ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

Farud Detection in Banking Data by Machine Learning Techniques

P. Siva Srinivasarao[1], Syed vahid[2], Bandla Sai Ranganadh[3], Bandla Sai Ranganadh[4], Shaik Mohammad Kafeel Basha[5]

[1] Assistant Professor, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

[2,3,4,5] Students, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

Email id: thrishiva123@gmail.com[1], pinnikasusmitha@gmail.com[2], syedvahid8185@gmail.com[3], ranganadhjul8@gmail.com[4], Kafeelsmd@gmail.com[5]

Detecting fraudulent transactions in the cybersecurity industry is crucial, especially in banking, where legitimate transactions vastly outnumber fraudulent ones, creating highly imbalanced datasets. This study compares the effectiveness of two approaches, random under-sampling and Synthetic Minority Over-sampling Technique, to evaluate the accuracy of fraud detection in such datasets. Random under-sampling excludes instances from the majority class, improving recall but compromising precision. Conversely, SMOTE generates synthetic examples for the minority class, yielding a balanced F1 score and improved overall accuracy but with slightly reduced recall. To further enhance fraud detection performance, multiple machine learning models were evaluated to balance recall, precision, and F1 scores effectively. This research provides insights into the nuances of under-sampling and oversampling in fraud detection, guiding cybersecurity professionals in selecting techniques that best suit their organizational needs. It underscores the importance of addressing class imbalances and emphasizes the role of ongoing model refinement in achieving reliable fraud detection.

Keywords: Fraud detection**,** Cybersecurity**,** Banking data**,** Imbalanced dataset**,** Random under-sampling

1.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 [1,2,5].

Thus, this study attempts to identify machine-learning-based techniques employed for financial transaction fraud and to analyse gaps to discover research trends in this area. Recently, some reviews have been conducted to detect fraudulent financial activities [11,12,13]. For instance, Delamaire et al. [11] conducted a review on different categories of fraudulent activities on credit cards, which include bankruptcy and counterfeit frauds, and suggested proper approaches to address them. Similarly, Zhang and Zhou [12] investigated ML methods for fraud transactions, which include the stock market and other fraud detection processes in financial sectors. Raj and Portia. [13] explored several ML approaches used for credit card fraud detection. Phua et al. [14] conducted a comprehensive survey to explore data mining and machine learning techniques to detect frauds in various aspects, including credit card fraud, insurance fraud, and telecoms subscription fraud.

Recently, there has been a significant increase in fraud activities in health sectors [15]. Abdallah et al. [16] introduced a review to investigate different approaches for uncovering fraudulent activities in the health care domain based on statistical approaches. Popat and Chaudhary [17] presented an extensive review work on credit card fraud detection. The authors provide a detailed analysis of various ML classification methods with their methodology and challenges. Ryman-Tubb et al. [6] reviewed several state-of-the-art methods for detecting payment card fraudulent activities using transactional volumes. The study showed that only eight approaches have a practical implication to be used in the industry. A study by Albashrawi and Lowell [3] analyzed several studies for one decade covering fraud detection in financial sectors using data mining techniques. However, this was not exhaustive and comprehensive enough as they ignored the method of evaluations and the pros and cons of data mining techniques, among others.

Despite several existing reviews in the field, however, most studies particularly focused on specific areas of finance, such as detecting credit card fraudulent activities [18], fraud in online banking [19], fraud in bank credit administration [20], and fraud in payment cards [21]. Hence, there is a need of a study that encompasses all popular areas of financial fraud activities to fill the gap in this aspect. More recently, a study was published to review fraud-detection methods in financial records [2]. The authors integrated the prior multi-disciplinary literature on financial statement fraud. However, there are several differences between their work and our review. First, their primary objective is to integrate research from several fields, including information systems, analytics, and accounting. On the other hand, we aim to identify financial fraud transactions based on machine learning methods and to discover datasets applied in the ML-based financial fraud detection. Furthermore, we considered conference articles in our study while they did not. This study reviews existing machine learning (ML)-based methods applied for financial transaction fraud detection. Furthermore, the SLR can guide researchers in their choice of applying ML-based financial transaction fraud-detection methods along with the datasets to be used for predicting fraudulent activities in financial transactions.

2. Existing System

Halvaiee & Akbari study a newmodel called the AIS-based fraud detection model (AFDM). They use the Immune System Inspired Algorithm (AIRS) to improve fraud detection accuracy. The presented results of their paper show that their proposed AFDM improves accuracy by up to 25%, reduces costs by up to 85%, and reduces system response time by up to 40% compared to basic algorithms [11]. Bahnsen et al. developed a transaction aggregation strategy and created a new set of features based on the periodic behaviour analysis of the transaction time by using the von Mises distribution. In addition, they propose a new cost-based criterion for evaluating credit card fraud detection's models and then, using a real credit card dataset, examine how different feature sets affect results. More precisely, they extend the transaction aggregation strategy to create new offers based on an analysis of the periodic behaviour of transactions [12]. Randhawa et al. study the application of machine learning algorithms to detect fraud in credit cards. They _rst use Naïve Bayes, stochastic forest and decision trees, neural networks, linear regression (LR), and logistic regression, as well as support vector machine standard models, to evaluate the available datasets. Further, they propose a hybrid method by applying AdaBoost and majority voting. In addition, they add noise to the data samples for robustness evaluation. They perform experiments on publicly available datasets and show that majority voting is effective in detecting credit card fraud cases [6]. Porwal and Mukund propose an approach that uses clustering methods to detect outliers in a large dataset and is resistant to changing patterns.

3. Proposed System

The system proposes an efficient approach for detecting credit card fraud that has been evaluated on publicly available datasets and has used optimized algorithms SVM and logistic regression individually, as well as majority voting combined methods, as well as deep learning and hyper parameter settings. An ideal fraud detection system should detect more fraudulent cases, and the precision of detecting fraudulent cases should be high, i.e., all results should be correctly detected, which will lead to the trust of customers in the bank, and on the other hand, the bank will not suffer losses due to incorrect detection. propose a group learning framework based on partitioning and clustering of the training set. Their proposed framework has two goals: 1) to ensure the integrity of the sample features, and 2) to solve the high imbalance of the dataset.

4. METHODOLOGY

The methodology encompasses a range of machine learning techniques tailored for diverse data analysis tasks. Decision tree classifiers partition data recursively to capture decision-making logic effectively. Gradient boosting iteratively improves model performance by optimizing a loss function through sequential addition of weak learners, typically decision trees. K-Nearest Neighbors (KNN) classifies data points by measuring similarities with neighboring instances, employing lazy learning for efficient classification without explicit training. Logistic regression models categorical outcomes by estimating probabilities based on independent variables, offering simplicity and interpretability. Naïve Bayes assumes feature independence to classify data efficiently, particularly useful for large datasets despite its assumption's simplicity. Random Forest mitigates overfitting by aggregating predictions from multiple decision trees, enhancing accuracy in both classification and regression tasks. Support Vector Machines (SVM) find optimal hyperplanes to separate classes in complex feature spaces, providing robust classification capabilities. This methodology integrates these techniques to empower data scientists in addressing a wide array of predictive modeling challenges across various domains effectively.

Architecture:

The architecture diagram illustrates the components and data flow of a credit card fraud detection system. At the core of the system is the Service Provider, which allows users to log in, browse, and train/test credit card data sets. Users can view the accuracy of these datasets through bar charts and detailed accuracy results. The Service Provider also enables users to view and download predictions of credit card fraud, including detection ratios and results. Additionally, it allows users to see all remote users within the system. Remote users can register, log in, predict the type of credit card fraud, and view their profile information. The Web Server plays a crucial role by accepting all information from users, storing dataset results, and processing user queries. It interacts closely with the Web Database, which stores and retrieves all necessary data, ensuring secure storage and access.The data flow begins with user interaction with the Service Provider, where they perform various actions such as logging in, browsing data sets, and viewing results. The Service Provider manages these interactions and requests data from the Web Server. The Web Server processes this information, accessing and storing data as needed from the Web Database. This architecture ensures efficient processing of user queries, secure data storage, and accessible predictions, providing a robust system for credit card fraud detection.

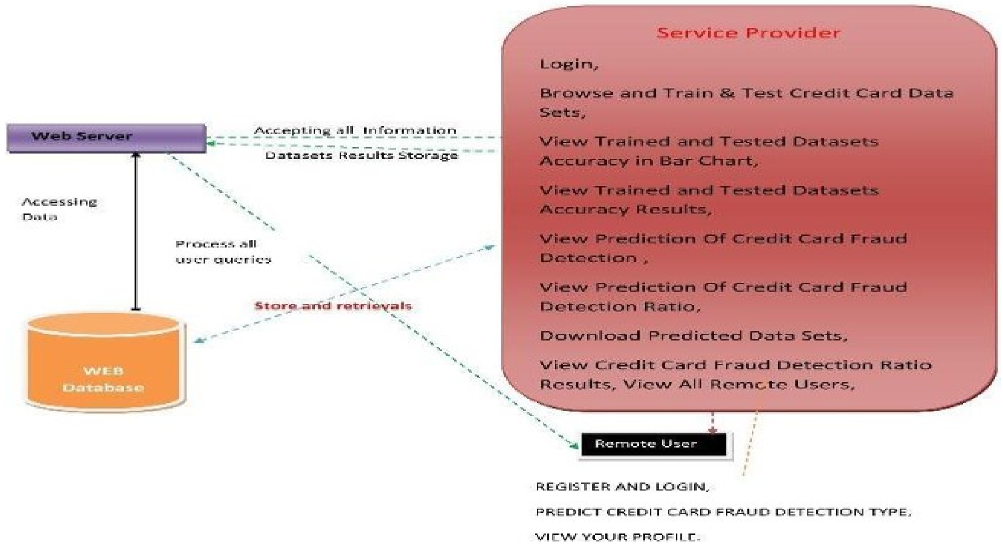

Service Provider
Login,
Browse and Train & Test Credit Card Data Sets,
View Trained and Tested Datasets Accuracy in Bar Chart,
View Trained and Tested Datasets Accuracy Results,
View Prediction Of Credit Card Fraud Detection ,
View Prediction Of Credit Card Fraud Detection Ratio,
Download Predicted Data Sets,
View Credit Card Fraud Detection Ratio Results, View All Remote Users,
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT CREDIT CARD FRAUD DETECTION TYPE,
VIEW YOUR PROFILE.

Figure: Architecture

DATASET

In this paper, we use a real dataset so that the outcome of the proposed algorithm can be used in practice. We consider a dataset named “credit card” that contains 284,807 records of two days of transactions made by credit card holders in September 2013. There are 492 fraudulent transactions, and the rest of the transactions are legitimate. The positive class(frauds) accounts for 0.172% of all transactions; hence, the dataset is highly imbalanced. This dataset is available and can be accessed through https://www.kaggle.com/mlg-ulb/creditcardfraud.This dataset contains only numerical input variables resulting from a principle component analysis (PCA) transformation. Unfortunately, the original features and background information about the data are not given due to confidentiality and privacy considerations. PCA yielded the following principal components: V1, V2, V28. The untransformed features with PCA are "time" and "amount." The "Time" column contains the time (in seconds) elapsed between each trans-action and the first transaction in the dataset

Algorithm: The architecture diagram illustrates the components and data flow of a credit card fraud detection system. At the core of the system is the Service Provider, which allows users to log in, browse, and train/test credit card data sets. Users can view the accuracy of these datasets through bar charts and detailed accuracy results. The Service Provider also enables users to view and download predictions of credit card fraud, including detection ratios and results. Additionally, it allows users to see all remote users within the system. Remote users can register, log in, predict the type of credit card fraud, and view their profile information. by accepting all information from users, storing dataset results, and processing user queries. It interacts closely with the Web Database, which stores and retrieves all necessary data, ensuring secure storage and access. The data flow begins with user interaction with the Service Provider, where they perform various actions such as logging in, browsing data sets, and viewing results. The Service Provider manages these interactions and requests data from the Web Server. The Web Server processes this information, accessing and storing data as needed from the Web Database. This architecture ensures efficient processing of user queries, secure data storage, and accessible predictions, providing a robust system for credit card fraud detection.

Decision tree classifiers are effectively used in various fields due to their ability to capture descriptive decision- making knowledge from the supplied data. These classifiers can be generated from training sets, where the procedure involves testing and partitioning data based on outcomes to recursively build a decision tree. Gradient boosting is another powerful technique used for regression and classification tasks, which builds an ensemble of weak prediction models, typically decision trees, to optimize a differentiable loss function.

K-Nearest Neighbors (KNN) is a simple yet powerful classification algorithm that classifies based on similarity measures and works in a non-parametric and lazy learning manner by finding the K-nearest neighbors from the training data. Logistic regression analysis is used to study the association between a categorical dependent variable and a set of independent variables. It competes with discriminant analysis for modeling categorical-response variables, offering versatility by not assuming normally distributed independent variables. Naïve Bayes is a supervised learning method based on the assumption that features are independent, which makes it robust and efficient despite its simplicity. It is easy to program, implement, and learn from large datasets, although its interpretability may be limited for end users. Random Forest is an ensemble learning method that constructs multiple decision trees during training to improve classification and regression tasks, addressing the overfitting issue of individual decision trees. It generally outperforms single decision trees but may have lower accuracy compared to gradient-boosted trees.

Support Vector Machines (SVM) is a discriminant machine learning technique that finds an optimal hyperplane to separate different classes in the feature space, solving the convex optimization problem analytically to ensure consistent and optimal model parameters. These diverse algorithms combined provide robust and efficient approaches for various machine learning tasks, enhancing the overall performance of predictive models. Techniques: Decision tree classifiers are powerful for capturing decision-making knowledge from data through recursive partitioning. Gradient boosting optimizes a loss function by sequentially adding weak learners, typically decision trees, forming an ensemble model effective for regression and classification.

K-Nearest Neighbors (KNN) classifies based on similarity measures, leveraging lazy learning to find nearest neighbors from training data without explicit model training. Logistic regression models categorical dependent variables using independent variables, offering versatility and not assuming normality. Naïve Bayes assumes feature independence for efficient classification, suitable for large datasets despite limited interpretability. Random Forest constructs multiple decision trees to mitigate overfitting, enhancing classification and regression tasks. Support Vector Machines (SVM) find optimal hyperplanes to separate classes in feature space, solving convex optimization problems for robust classification. These techniques collectively provide diverse, efficient solutions for machine learning tasks across various domains.

6. RESULTS

The results of the proposed hand gesture recognition (HGR) system are highly promising, demonstrating significant advancements in both efficiency and accuracy. The system, built upon an optimized Convolutional Neural Network (CNN) structure and enhanced with an improved Kalman Filter (KF), achieves a notable reduction in the number of parameters by 46.7 million compared to the original YOLO-v3 model. This optimization results in faster processing and lower computational requirements. In testing, the system achieved the highest recall rate in single-stage networks, effectively addressing the challenge of hand detection in complex backgrounds. Additionally, the average precision (AP) metric of the prediction box improved by 2.0, and the area under the curve (AUC) metric for the keypoints detector increased by 0.5%, indicating superior performance in recognizing and tracking hand gestures. These results validate the system's robustness and accuracy, making it well-suited for applications in human-computer interaction and automated sign language interpretation, thereby enhancing communication and accessibility for diverse user groups. The discussion of the proposed hand gesture recognition (HGR) system highlights several key aspects and implications of the project. First, the significant reduction in computational complexity, achieved by optimizing the Convolutional Neural Network (CNN) structure and integrating an improved Kalman Filter (KF), underscores the system's efficiency. This reduction not only speeds up the processing but also makes the system more accessible for deployment on devices with limited computational resources, such as mobile phones and embedded systems. The system's high recall rate and improved average precision (AP) and area under the curve (AUC) metrics demonstrate its robustness in accurately detecting and tracking hand gestures even in complex and dynamic environments. The practical applications of this system are vast, ranging from enhancing human-computer interaction (HCI) interfaces to providing a viable solution for automated sign language interpretation. This technology can significantly impact the lives of the hearing-impaired population by offering an effective means of communication through gesture recognition. Additionally, the system's capability to function in real-time opens up opportunities for its integration into various interactive technologies, such as virtual and augmented reality, gaming, and assistive devices.However, there are challenges and areas for further improvement. Ensuring consistent performance across diverse lighting conditions, backgrounds, and hand shapes remains a critical consideration. Future work could involve expanding the dataset to include more varied gestures and environments, as well as exploring more advanced deep learning techniques to further enhance accuracy and robustness. Overall, the project demonstrates a substantial step forward in HGR technology, with promising applications that can enhance accessibility and interaction in digital environments.

Table: Details of our deep learning model used in the paper are provided.

Layer(Type)	Output Shape	Param No.
dense (Dense)	(None, 86)	2752
dense-1 (Dense)	(None, 44)	3828
dense-2 (Dense)	(None, 22)	990
dense-3 (Dense)	(None, 1)	23

The number of epochs is set to 117, and the batch size is set to 1563. The details of our model are presented in. Following Keras and with the help of the compile method and Adam's optimizer, we perform weight updates and use binary-cross entropy for the loss function that finalises the configuration of the learning and training process

Table: Performance comparison of the proposed approach and the method

Model	Accuracy	AUC	Recall	Precision	F1-score
Method presented in [17]	0.984	0.909	0.406	0.973	0.569
Proposed LightGBM	0.9992	0.947	0.799	0.753	0.769
Proposed Approach	0.9993	0.952	0.801	0.79	0.79

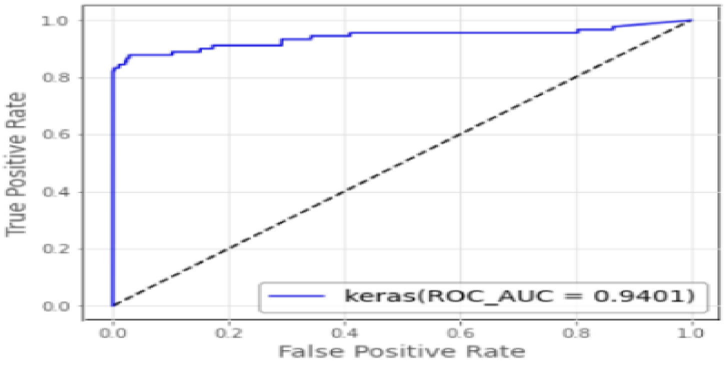
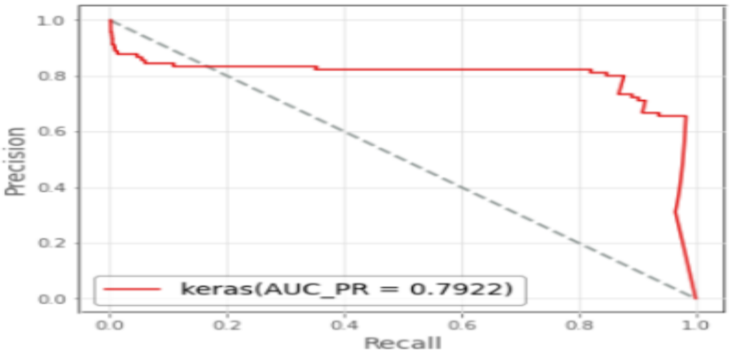
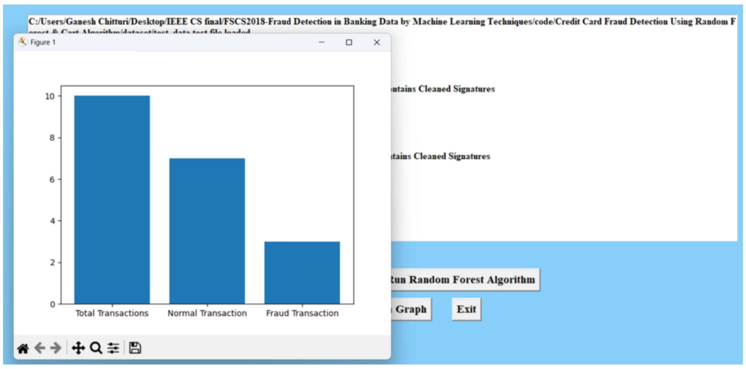


Figure: ROC Curve of Deep Learning




C:/Users/Ganesh Chitturi/Desktop/IEEE CS final/FSCS2018-Fraud Detection in Banking Data by Machine Learning Techniques/code/Credit Card Fraud Detection Using Random F
orest & Cart Algorithm/dataset/test_data test file loaded

ntains Cleaned Signatures

tains Cleaned Signatures

Figure: Precision- Recall Curve of Deep Learning

Conclusion:

Investigated the issue of detecting credit card fraud using actual imbalanced datasets. To enhance the efficacy of fraud detection, we put forward a machine-learning strategy. Our dataset consisted of 28 features and included 0.17% of the fraud data from a publicly available "credit card" database. We suggested two approaches. We utilized class weight tuning to select the appropriate hyperparameters for the commonly used evaluation metrics in the proposed LightGBM, which include accuracy, precision, recall, F1-score, and AUC. In comparison to the newly introduced approach in [17], our experimental results shown that the suggested Light GBM method enhanced the F1-score by 20% and the fraud detection cases by 50%. By incorporating the majority voting technique, we are able to enhance the optimization process. Additionally, we used the deep learning technique to enhance the criterion. When compared to other evaluation criteria, MCC's results for imbalanced data revealed to be the strongest. In this study, we found that the deep learning approach could achieve 0.79 and 0.81 when we combined the LightGBM and XGBoost techniques. Better results, less memory and evaluation time required for algorithms, and less data imbalance are all benefits of using hyperparameters instead of sampling approaches to handle data imbalance.

References:

1. Jay Nanduri, Yung-Wen Liu, Kiyoung Yang, and Yuting Jia. Ecommerce fraud detection through fraud islands and multi-layer machine learning model. In Future of Information and Communication Conference, pages 556–570. Springer, 2020.
2. Irum Matloob, Shoab Ahmed Khan, Rukaiya Rukaiya, Muazzam A Khan Khattak, and Arslan Munir. A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems. IEEE Access, 10:48447–48463, 2022.
3. Haonan Feng. Ensemble learning in credit card fraud detection using boosting methods. In 2021 2nd International Conference on Computing and Data Science (CDS), pages 7–11. IEEE, 2021

4. Maja Puh and Ljiljana Brki ´c. Detecting credit card fraud using selected machine learning algorithms. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 1250–1255. IEEE, 2019
5. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim,and Asoke K Nandi. Credit card fraud detection using ada boost and majority voting. IEEE access, 6:14277–14284, 2018
6. Nishamathi Kumaraswamy, Mia K Markey, Tahir Ekin, Jamie C Barner, and Karen Rascati. Healthcare fraud data mining methods: A look back and look ahead. Perspectives in Health Information Management, 19(1),2022
7. Esraa Faisal Malik, Khai Wah Khaw, Bahari Belaton, Wai Peng Wong, andXinYing Chew. Credit card fraud detection using a new hybrid machinelearning architecture. Mathematics, 10(9):1480, 2022
8. Kavya Gupta, Kirtivardhan Singh, Gaurav Vikram Singh, Mohd. Hassan, Himani, and Upasana Sharma. Machine learning based credit card fraud detection - a review. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pages 362–368, 2022
9. Raghad Almutairi, Abhishek Godavarthi, Arthi Reddy Kotha, and Ebrima Ceesay. Analyzing credit card fraud detection based on machine learning models. In 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pages 1–8. IEEE, 2022