International Journal of Advanced Research and Review

www.ijarr.org

Fraud Detection in Banking Data by using Machine Learning Techniques

G.Sushmitha[1], A.Sharanya[2], M.Sriharshitha[3], Mrs. D.Madhuri[4]
[1,2,3]UG Scholar, Dept. of AI&DS, St. Martin's Engineering College, Secuenderabad, Telangana, India – 500100
[4] Assistant Professor, Dept. of AI&DS, St. Martin's Engineering College, Secuenderabad, Telangana, India – 500100
sushmithagurrala3@gmail.com

Abstract: As technology advanced and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in an increase in the volume of banking transactions. Furthermore, the significant increase in fraud requires high banking transaction costs. As a result, detecting fraudulent activities has become a fascinating topic. In this study, we consider the use of class weight-tuning hyperparameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization in particular to optimize the hyperparameters while preserving practical issues such as unbalanced data. We propose weight-tuning as a pre-process for unbalanced data, as well as CatBoost and XGBoost to improve the performance of the LightGBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, we use deep learning to fine-tune the hyperparameters, particularly our proposed weight-tuning one. We perform some experiments on real-world data to test the proposed methods. To better cover unbalanced datasets, we use recall-precision metrics in addition to the standard ROC-AUC. CatBoost, LightGBM, and XGBoost are evaluated separately using a 5-fold cross-validation method. Furthermore, the majority voting ensemble learning method is used to assess the performance of the combined algorithms. LightGBM and XGBoost achieve the best level criteria of ROC-AUC = 0.95, precision 0.79, recall 0.80, F1 score 0.79, and MCC 0.79, according to the results. By using deep learning and the Bayesian optimization method to tune the hyperparameters, we also meet the ROC-AUC = 0.94, precision = 0.80, recall = 0.82, F1 score = 0.81, and MCC = 0.81. This is a significant improvement over the cutting-edge methods we compared it to.

Keywords: Fraud detection,Banking fraud,Financial fraud,Machine learning,Data analysis,Predictive modeling,Bayesian optimization, data Mining, deep learning, ensemble learning, hyper parameter, unbalanced data, machine learning.

1.INTRODUCTION

In recent years, there has been a significant increase in the volume of financial transactions due to the expansion of financial institutions and the popularity of web-based e-commerce. Fraudulent transactions have become a growing problem in online banking, and fraud detection has always been challenging . Along with credit card development, the their best to make it look legitimate, and credit card fraud has always been updated. Fraudsters do their best to make it look The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu . legitimate. They try to learn how fraud detection systems work and continue to stimulate these systems, making fraud detection more complicated. Therefore, researchers are constantly trying to find new ways or improve the performance of the existing methods . People who commit fraud usually use security, control, and monitoring weaknesses in commercial applications to achieve their goals. However, technology can be a tool to combat fraud . To prevent further possible fraud, it is important to detect the fraud right away after its occurrence. Fraud can be defined as wrongful or criminal deception intended to result in financial or personal gain. Credit card fraud is related to the illegal use of credit card information 3034 VOLUME 11, 2023 IEEE Transaction on Machine Learning,Volume:11,Issue Date:Jan.2023 for purchases in a physical or digital manner. In digital transactions, fraud can happen over the line or the web, since the cardholders usually provide the card number, expiration date, and card verification number by telephone or website. There are two mechanisms, fraud prevention and fraud detection, that can be exploited to avoid fraud-related losses. Fraud prevention is a proactive method that stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudster attempts a fraudulent transaction. Fraud detection in banking is considered a binary classification problem in which data is classified as legitimate or fraudulent. Because banking data is large in volume and with datasets containing a large amount of transaction data, manually reviewing and finding patterns for fraudulent transactions is either impossible or takes a long time. Therefore, machine learning-based algorithms play a pivotal role in fraud detection and prediction. Machine learning algorithms and high processing power increase the capability of handling large datasets and fraud detection in a more efficient manner. Machine learning algorithms and deep learning also provide fast and efficient solutions to real-time problems. In this paper, we propose an efficient approach for detecting credit card fraud that has been evaluated on publicly available datasets and has used optimised algorithms LightGBM, XGBoost, CatBoost, and logistic regression individually, as well as majority voting combined methods, as well as deep learning and hyperparameter 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.

2. LITERATURE SURVEY

Corruption or frauds has become common terms which are associated with government bodies working across the globe. It often leads to several social and economic problems, if remain unchecked. Increase in the rate of corruption adversely affects the development of any country. The government funds or money which is intended for the welfare of the public goes in the pocket of greedy officers. This research work is aimed to reduce corruption or frauds using blockchain technology. To establish our framework, we have worked on a generic scenario in which a government has various schemes running for the welfare of common people and the funds are disbursed through a layered architecture of government passing through various organisations. Non-transparency, poor management of government records, delay in verification process can lead to corruption in various schemes at various levels. Blockchain being a transparent, immutable and decentralized mechanism is found to be a mightier technology which can help fighting corruption in the experimental generic scenario.

The reputation system has been designed as an effective mechanism to reduce risks associated with online shopping for customers. However, it is vulnerable to rating fraud. Some raters may inject unfairly high or low ratings to the system so as to promote their own products or demote their competitors.

This paper has considered the problem of Video Fraudulence, i.e., attackers can tamper with the original video and can create a fake video of their own. This problem is of great practical importance given the massive volume of online videos available through the World Wide Web, Internet news feeds, electronic mail, corporate databases, and digital libraries. To the best of our knowledge, no such video fraud detection algorithm has been proposed in the literature that can detect, using Blockchain, whether the video has been tampered with. This paper is a comparative study and provides a prototype of how it applies Blockchain to detect video fraudulence. The focus is on the usability of Blockchain and how to implement it to get the desired result. Three features of Blockchain that are Decentralization, Data transparency, and Security and privacy are being used to provide a reliable solution. Some cryptographic algorithms are used to find unique feature in all of the videos that can act as the hash value of the video because, in Blockchains, every node stores the data in the form of the hash value. If the video is tampered with, then the hash value changes and hence any fraud in the video can be detected.

Since its inception in 2009, Bitcoin is mired in controversies for providing a haven for illegal activities. Several types of illicit users hide behind the blanket of anonymity. Uncovering these entities is key for forensic investigations. Current methods utilize machine learning for identifying these illicit entities. However, the existing approaches only focus on a limited category of illicit users. The current paper proposes to address the issue by implementing an ensemble of decision trees for supervised learning. More parameters allow the ensemble model to learn discriminating features that can categorize multiple groups of illicit users from licit users. To evaluate the model, a dataset of 1216 real-life entities on Bitcoin was extracted from the Blockchain. Nine Features were engineered to train the model for segregating 16 different licit-illicit categories of users. The proposed model provided a reliable tool for forensic study. Empirical evaluation of the proposed model vis-a-vis three existing benchmark models was performed to highlight its efficacy. Experiments showed that the specificity and sensitivity of the proposed model were comparable to other models. Due to higher parameters of the ensemble tree model, the classification accuracy was 0.91, with 95% CI - 0.8727, 0.9477. This was better than SVM and Logistic Regression, the two popular models in the literature and comparable to the Random Forest and XGBOOST model. CPU and RAM utilization were also monitored to demonstrate the usefulness of the proposed work for real-world deployment. RAM utilization for the proposed model was higher by 30-45% compared to the other

three models. Hence, the proposed model is resource-intensive as it has higher parameters than the other three models. Higher parameters also result in higher accuracy of predictions.

Applications of blockchain technologies got a lot of attention in recent years. They exceed beyond exchanging value and being a substitute for fiat money and traditional banking system. Nevertheless, being able to exchange value on a blockchain is at the core of the entire system and has to be reliable. Blockchains have built-in mechanisms that guarantee whole system's consistency and reliability. However, malicious actors can still try to steal money by applying well known techniques like malware software or fake emails. In this paper we apply supervised learning techniques to detect fraudulent accounts on Ethereum blockchain. We compare capabilities of Random Forests, Support Vector Machines and XGBoost classifiers to identify such accounts basing on a dataset of more than 300 thousands accounts. Results show that we are able to achieve recall and precision values allowing for the designed system to be applicable as an anti-fraud rule for digital wallets or currency exchanges. We also present sensitivity analysis to show how presented models depend on particular feature and how lack of some of them will affect the overall system performance.

We design a distributed platform with blockchain as a system service for supporting transaction execution in insurance processes. The insurance industry is heavily dependent on multiple processes between transacting parties for initiating, maintaining and closing diverse kind of policies. Transaction processing time, payment settlement time and security protection of the process execution are major concerns. Blockchain technology, originally conceived as an immutable distributed ledger for detecting double spending of cryptocurrencies, is now increasingly used in different FinTech systems to address productivity and security requirements. The application of blockchain in FinTech processing requires a deep understanding of the underlying business processes. It supports automated interactions between the blockchain and existing transaction systems through the notion of smart contracts. In this paper, we focus on the design of an efficient approach for processing insurance related transactions based on a blockchain-enabled platform. An experimental prototype is developed on Hyperledger fabric, an open source permissioned blockchain design framework. We discuss the main design requirements, corresponding design propositions, and encode various insurance processes as smart contracts. Extensive experiments were conducted to analyze performance of our framework and security of the proposed design.

The private insurance sector is recognized as one of the fastest-growing industries. This rapid growth has fueled incredible transformations over the past decade. Nowadays, there exist insurance products for most high-value assets such as vehicles, jewellery, health/life, and homes. Insurance companies are at the forefront in adopting cutting-edge operations, processes, and mathematical models to maximize profit whilst servicing their customers claims. Traditional methods that are exclusively based on human-in-theloop models are very time-consuming and inaccurate. In this paper, we develop a secure and automated insurance system framework that reduces human interaction, secures the insurance activities, alerts and informs about risky customers, detects fraudulent claims, and reduces monetary loss for the insurance sector. After presenting the blockchain-based framework to enable secure transactions and data sharing among different interacting agents within the insurance network, we propose to employ the extreme gradient boosting (XGBoost) machine learning algorithm for the aforementioned

insurance services and compare its performances with those of other state-of-the-art algorithms. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. Furthermore, we propose an online learning solution to automatically deal with real-time updates of the insurance network and we show that it outperforms another online state-of-the-art algorithm. Finally, we combine the developed machine learning modules with the hyperledger fabric composer to implement and emulate the artificial intelligence and blockchain-based framework.

Lakhs of people getting Degrees year after year, due to the lack of effective anti-forge mechanism, events that cause the graduation certificate to be forged often get noticed. In order to solve the problem of counterfeiting certificates, the digital certificate system based on block chain technology. All the illegal activities filled against a person and all the activities are updated in the Personal ID. Using the modification process we would monitor not only the degree cortication alone but also entire personality and behavioral activities of that person. We deploy Unique based monitoring using this system.

In the Bitcoin network, lack of class labels tend to cause obscurities in anomalous financial behaviour interpretation. To understand fraud in the latest development of the financial sector, a multifaceted approach is proposed. In this paper, Bitcoin fraud is described from both global and local perspectives using trimmed k-means and kd-trees. The two spheres are investigated further through random forests, maximum likelihood-based and boosted binary regression models. Although both angles show good performance, global outlier perspective outperforms the local viewpoint with exception of random forest that exhibits nearby perfect results from both dimensions. This signifies that features extracted for this study describe the network fairly.

3. PROPOSED METHODOLOGY

Many Machine Learning techniques have been proposed to deal with this problem, some results appear to be quite promising [4], but there is no obvious superior method. This paper compares the performance of various supervised machine learning models like SVM, Decision Tree, Naive Bayes, Logistic Regression, and few deep learning models in detecting fraudulent transactions in a blockchain network. Such comparative study will help decide the best algorithm based on accuracy and computational speed trade-off. Our goal is to see which users and transactions have the highest probability of being involved in fraudulent transactions.

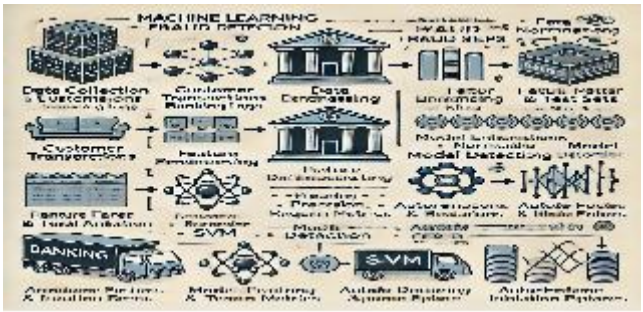

Figure 1: Proposed Machine Learning Fraud Detection System.

The proposed methodology typically includes the following key components:

- Data Collection: Gathering transactional data, customer profiles, and historical fraud cases.
- Data Preprocessing: Handling missing values, removing duplicates, and normalizing/standardizing numerical data.
- Feature Engineering: Extracting useful features like transaction frequency, amount patterns, and behavioral analytics.
- Data Splitting: Dividing data into training and testing sets for model validation.
- Model Selection: Choosing ML algorithms such as Decision Trees, Random Forest, SVM, or Neural Networks.
- Model Training & Evaluation: Training models using labeled fraud and non-fraud data, then evaluating performance using metrics like Accuracy, Precision, Recall, and F1-score.
- Anomaly Detection: Using unsupervised learning techniques like Autoencoders or Isolation Forests to detect unusual transactions.
- Deployment & Monitoring: Implementing the trained model in real-time fraud detection systems and updating it periodically for improved accuracy.
- Metric Evaluation: To assess the quality of the enhancement, the project often calculates various image quality metrics, such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and MSE (Mean Squared Error), to measure the similarity between the original and enhanced images.
- Output: The fraud detection system outputs a fraud probability score, transaction classification, real-time alerts, and risk assessment reports

- Evaluation and Benchmarking: It compares the model with rule-based systems, other ML techniques, and industry standards using real-world datasets. Computational efficiency is assessed to guarantee real-time fraud detection capabilities. Optimization techniques like hyperparameter tuning, feature selection, and ensemble learning enhance accuracy and adaptability to evolving fraud patterns.

Applications:

The enhanced fraud detection banking data images can be used in a wide range of applications, including:

- Identifies suspicious transactions and prevents financial losses in banking.
- Analyzes customer behavior to detect anomalies and flag fraudulent activities.
- Enhances security in online payments and prevents credit card fraud.
- Assists in risk assessment for loan approvals and financial decision-making.
- Supports anti-money laundering (AML) compliance and financial crime prevention.
- Protects e-commerce platforms, digital wallets, and fintech services from fraud.

Advantages:

LIME is a technique that leverages deep learning and image processing to enhance images captured in low-light conditions. It offers several advantages, making it a valuable solution for various applications:

- Fraud Detection: Identifies fraudulent transactions in real time, preventing financial losses.
- Improved Accuracy: Machine learning models reduce false positives and enhance detection precision.
- Automated Monitoring: Continuously analyzes transactions without human intervention.
- Adaptability: Learns from new fraud patterns and adapts to evolving threats.
- Cost-Effective: Reduces manual fraud investigations and operational costs.
- Scalability: Handles large volumes of banking transactions efficiently.
- Enhanced Security: Strengthens cybersecurity measures and builds customer trust.

4. EXPERIMENTAL ANALYSIS

We use the stratified 5-fold cross validation method and the boosting algorithms with the Bayesian optimization method to evaluate the performance of the proposed framework. We extract the hyperparameters and evaluate each algorithm individually before using the majority voting method. We examine the algorithms in triple and double precision

Figure 1 shows a collection of original banking transaction records, including both fraudulent and legitimate transactions. These raw transactions serve as the input to the fraud detection model. The dataset contains various features such as transaction amount, location, time, device ID, and merchant details. The purpose of this figure is to provide a visual representation of the real-world banking data that the model will analyze to detect fraudulent activities

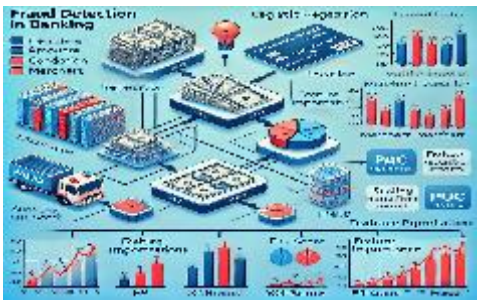

Figure 2: Original Transaction Data

Figure 2 displays a set of preprocessed transactions used for training the fraud detection model. These transactions have undergone data cleaning, normalization, feature selection, and encoding to improve the efficiency and accuracy of fraud detection. The figure helps to demonstrate the structured input format used for machine learning analysis. It presents the transactions flagged as fraudulent by the machine learning model. The model analyzes patterns such as unusually high transaction amounts, multiple transactions in a short period, and deviations from normal spending behavior. Fraudulent transactions are highlighted in red to indicate potential financial fraud. This illustrates the confusion matrix used to evaluate the fraud detection model's performance. The confusion matrix categorizes predictions into:

- **True Positives (TP)**: Correctly identified frauds.
- **False Positives (FP)**: Legitimate transactions incorrectly classified as fraud.
- **True Negatives (TN)**: Correctly classified legitimate transactions.
- **False Negatives (FN)**: Fraudulent transactions missed by the model.

This visualization helps in understanding the model's accuracy and identifying areas for improvement. It presents the Receiver Operating Characteristic (ROC) curve and Precision-Recall scores of the fraud detection model. These metrics provide insights into the model's effectiveness in distinguishing fraudulent transactions from legitimate ones.

Figure 3: Sample Transaction Data

The precision-recall curve is illustrated in Fig. 5 and shows the system performance in a more precise manner compared with the ROC-AUC curve. However, the results cannot be cited because false negatives are far from the view of this diagram. As Fig. 5 shows, the highest value belongs to the combination of the CatBoost and LightGBM algorithms with a value of 0.7672, and the lowest value belongs to logistic regression and is 0.7361. Comparing the precision, recall, and F1-score as well as the MCC, the algorithms used are shown in Fig. 6. The best performance is related to the combination of lightGBM and XGBoost algorithms, which have an MCC value of 0.79 and an F1-score of 0.79. In individual algorithms, XGBoost has the highest values. According to the digits obtained in Table 5, deep learning has achieved better performance compared with individual algorithms and majority voting ensemble learning. The MCC and F1-score metrics have values of 0.8129 and 0.8132, respectively. The area under the ROC curve in the deep learning method is illustrated in Fig. 7 and shows a value of 0.9401.

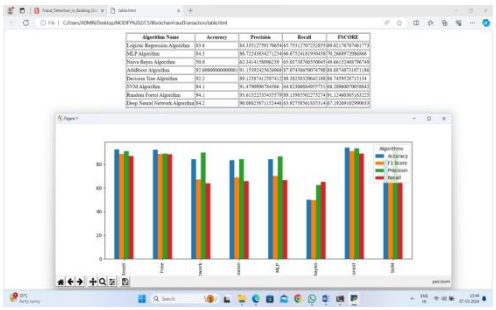

Figure 4: Algorithms Metrics

5. CONCLUSION

This blockchain technology and its defining characteristics. They also review state-of-the-art technologies for detecting online fraud and intrusions, identify certain fraud and malicious activities that blockchain technology can effectively prevent, and make recommendations for strategically fighting various attacks to which blockchain technology may be vulnerable. Existing machine learning and data-mining algorithms could find new uses in identifying fraud and intrusions in blockchain-based transactions. Guided machine learning methods like deep-learning neural networks, support vector machines, and Bayesian belief networks may help detect outlier behaviors by profiling, monitoring, and detecting behavioral trends based on people's transaction histories. Despite the advancement in technology, still, the problems regarding Video Fraudulence are faced and there is no concrete solution for this problem. To improve the technology and related anti-attack methods, more research is required. Machine learning-based fraud detection in banking plays a crucial role in identifying and preventing fraudulent activities in real time. By leveraging advanced algorithms, it enhances accuracy, reduces manual intervention, and adapts to evolving fraud patterns. Its ability to analyze large transaction volumes, provide real-time alerts, and ensure compliance with financial regulations makes it an essential tool for financial security. The integration of fraud detection systems across banking and digital platforms helps protect customers, reduce financial losses, and improve overall trust in financial institutions. With continuous advancements, machine learning will further strengthen fraud prevention, making banking transactions safer and more efficient. Fraud detection will focus on improving accuracy, efficiency, and real-time detection capabilities. Advanced deep learning models, such as Graph Neural Networks (GNNs) and Transformers, will enhance the detection of complex fraud patterns and relationships within transactional data. Explainable AI (XAI) techniques, like SHAP and LIME, will make fraud detection models more transparent, helping financial institutions understand and trust AI-driven decisions. Real-time fraud detection will be significantly improved through federated learning, which enables multiple banks to train models collaboratively without sharing sensitive data, and Edge AI, which allows fraud detection directly on devices to reduce latency. Hybrid and ensemble learning approaches, combining supervised, unsupervised, and reinforcement learning, will enhance adaptability against evolving fraudulent techniques. Blockchain integration will add an extra layer of security by ensuring immutable transaction records, while AI-driven smart contracts will enable automated fraud prevention. Additionally, behavioral biometrics, such as keystroke dynamics and device interaction, are likely to focus on several key areas.

REFERENCES

[1] Joshi, P.,Kumar, S.,Kumar, D.,&Singh,A. K. (2019,September). A blockchain based framework for fraud detection. In 2019 Conference on Next Generation Computing Applications (NextComp) (pp. 1-5). IEEE.

[2] Cai, Y., & Zhu, D. (2016).Fraud detections for online businesses: a perspective from blockchain technology. Financial Innovation, 2(1), 1-10.

[3] Dhiran, A., Kumar, D., & Arora, A. (2020, July). Video Fraud Detection using Blockchain. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 102-107). IEEE.

[4] Nerurkar P, Bhirud S, Patel D, Ludinard R, Busnel Y, Kumari S. Supervised learning model for identifying illegal activities in Bitcoin. Appl Intell. 2020;209(1):1- 20.

[5] Ostapowicz, M., & Żbikowski, K. (2020, January). Detecting fraudulent accounts on blockchain: a supervised approach. In International Conference on Web Information Systems Engineering (pp. 18-31). Springer, Cham.

[6] Raikwar, M., Mazumdar, S., Ruj, S., Gupta, S. S., Chattopadhyay, A., & Lam, K. Y. (2018, February). A blockchain framework for insurance processes. In 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS) (pp. 1-4). IEEE.

[7] Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. IEEE Access, 8, 58546-58558.

[8] Shanmuga Priya P and Swetha N, "Online Certificate Validation using Blockchain", Special Issue Published in Int. Jnl. Of Advanced Networking and Applications (IJANA).

[9] Monamo, P. M., Marivate, V., & Twala, B. (2016, December). A multifaceted approach to bitcoin fraud detection: Global and local outliers. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 188-194). IEEE.

[10] Xu, J. J. (2016).Are blockchains immune to all malicious attacks? Financial Innovation, 2(1), [11] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[12] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[13]X. Kewei, B. Peng, Y. Jiang and T. Lu, "A hybrid deep learning model for online fraud detection", Jan. 2021.

[14]A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine", IEEE Access , vol. 8, 2020.

[15]E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong and X. Chew, "Credit card fraud detection using a new hybrid machine learning architecture", Mathematics, vol. 10, no. 9, pp. 1480, Apr. 2022.