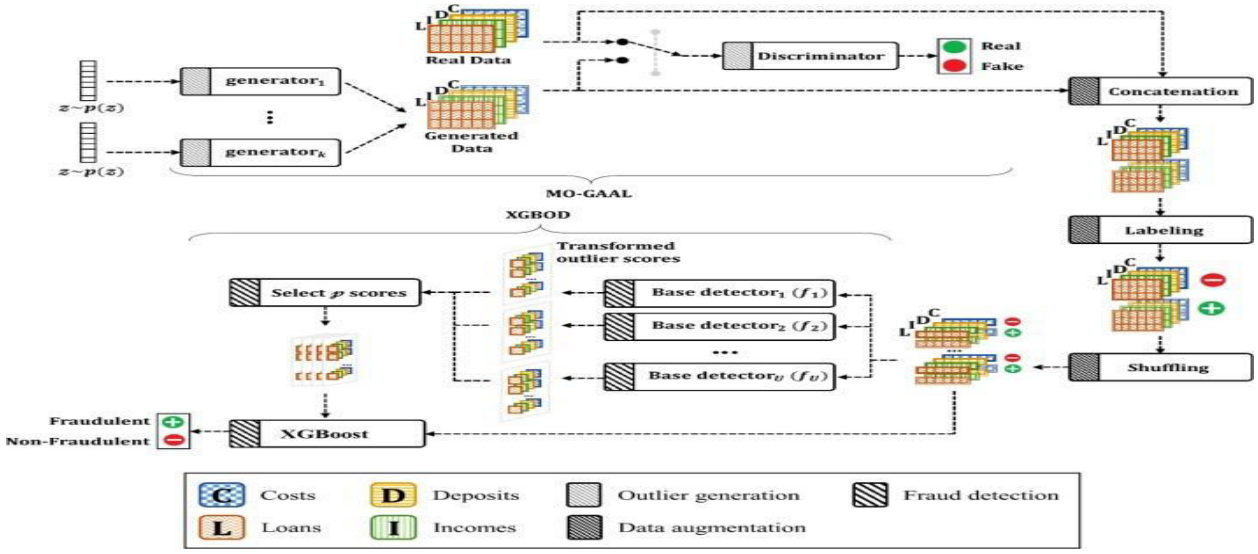

Fig 1: proposal work flow

The characteristics of XGBoost fit very well with our small dataset, which is characterized by many missing values and high class imbalance. XGBoost facilitates the tuning of a range of hyperparameters in order to further improve the efficiency of the model. Because of the high class imbalance in this one-of-a-kind domain, the samples selected in past studies have a tendency to be processed less realistically. XGBoost is able to effectively learn from imbalanced data with the help of Synthetic Minority Oversampling Technique (SMOTE) sampling. We chose expert-defined financial ratios [13] along with raw financial data for our research. FSF detection models based on financial ratios may be more effective, as the ratios determined by domain experts are mostly based on assumptions that provide a strong prediction of situations in which corporate managers are encouraged to commit fraud [14]. Fernandez-Delgado, Cernadas, Barro, and Amorim [15] demonstrate that there may be no single right model that can be applied to all data environments; therefore, there is uncertainty about whether ensemble algorithms perform better than conventional financial fraud detection methods in our particular context. Hence, we selected five other ML techniques that are widely used in this area and modeled them for performance analysis and comparison with our model: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), AdaBoost, and Random Forest (RF). We think that the results of this study can be useful to investors, regulators, stock exchanges, boards of directors, external auditors, and other key stakeholders as they seek to prevent, deter, and detect fraudulent financial reporting. Benford's Law is an effective method and analytical technique to help detect accounting fraud; see, for instance, [16,17,18]. Benford's Law is a mathematical tool used to determine whether investigated financial statements contain unintentional errors or fraud. When using Benford's law, counterfeit numbers have a slightly different pattern than valid or random samples.

II. RELATED RESEARCH

Financial Statement Fraud (FSF) detection has been under the limelight for the past few decades, with emphasis placed on accounting anomalies broadly and on financial statement fraud specifically. While initial research made use of statistical or traditional techniques that are both time-consuming and expensive, more recently the focus has drifted with the emergence of big data and ML [19]. The statistical approaches are centered on traditional mathematical methods, while methods involving ML are focused on learning and intelligence. While both categories have many similarities, the key difference between them is that statistical methods are more rigid, while ML methods are able to learn from and adapt to the problem domain.

In addition, previous studies have shown the superior efficiency of ML approaches over conventional statistical approaches. According to the literature, there is no one-size-fits-all strategy for detecting financial statement fraud. The findings of show that models built using ML approaches can efficiently detect financial statement fraud, keep up with

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

the continuous evolution of financial reporting fraudulent behavior, and respond with the most up-to-date technology. In, the authors demonstrate that detection models developed using ML approaches are more accurate than traditional methods. Therefore, our review of past research is limited to papers that have used only ML techniques for FSF detection. Most of the research in the previous literature has formulated the detection of FSF as a binary classification problem, sometimes as a multi-class problem and other times as a clustering problem. Researchers have conducted both quantitative and qualitative FSF analyses. Text mining has been used extensively for qualitative research. Here, we focus on papers that perform quantitative analysis using ML techniques. In the initial stages, research mainly included the Neural Network (NN).

In the case of fraud detection problems the minority class needs to receive special consideration, as it defines the phenomena which we aim to anticipate from a multitude of majority class structures that reflect correct processes. The performance of standard classifiers is biased towards the majority class, as they are programmed to minimize the overall inaccuracy of classification regardless of class distribution. This bias problem can be overcome by excluding examples of the majority class, known as undersampling, or by including new examples of the minority class, known as oversampling. Due to the small size of our dataset, we chose the latter.

DTs are among the most successful ML algorithms thanks to their intelligibility and clarity. DT approaches are used for estimation, clustering, and classification tasks. The best attribute is placed at the root node. This attribute is selected based on a measure of information gain that is subsequently used at each stage of tree building. The training dataset is then split into two or more subsets based on the values of the chosen attribute. This process is repeated for each of the subsets, selecting the best attribute for each and creating child nodes. The process continues until a stopping criterion is met, such as reaching a maximum depth or all instances belonging to the same class. The leaf nodes in the decision tree reflect the class, while the decision nodes determine the rules. The test data class is predicted by the decision rules. Among the key benefits of Decision Trees are that they offer a model that meaningfully describes the acquired knowledge and enables the extraction of if–then classification rules

III. METHODS

Random Forest (RF) is a bagging ensemble technique proposed particularly for trees. The base model of an RF is a DT. RF addresses the issue of high variance in DTs by combining the predictions of multiple DTs, with each DT trained on a different subset of the data and a different subset of the features. The resulting combination of trees leads to a more robust and less variable model. Random subsets of the data are generated via replacement, and each subset is trained with the help of a DT. While expanding the trees, RF adds more randomness to the structure. As a result, there is a wide range of diversity, which contributes to a successful model in general. As a result, in an RF the algorithm only considers a random subset of the features when dividing a node. The randomness of the trees can be increased by using additional random thresholds for every feature, instead of looking for the highest suitable thresholds as in a normal DT.

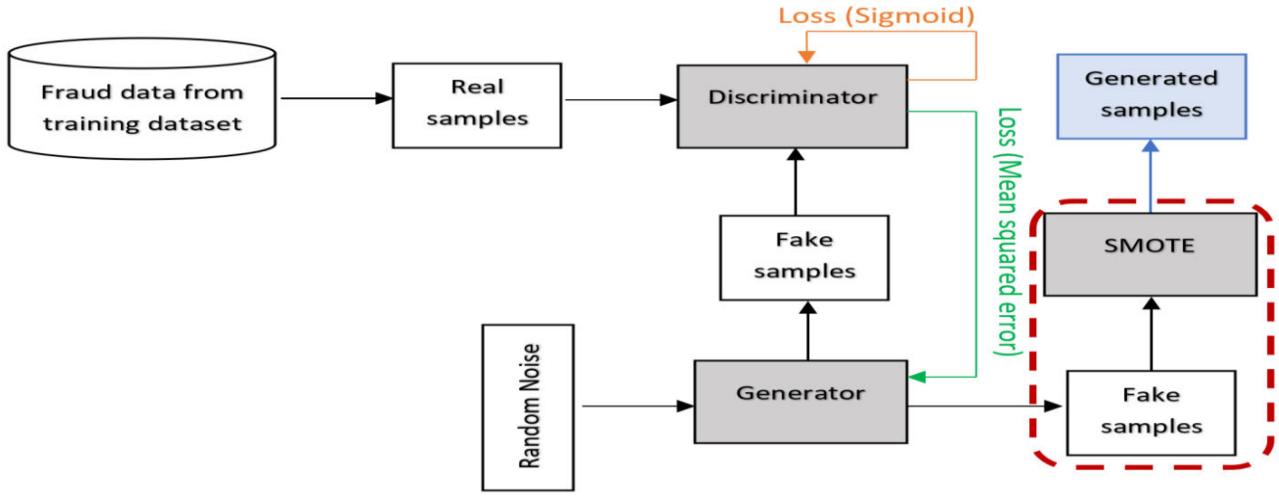

Fig 2: Fake Detection

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

AdaBoost, or Adaptive boosting, was the first really successful boosting algorithm developed for binary classification. It is an approach to minimize the error of a weak learning algorithm. Theoretically, any algorithm can be a weak learning algorithm if it can produce classifiers that need to be slightly more consistent than random guessing. AdaBoost helps to combine multiple "weak classifiers" into a single "strong classifier". The most common algorithms used with AdaBoost are DTs. A weak classifier (decision stump) is prepared using weighted samples from the training data. Only binary (two-class) classification problems are supported; thus, each decision stump makes a decision on one input variable and outputs a value of +1 or −1 for the first or second class. Weak models are sequentially added and trained with weighted training data. The method progresses until a predetermined number of weak learners has been generated or no more improvements can be achieved on the training dataset.

XGBoost, or eXtreme Gradient Boosting, is a tree-based algorithm. XGBoost has shown great success in terms of both performance and speed. Boosting is an ensemble strategy with the key goal of reducing bias and variance. The aim is to sequentially build weak trees in such a way that each new tree (or learner) works on the flaw (misclassified data) of the preceding tree. The data weights are re-adjusted, known as "re-weighting", whenever a weak learner is added. Because of this auto-correction after every new learner is introduced, the whole forms a strong model after convergence. The loss function of the model is characterized as penalizing the complexity of the model with regularization in order to decrease the possibility of overfitting. This technique performs well even when there are missing values or many zero values, demonstrating a good ability to deal with sparsity. XGBoost uses an algorithm called the "weighted quantile sketch algorithm" that helps the classifier to concentrate on data that are incorrectly classified. In each iteration, the aim of each new learner is to learn how to classify the incorrect data.

IV. RESULT ANALYSIS

In this report, the problem of financial statements data fraud data detection is formulated as Time Series Classification problems. To handle the problems, RNN based approaches had been implemented using real data sets from US Security and Exchange Commission (SEC). First, the three main RNN based approaches Simple RNN, LSTM and GRU were reviewed. Then, some experimental settings are designed to compare performances of Simple RNN, LSTM and GRU using three types of hidden unit which include: 6, 10 and 14 in the experiments. The evaluation of performances is evaluated using training and test accuracy as well as loss and accuracy curves as visualization tools. Overall, the results show that when the hidden unit number increased the performances of the model improved. The GRU model, except for 10 hidden unit number, consistently outperforms Simple RNN and LSTM. In contrast, the results show that LSTM model to be the worst for all types of hidden unit number. From above results, the GRU model or Simple RNN model is a recommended approach for short time series data, the LSTM is not suitable for this type of time series data. However, the theoretical research of the proposed methods needed to be explored in the future.

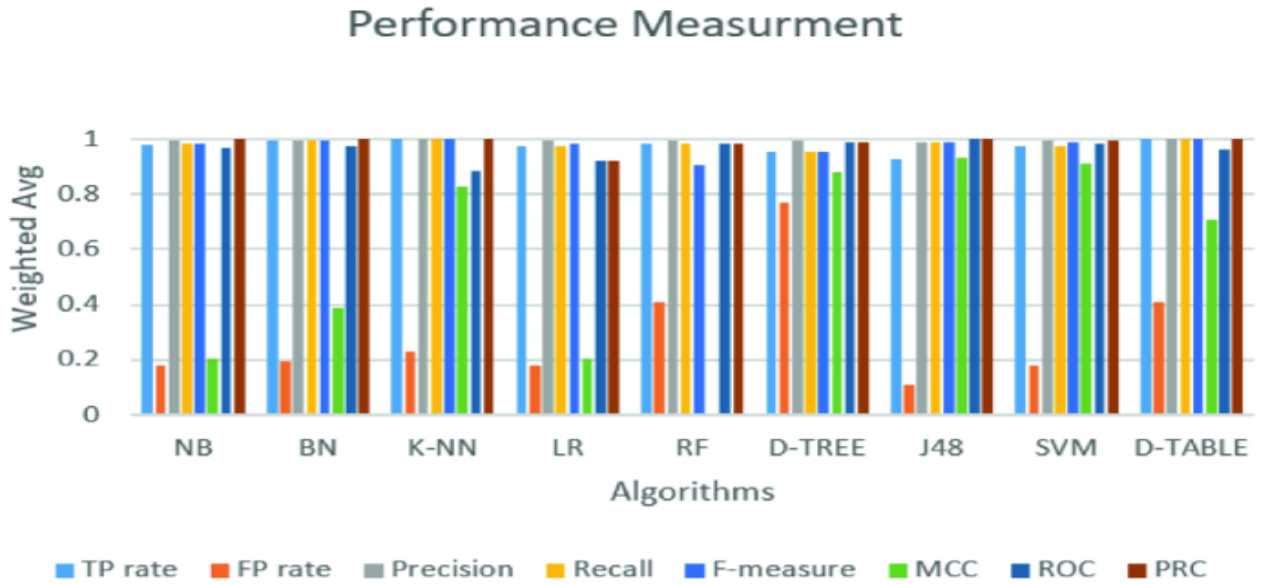

Fig 3: Result Analysis

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

The financial statements data collected from Accounting and Auditing Enforcement Releases (AAERs), which are federal materials issued by US Security and Exchange Commission (SEC) at https://www.sec.gov/ for alleging accounting and/or auditing misconduct in U.S.-listed companies. The twenty-three companies are selected such that they have full yearly report K-10 form between year 2012 to 2020. Among twenty-three companies there are seven companies confirmed to have fraud issues. For the analysis purposes, the financial statements will be arranged into time series data structure. The next step, the variables will be used to selected. Those variables which are available for all selected companies including: Cash and Cash Equivalents at Carrying Value, Assets Current, Liabilities Current, Liabilities And Stockholders Equity,

V. CONCLUSION

Financial Fraud can be committed in different financial aspects such as insurance, banking, taxation, and corporate sectors [3]. Recently, financial fraud has become increasingly worrisome among companies and industries [4]. Despite several efforts to eradicate financial fraud, its persistence adversely affects the economy and society as very large amounts of money are lost to fraud every day [6]. With the advent of artificial intelligence, machine-learning-based approaches can be used intelligently to detect fraudulent transactions by analyzing a large number of financial data. In this paper, we presented a study that systematically reviewed and synthesized the existing literature on ML-based fraud detection. In particular, this paper adopted the Kitchenham methodology, which uses well-defined protocols to extract, synthesize, and report results. Several studies have been gathered based on the specified search strategies for popular electronic libraries. After the inclusion/exclusion criteria, 87 were selected. In this review, popularly used ML techniques for fraud detection, the most common fraud type, and the evaluation metrics are summarized. Based on the reviewed articles, results showed that SVM and NN are the popular ML algorithms used for fraud, and credit card fraud is the most popular fraud type in the literature. The paper finally presented the key issues, gaps, and limitations in the area of financial fraud detection and suggests areas for future research. We identified gaps in the research by examining unexplored or less studied algorithms. Previous studies in financial fraud detection focused on supervised classification and regression methods, such as SVM, neural networks, and logistic regression. The use of ensemble methods that take advantage of multiple algorithms to classify samples is a rising trend in the field. Interestingly, we discovered that unsupervised learning approaches, such as clustering, were less employed in the present literature. Clustering is beneficial for investigating latent relations and resemblances. In addition, since there are a small number of fraud cases that have to be identified, clustering could be effective. We recommend that future studies pay more attention to unsupervised practices, such as anomaly detection, which can uncover new insights. Additionally, another avenue for future research would be to use emerging text-mining techniques and word-embedding techniques such as Word2Vec, Doc2Vec, or BERT to transform financial texts into vectors of features, which will then be used to build machine learning models.

REFERENCES

1. Hilal, W.; Gadsden, S.A.; Yawney, J. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Syst. Appl.* **2021**, *193*, 116429. [Google Scholar] [CrossRef]
2. Ashtiani, M.N.; Raahemi, B. Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access* **2021**, *10*, 72504–72525. [Google Scholar] [CrossRef]
3. Albashrawi, M. Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. *J. Data Sci.* **2016**, *14*, 553–570. [Google Scholar] [CrossRef]
4. Choi, D.; Lee, K. An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation. *Secur. Commun. Netw.* **2018**, *2018*, 1–15. [Google Scholar] [CrossRef]
5. Ngai, E.W.T.; Hu, Y.; Wong, Y.H.; Chen, Y.; Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **2011**, *50*, 559–569. [Google Scholar] [CrossRef]
6. Ryman-Tubb, N.F.; Krause, P.; Garn, W. How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Eng. Appl. Artif. Intell.* **2018**, *76*, 130–157. [Google Scholar] [CrossRef]
7. Al-Hashedi, K.G.; Magalingam, P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Comput. Sci. Rev.* **2021**, *40*, 100402. [Google Scholar] [CrossRef]

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303269 |

8. Chaquet-ulldemolins, J.; Moral-rubio, S.; Muñoz-romero, S. On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders. *Appl. Sci.* **2022**, *12*, 3856. [Google Scholar] [CrossRef]
9. Da'U, A.; Salim, N. Recommendation system based on deep learning methods: A systematic review and new directions. *Artif. Intell. Rev.* **2019**, *53*, 2709–2748. [Google Scholar] [CrossRef]
10. Zeng, Y.; Tang, J. RLC-GNN: An Improved Deep Architecture for Spatial-Based Graph Neural Network with Application to Fraud Detection. *Appl. Sci.* **2021**, *11*, 5656. [Google Scholar] [CrossRef]
11. Delamaire, L.; Hussein, A.; John, P. Credit card fraud and detection techniques: A review. *Banks Bank Syst.* **2009**, *4*, 57–68. [Google Scholar]
12. Zhang, D.; Zhou, L. Discovering Golden Nuggets: Data Mining in Financial Application. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2004**, *34*, 513–522. [Google Scholar] [CrossRef]
13. Raj, S.B.E.; Portia, A.A. Analysis on credit card fraud detection methods. In Proceedings of the 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), Tirunelveli, India, 18–19 March 2011; pp. 152–156. [Google Scholar] [CrossRef]
14. Phua, C.; Lee, V.; Smith, K.; Gayler, R. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *arXiv* **2010**, arXiv:1009.6119. [Google Scholar]
15. West, J.; Bhattacharya, M. Intelligent financial fraud detection: A comprehensive review. *Comput. Secur.* **2016**, *57*, 47–66. [Google Scholar] [CrossRef]
16. Abdallah, A.; Maarof, M.A.; Zainal, A. Fraud detection system: A survey. *J. Netw. Comput. Appl.* **2016**, *68*, 90–113. [Google Scholar] [CrossRef]
17. Popat, R.R.; Chaudhary, J. A Survey on Credit Card Fraud Detection Using Machine Learning. In Proceedings of the 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 11–12 May 2018; pp. 1120–1125. [Google Scholar]
18. Gyamfi, N.K.; Abdulai, J. Bank Fraud Detection Using Support Vector Machine. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 37–41. [Google Scholar]
19. Carneiro, E.M.; Dias, L.A.V.; Da Cunha, A.M.; Mialaret, L.F.S. Cluster Analysis and Artificial Neural Networks: A Case Study in Credit Card Fraud Detection. In Proceedings of the 2015 12th International Conference on Information Technology-New Generations, Mumbai, India, 11–14 December 2011; pp. 122–126. [Google Scholar] [CrossRef]
20. Iyer, D.; Mohanpurkar, A.; Janardhan, S.; Rathod, D.; Sardeshmukh, A. Credit card fraud detection using Hidden Markov Model. In Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, 11–14 December 2011; pp. 1062–1066. [Google Scholar]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details