IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 2, Issue 2, February 2021
Impact Factor: 5.731

Anomaly Detection for Banking Fraud Prevention Using Advanced Machine Learning Techniques

Deepak Reddy Suram
Senior Software Engineer & Cloud Data Architect
H&R Block, Inc
reddydeepaksuram@gmail.com
ORCID: 0009-0004-9698-0791

Abstract: *Over the past few years, because cashless transactions and digital banking are developing so quickly, the risk of financial fraud has risen significantly, which has left a high demand for the accuracy and real-time system of detecting anomalies. The effective machine learning-based banking fraud detection system suggested by this study should be able to recognize the presence of odd credit card transactions. To handle data noise and redundancy, as well as extreme class imbalance, the suggested technique would include sound data preprocessing, Random under sampling is used for class balance, while Principal Component Analysis (PCA) is employed to diminish dimensionality. Kaggle provides a sizable dataset for testing on credit card fraud detection (CCFD). Two very complicated models, Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost), are trained and evaluated using common performance metrics, including accuracy (acc), precision (pre), recall (rec), F1-score (F1), ROC, and AUC. The experiment shows that the XGBoost model performs marginally better with 99.8% acc, rec, F1, and an AUC of 0.999, whereas the recommended LSTM model obtains 99.7% acc, pre, rec, and F1. These findings prove the power, efficiency and great predictability of the proposed framework and suggest its suitability in actual-time application in the modern banking system to identify and thwart fraud in a successful way.*

Keywords: Fraud Detection, Machine Learning, Financial Security, Fraudulent Transactions, Anomaly Detection, Financial Transactions, Fraud Prevention

I. INTRODUCTION

In recent years, the widespread use of electronic financial services has resulted in remarkable ease of use and appeal. Nevertheless, this has also brought about different types of online banking fraud and financial cybercrimes that have become more complex [1]. With users migrating to online banking platforms, real-time payment, and networked transaction platforms, bad actors are taking advantage of the system vulnerabilities, causing an urgent demand for smart, responsive, and privacy-sensitive fraud detection systems [2]. The growing amounts and sophistication of financial operations have made conventional systems of fraud detection ineffective, leading to the advancement of Big Data-powered fraud and anomaly detection technologies. The accuracy, potency, and adaptability of fraud protection systems are all enhanced by these tactics by tracking abnormal tendencies that are vital in making effective decisions [3].

Financial institutions are faced with many challenges of detecting banking fraud [4]. Fraudsters are continuously upgrading their tactics to exploit the gaps of the systems and it is not simple to keep traditional rule-based systems abreast with the Method, equipment's and computer readable media of executing a data exchange on a foundation for data exchange. The sheer amount of transactions also makes it difficult to detect, because a manual analysis is time-consuming, and may not identify the fine details of a pattern that would indicate fraud [5]. The process, tools, and computer-readable medium used to share data with a data exchange framework. Financial institutions should also reconcile security and customer convenience, too much security can lead to false alarms and disturb the normal operations of the customer and may destroy trust and customer loyalty.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 2, Issue 2, February 2021 Impact Factor: 5.731

The conventional advanced fraud schemes are too complex for rule-based fraud detection systems to handle due to their extension to a limited set of rules that cannot be updated on a regular basis and need to be handled manually [6]. ML-based systems offer a dynamic solution, which utilizes sophisticated algorithms to identify the presence of anomalies, complex transactions, and reduce the number of false positives [7][8]. ML-based solutions effectively identify fraud recurring patterns in large monetary datasets as compared to traditional algorithms. Furthermore, the degree of precise identification has increased due to recent advancements in DL and hybrid models, through the capabilities of minimizing false alerts and collecting intricate transactional behaviors.

A. Motivation and Contribution

Digital banking has escalated the frequency and complexity of frauds and the inefficiency of traditional rule-based fraud detection systems has been proven. The dynamism in the trends of fraud, extensive transactions and the need to minimise false positives necessitates clever and adaptive solutions. As a result, advanced machine learning-based fraud detection methods are used to effectively spot dishonest behaviour, as well as offer scalable, real-time and consumer-friendly banking fraud detection. In addition, the financial systems are more secure against the new cyber threats because big data analytics and deep learning models enable detecting the hidden and not yet observed fraudulent patterns. These smart structures are also useful in preemptive risk management and decision-making in modern-day financial establishments. This research offers several key contributions as listed below:

- Utilizing of CCFD Dataset, which is hosted in Kaggle and includes real-life transactional data to validate practically.
- End-to-end data cleaning, data normalization, feature selection via PCA and data balancing via random under-sampling.
- The development of a strong framework for detecting fraud utilizing deep learning and advanced ML models (XGBoost and LSTM) to increase the precision of detection.
- Reliable model validation is ensured by evaluation utilizing standard performance indicators, such as rec, acc, pre, F1, ROC and AUC.

B. Justification and Novelty

The originality of the given piece of work is the realisation of a complex framework of anomaly detection that introduces both deep learning-based LSTM-based models and ensemble models based on XGBoost, which help address the issues caused by extremely skewed CCFD. The recommended method combines the advantages of LSTM and XGBoost to improve detection performance to represent sequential and temporal transaction patterns and nonlinear decision boundaries, respectively, as well as class imbalance. PCA-based feature selection and random under-sampling are also added, which helps improve the data presentation and the strength of the model. The complexity and rarity of fraudulent transactions justify this methodology, as single-model techniques are not particularly effective. The experiment results, with a 99.8% accuracy and low-cost computations, demonstrate the efficiency, dependability, and suitability of the provided framework for detecting financial fraud in real time.

C. Organization of the Paper

The paper is organized as follows: Section II presents a comprehensive review of existing studies, Section III describes the proposed methodology, Section IV discusses the experimental results, performance evaluation, and comparative analysis of different algorithms. Finally, Section V provides an interpretation of the results, the conclusion, and the future scope of the proposed approach.

II. LITERATURE REVIEW

This section examines studies on various ML and DL techniques for anomaly detection (AD) in the fight against financial fraud. A synopsis of this research is shown in Table I.

Alghofaili, Albattah, and Rassam (2020) proposed using the LSTM approach for financial fraud detection. This model

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 2, Issue 2, February 2021 Impact Factor: 5.731

is geared towards improving the existing detecting methods, besides improving the detecting accuracy in the face of big data. The proposed approach is tested using an actual credit card theft dataset, and the results are contrasted with those of various other machine learning techniques, including an established DL model known as the Auto-encoder model. The trial's outcomes demonstrated that LSTM operated faultlessly, with 99.95% accuracy in under one minute [9].

Rai and Dwivedi (2020) presents a technique that makes use of NN and utilizing unsupervised learning to CCFD. The K-Means, AE, LOF, and IF clustering techniques are outperformed by the proposed method. The suggested neural network-based FD system outperforms the other methods—The AE, IF, LOF, and K Means have accuracy ratings of 97%, 98%, 98%, and 99.75%, respectively[10].

Malaiya et al. (2019) constructed and examined DL architectures using Sequence-to-Sequence (Seq2Seq), Variational AutoEncoder (VAE), and Fully Connected Networks (FCNs). In order to do the comprehensive evaluation, utilize a sizable sample of datasets with unique characteristics provided by the population. The work of deep learning-based networks to detect network anomalies is plausible, and the enhanced accuracy over the traditional methods of learning is supported by experiment. Specifically, the Seq2Seq LSTM-based detection model has great potential, having more than 99 percent accuracy in detecting network anomalies based on assessing all the datasets used in the tests [11].

Mubalaike and Adali (2018) designed to understand the application of DL models to create a charitable strategy to fraudulent transactions detection in a highly accurate manner. Evaluations of the produced classifier models are conducted using the confusion matrix, ROC curves, and parameters including precision, sensitivity, specificity, and accuracy. The best accuracy scores are 80.52, 91.53, and 90.49, respectively. The relative performance is such that limited Boltzmann machine can get much better performance as compared to the other procedures [12].

Yee et al. (2018) There were several Bayesian network classifiers used to demonstrate supervised categorization, including the logistics, J48, K2, TAN and NB classifiers. Normalisation and principal component analysis were used to preprocess the dataset, and the results indicated that all classifiers had an accuracy of above 95.0%, which is much higher than the results prior to pre-processing [13].

Yau et al. (2017) proposed a unique technique to identify aberrant PLC events using OCSVM. The simulated traffic lights control system was exposed to the methodology in order to show its efficiency and accuracy. They find that high accuracy of detecting abnormal PLC operations is achieved that can assist investigators to conduct PLC forensics effectively and efficiently Semi-supervised machine learning for identifying PLC anomalies [14].

Table 1: Recent Studies on Anomaly Detection for Banking Fraud Prevention using Machine Learning

Author	Proposed Work	Results	Key Findings	Limitations & Future Work
Alghofaili, Albattah & Rassam (2020)	Detecting credit card theft from large data sets using a deep learning model based on LSTM	99.95% accuracy, execution time < 1 minute	LSTM significantly outperforms Auto-Encoder and traditional ML models in both accuracy and speed	Evaluated on a single real-world dataset; future work can focus on cross-dataset validation and real-time deployment
Rai & Dwivedi (2020)	Detecting credit card theft with unsupervised neural networks	99.87% accuracy	NN-based unsupervised learning outperforms AE, LOF, IF, and K-Means	Performance tested on limited data distributions; future work may include hybrid or semi-supervised approaches
Malaiya et al. (2019)	Deep learning-based anomaly detection using FCN, VAE, and Seq2Seq (LSTM) models	>99% accuracy across multiple datasets	Seq2Seq with LSTM consistently delivers superior anomaly detection performance	High computational complexity; future work can optimize models for scalability and real-time environments
Mubalaike & Adali (2018)	Deep learning classifiers including Restricted Boltzmann Machine (RBM) for	Accuracy: 90.49%, 80.52%, 91.53%	RBM outperforms other DL techniques in fraud detection accuracy	Lower performance compared to newer DL models; future research can integrate RBM with advanced deep

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 2, Issue 2, February 2021 Impact Factor: 5.731

	fraud detection			architectures
Yee et al. (2018)	Supervised classification using Bayesian Networks (K2, TAN, NB), Logistic Regression, and J48	>95% accuracy after preprocessing	Data normalization and PCA significantly enhance classifier performance	Focus limited to supervised learning; future work may explore imbalanced data handling and deep learning methods
Yau et al. (2017)	OCSVM-based anomaly detection for PLC systems	High anomaly detection accuracy (qualitative)	Effective identification of anomalous PLC behavior supports digital forensics	Evaluated on simulated systems; future work should validate on real-world industrial PLC environments

III. RESEARCH METHODOLOGY

This method's objective is to developed a fraud detection system that is effective and precise in identifying abnormal credit card transactions within banks. The offered solution combines data preprocessing, class-balancing strategies, and reducing dimensionality to improve data quality and lessen the existing class imbalance in fraud datasets. The usefulness of the suggested methodology in detecting financial fraud in real time is demonstrated using standard evaluation measures. Figure 1 shows the procedure's fundamental flow, while the methodology's specific instructions are listed below:

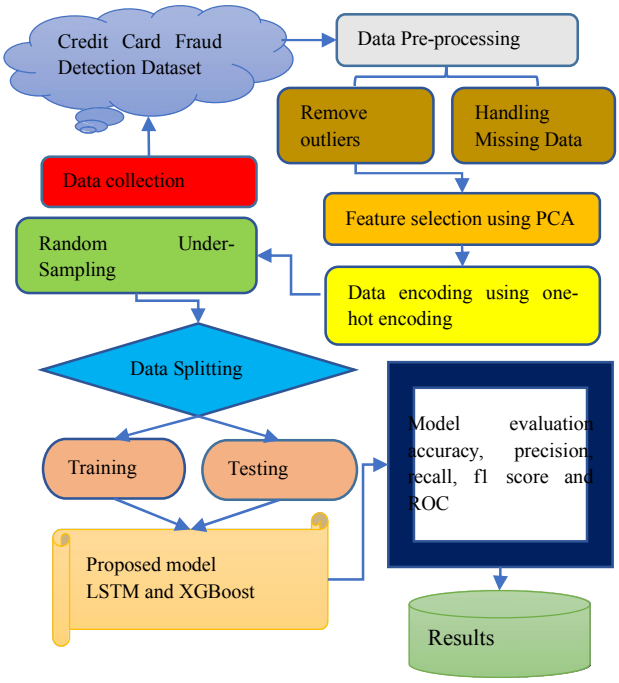

Fig. 1. Proposed Flowchart for Anomaly Detection for Banking Fraud Prevention Using Machine Learning

The following critical stages comprise the Anomaly Detection for Banking Fraud Prevention methodology:

A. Data Gathering and Analysis

The research of the dataset used in this study must be utilized to identify credit card fraud, which was obtained via Kaggle. There are 87,403 cases of fraudulent transactions and 912,597 instances of legitimate transactions in the first dataset. Below are data visualizations that were employed to examine feature correlations, attack distribution, etc.,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 2, Issue 2, February 2021
Impact Factor: 5.731

including bar graphs and heatmaps:

Figure. 2 illustrates a class imbalance in a dataset, probably for a task involving fraud detection or transactions. The data indicates that 80.0% of transactions, are Legitimate (colored green), while only a small minority, 20.0%, are Fraudulent (colored red and slightly exploded for emphasis).

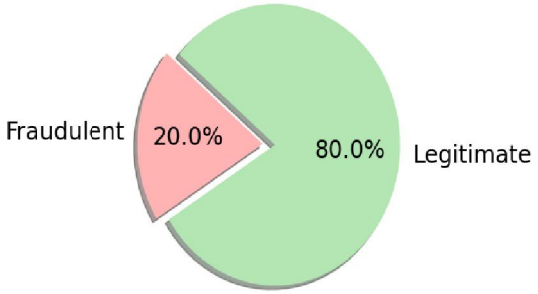

Fig. 2. Dataset Class Distribution for Credit Card Fraud Detection Dataset from Kaggle

This significant discrepancy draws attention to a common problem with fraud detection, where the target class (fraud) is rare, this may have an adverse effect on ML models' performance and training if not addressed using methods such as specialized loss functions or resampling.

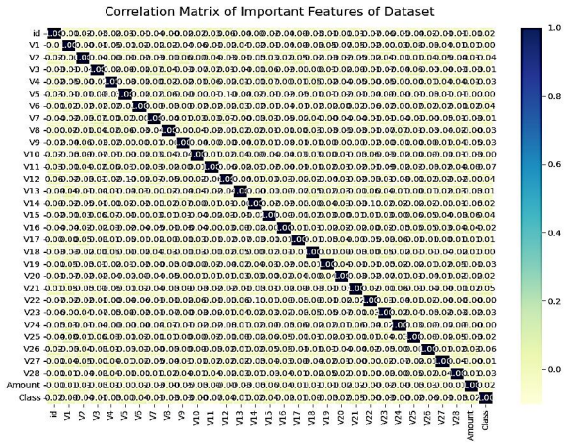

Fig. 3. Correlation Matrix of Important Features using Credit Card Fraud Detection Dataset

Figure 3 presents the pairwise linear relationships between numerous features labeled V1 through V28. The diagonal of the matrix is colored white and shows a value of 1, it is to be anticipated as it shows how each property is correlated with itself. The rest of the matrix is primarily colored in a dark pink/red, which, as per the color bar provided, represents the very low magnitude of correlation, with most being close to 0 (identified by the white text labels such as 0, -0.1, or 0.1), in the cells. This near-zero correlation in most pairs of V-features is an indication that the features are largely linearly independent of each other, which is a desirable property of numerous machine learning models since it reduces multicollinearity.

B. Data Pre-processing

The CCFD Dataset from Kaggle was used to prepare the data, it consisted of data concatenation, cleaning, and feature engineering. The preprocessing involved the treatment of missing data, outliers and leveling and normalization of the data. The most important preprocessing processes are recapped in the following manner:

- **Handling Missing Data:** Handling missing data involves deleting affected cases (deletion) or replacing missing values with estimated ones (imputation), with methods like multiple imputation being a recommended advanced

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 2, Issue 2, February 2021 Impact Factor: 5.731

technique that uses models to predict missing values, whereas simple methods include mean, median, or mode imputation.

- **Remove outliers:** Data preprocessing technique is the removal of outliers, which enhances the quality and precision of data analysis and ML models. Outliers refer to data that are far apart, in comparison with most data, and can have a perverse influence on the measures and model of statistics.

C. Feature Selection using PCA

The dimensionality reduction and component interpretation steps in component selection are a two-step process of component selection and analysis of the most significant original features using Principal Component Analysis (PCA). The time feature keeps track of each transaction's time in seconds since the first one. Principal component analysis (PCA) yielded the main components, which are the V1~V28 characteristics.

D. Data Encoding using One-Hot Encoding

One method of data preparation that generates a new binary column is called "one-hot encoding" in the table with distinct categories to represent categorical data as numerical data. The anticipated likelihood that samples i belongs to class c is known as the one-hot representation, or p_{ic}, and it is computed as in Equation (1):

$$p_{ic} = argmax(\hat{y}_{out}) \quad (1)$$

E. Data Balancing using Random Under-Sampling

In data balancing, an unbalanced dataset, i.e., compared to the other classes, one class contains many fewer samples, is corrected by balancing the class distributions to enhance ML models' performance. The first solution, often used, is random under-sampling, it reduces imbalance by choosing examples at random from the dominant class until the classes are roughly balanced.

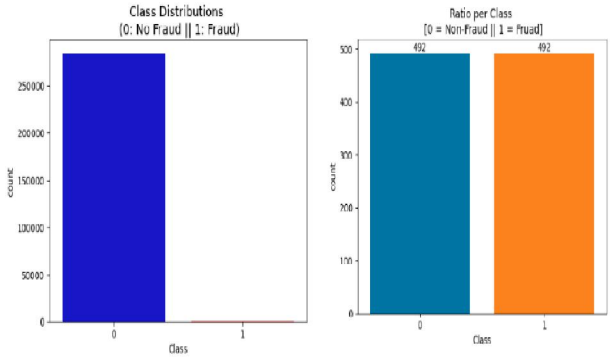

Fig. 4.Class Distribution Before and After Resampling

The distribution of a classification variable that would likely be employed in fraud detection is displayed in Figure. 4. With the great majority of observations falling into Class 0 (No Fraud) and a tiny percentage falling into Class 1 (Fraud), The data in the Class Distributions graphic (0: No Fraud || 1: Fraud) on the left is not balanced. The chart on the right, which is called Ratio per Class (0 = Non-Fraud || 1 = Fraud), is a very different story, where it is observed that Class 0 (non-fraud) and Class 1 (fraud) have the same amount of observations (492 in each class). This suggests that the second chart is likely the result of an oversampling, under sampling, or synthetic sampling technique applied to the initially unbalanced data to produce a balanced dataset for training classification models.

F. Data Splitting

To assess 70% of the dataset was allocated for training and 30% for assessing the performance of the model on unknown data. Employing a stratified split method, both subsets were designed to maintain the original dataset's class distribution.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 2, Issue 2, February 2021 Impact Factor: 5.731

G. Proposed Long Short-Term Memory (LSTM) Model

Text categorisation is an area where LSTMs are an expert because they can detect long-term associations in text. An RNN, or recurrent neural network, is an example of an LSTM classifier. Layered networks called RNNs employ their previous outputs as inputs for layers that come after them. Instead of working with individual data points, LSTM can handle data sequences thanks to its feedback connections. An input gate, an output gate, a forget gate, and a cell make up an LSTM node. Long-term value retention is managed by three gates that regulate the cell's information flow. Each of the three multiplicative gates that make up the LSTM layers is made up of memory blocks that are recurrently coupled. For a certain amount of time, gates continuously write, read, and reset data to guarantee that the temporary data is used. Here are the updated values for the unit's input x_t, h_{t-1}, c_{t-1} and output h_t, c_t.Gates

$$i_t = \sigma(w_i x_t + U_i h_{t-1} +) b_i \quad (2)$$

$$f_t = \sigma(w_f x_t + U_i h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(w_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

Input transform:

$$g_t = tanh(w_g x_t + U_t h_{t-1} = b_g) \quad (5)$$

State update:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

The logistic sigmoid function with multiplication by elements is shown by the symbols σ and $\odot$ in the equations (2-7) above, respectively. The LSTM unit has a memory cell c_t, a hidden unit h_t, an input gate i_t, a forget gate f_t, and an output gate o_t at each time step t. The learning parameters are W and U, whereas the additional bias is represented by b. The amount of internal memory state exhibited is managed by The forget gate, the output gate regulates the amount of memory cell erasure, whereas the input gate governs the degree of each unit's updating.

H. Proposed Extreme Gradient Boosting (XGBoost) Model

An ensemble learning technique called XGBoost uses decision trees to create predictions. To solve regression issues, a loss function that determines the difference between the target values' actual and expected values can be reduced. Represents the mathematical model for XGBoost regression Equation (8):

$$y = f(x) \quad (8)$$

where $f(x)$ is the XGBoost model, x is the expected property price, and y is the vector of input factors (such as square footage, number of bedrooms, etc.). An ensemble of decision trees is produced by XGBoost in order to compute $f(x)$ by training decision trees to minimise the MSE loss function. The model aggregates the forecasts from several decision trees to provide a final prediction. All things considered, the XGBoost regression model can look like this Equation(9):

$$y = \sum(k = 1\ to\ K)\ fk(x) \quad (9)$$

where $fk(x)$ is the forecast of the $K$the decision tree, where k is the ensemble's total number of trees. A weighted average of the leaf values gained during training is used to forecast individual trees. All of the ensemble's DT predictions are added to determine the XGBoost model's prediction for an input x.

I. Evaluation Metrics

The model's accuracy and classification performance were evaluated using the F1-score, recall, and validation accuracy since they are appropriate metrics for assessing models trained on balanced datasets. The trade-off between TP and FPR is depicted by the ROC curve, after which they were tested using standard evaluation metrics. True Positive (TP) measures how well the model can detect and classify harmful behavior, whereas True Negative (TN) measures how well it can detect and classify normal, non-malicious activity. These basic measurements are used as the basis of the derivation of overall performance measures. Equations (10) to (13) are developed as follows.

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

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 2, Issue 2, February 2021
Impact Factor: 5.731

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (12)$$

$$\text{F1} - \text{score} = 2 * \frac{(\text{precision} * \text{recall}}{(\text{precision+recall})} \quad (13)$$

Accuracy is very effective with balanced datasets. It is the percentage of all samples with predictions that were correct. Recall is the percentage of TP that is successfully recognised, whereas accuracy is the percentage of properly identified positive reports to all TP. The model's overall dependability is shown by the F1-score, which is the harmonic mean of recall and accuracy. Furthermore, the trade-off between TP and FPR is depicted by the ROC curve, whereas the AUC value assesses the classifier's overall effectiveness.

IV. RESULTS AND DISCUSSION

The section includes a brief explanation of the performance testing and experimental organization of the proposed models. Experiments were conducted on the system using GeForce RTX 3070 laptop graphics, and it took 32 seconds to train and 6 seconds to test. It was coded in PyTorch (v2.6.0+cu118) and scikit-learn (v1.5.2) and was trained on the CCFD Dataset on Kaggle. While the LSTM model received a score of 99.7% on all performance assessment metrics, including acc, pre, rec, and F1, and the applicability of the proposed models to performance-based real-time banking fraud detection in Table II, XGBoost achieved a slightly higher score of 99.8% on acc, rec, and the F1 of competence and efficiency.

Table 2: Classification Results of the Proposed Deep Learning and Machine Learning Model for Anomaly Detection for Banking Fraud Prevention using Ccfd Dataset

Matrix	LSTM	XGBOOST
Accuracy	99.7	99.8
Precision	99.7	99.7
Recall	99.7	99.8
F1-score	99.7	99.8

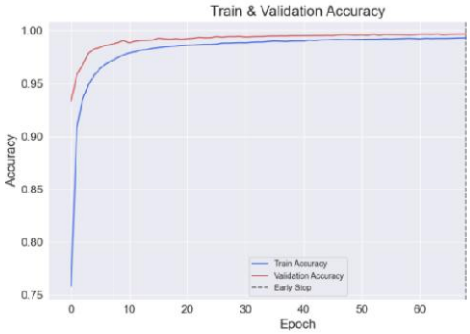

Fig. 5. Accuracy curve for the LSTM Model

Figure 5 illustrates the Validation and Training model accuracy across 80 epochs. Both the blue (Train Accuracy) and red (Validation Accuracy) lines demonstrate a sharp rise in accuracy throughout the early epochs, quickly reaching above 95%. The validation accuracy briefly surpasses the training accuracy around epoch 10, but then both lines converge, staying very close to 1.00 (or 100%) accuracy for the majority of the training run, demonstrating that there is no discernible overfitting and that the model's remarkable performance is shown by both the training and validation data sets. A dashed vertical line around epoch 69 marks where an Early Stop mechanism was triggered, suggesting the training process was halted at this point because further epochs were unlikely to provide substantial improvement or might risk starting to degrade performance, although the graph shows consistently high accuracy up to epoch 80.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 2, Issue 2, February 2021
Impact Factor: 5.731

Fig. 6. LSTM Model Loss Curve

The model's Training and Validation Loss across 80 epochs is shown in Figure 6. The blue line represents training loss in the early epochs, and the red line represents validation loss. Both lines rapidly decline, quickly falling from values around 0.5 to below 0.1, indicating the model is learning effectively. Following this early steep fall, the loss curves then become smooth with a much slower rate of descent reaching levels around zero towards epoch 40. The Validation Loss (red line) is always able to be smaller than the Train Loss (blue line) over the training process, a strange but not unprecedented phenomenon, indicating that the validation data may be less complicated or the regularization of the model works really well with the training set. An Early Stop event is denoted by a dashed vertical line at the epoch 69, at which point training was stopped due to the minimal improvement, though the loss was technically decreasing to epoch 80.

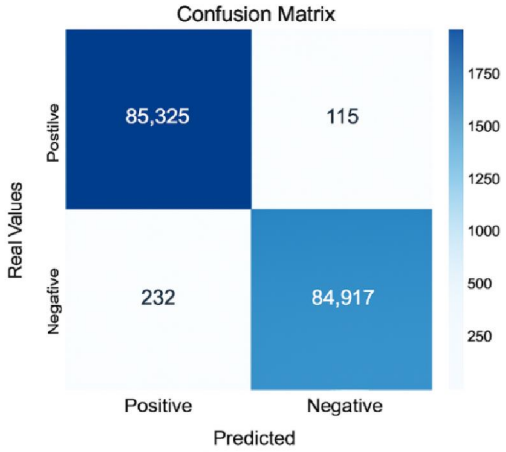

Fig. 7. XGBoost Model Confusion Matrix

The confusion matrices in Figure 7 compare Real Values (actual class) with Predicted Values. The model is very accurate on a global scale, and most of the classifications are concentrated on the main diagonal. In the Positive class, the model was able to identify 85325 true positives with a false negative of 115. In the case of the Negative, it was able to identify 84,917 cases as True Negatives, yet projected 232 as Positive (False Positives). The model is quite effective at differentiating between the two groups, as evidenced by the large numbers under TP and TN and the very low error values.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 2, Issue 2, February 2021
Impact Factor: 5.731

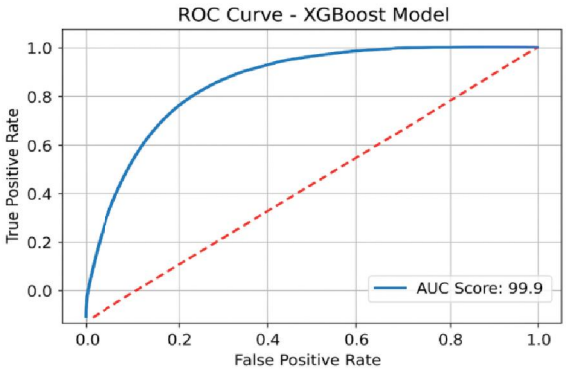

Fig. 8. XGBoost model ROC Curve

The ROC Curve of the XGBoost Model, which was used for binary classification, is displayed in Figure 8. The curve shows the ratio of the TPR (Sensitivity) to the FPR (1-Specificity) at different classification criteria. The red line that is dashed indicates a random classifier, is significantly higher than the blue line, which rises abruptly in the upper-left direction. The XGBoost model's Area Under the Curve (AUC) Score of 99.9 (likely 0.999) demonstrates its remarkable capacity to discriminate between the negative and positive groups at practically every threshold.

A. Comparative Analysis

The effectiveness of the suggested LSTM and XGBoost models can be assessed by providing a comparative analysis of their accuracy with other existing models in Table III. According to the results, the common ML models, such as the DT, K2, and SVM have accuracies of 95.5, 95.8, and 93.7, respectively, which are quite reasonable, but not very capable of modelling complicated data patterns. Contrarily, the proposed LSTM model increases the detection performance to 99.7, whereas XGBoost also improves the performance to the highest accuracy of 99.8, which indicates a higher efficiency of DL and ensemble-based methods in predicting results with high accuracy and reliability.

Table 3: Comparison of Different Machine Learning Models for Anomaly Detection for Banking Fraud Prevention on CCFD Dataset

Models	Accuracy
DT[15]	95.5
K2[13]	95.8
SVM[16]	93.7
LSTM	99.7
XGBoost	99.8

The proposed LSTM and XGBoost models have great benefits because they have high predictive accuracy and high capacity to forecast complex patterns of data. The features of LSTM are that it can be used to accurately model the sequential and temporal dependencies and, therefore, learn time-dependent features thus achieving a high accuracy of 99.7%. XGBoost also maximizes its performance by using ensemble learning and gradient boosting to address nonlinear relationships, eliminate overfitting, and enhance generalization and reaches accuracy of 99.8. These models rather collectively present a powerful, scalable, and highly accurate framework that would be well adapted to the real-world application that demands reliable and accurate prediction.

REFERENCES

[1] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *J. Big Data*, vol. 7, no. 1, p. 42, Dec. 2020, doi: 10.1186/s40537-020-00320-x.

[2] B. R. Cherukuri, "Ethical AI in cloud: Mitigating risks in machine learning models," *World J. Adv. Eng. Technol. Sci.*, vol. 1, no. 01, pp. 096–109, 2020, doi: 10.30574/wjaets.2020.1.1.0018.

[3] M. B. S. et al. . Ms. Bhavna Sharma et al., "An Investigation of Challenges Faced in Detecting Frauds," *Int. J.*

Mech. Prod. Eng. Res. Dev., 2020, doi: 10.24247/ijmperdjun20201129.

[4] R. R. Popat and J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning," in *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 2018. doi: 10.1109/ICOEI.2018.8553963.

[5] S. Achouche, U. B. Yalamanchi, and N. Raveendran, "Method, apparatus, and computer-readable medium for performing a data exchange on a data exchange framework," 2019

[6] S. Omar, A. Ngadi, and H. H. Jebur, "Machine Learning Techniques for Anomaly Detection: An Overview," *Int. J. Comput. Appl.*, vol. 79, no. 2, pp. 33–41, Oct. 2013, doi: 10.5120/13715-1478.

[7] I. G. A. Poornima and B. Paramasivan, "Anomaly detection in wireless sensor network using machine learning algorithm," *Comput. Commun.*, vol. 151, pp. 331–337, Feb. 2020, doi: 10.1016/j.comcom.2020.01.005.

[8] M. Rezapour, "Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 11, p. 9637, Sep. 2019, doi: 10.14569/IJACSA.2019.0101101.

[9] Y. Alghofaili, A. Albattah, and M. A. Rassam, "A Financial Fraud Detection Model Based on LSTM Deep Learning Technique," *J. Appl. Secur. Res.*, vol. 15, no. 4, pp. 498–516, Oct. 2020, doi: 10.1080/19361610.2020.1815491.

[10] A. K. Rai and R. K. Dwivedi, "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Jul. 2020, pp. 421–426. doi: 10.1109/ICESC48915.2020.9155615.

[11] R. K. Malaiya, D. Kwon, S. C. Suh, H. Kim, I. Kim, and J. Kim, "An Empirical Evaluation of Deep Learning for Network Anomaly Detection," *IEEE Access*, vol. 7, pp. 140806–140817, 2019, doi: 10.1109/ACCESS.2019.2943249.

[12] A. M. Mubalaike and E. Adali, "Deep Learning Approach for Intelligent Financial Fraud Detection System," in *UBMK 2018 - 3rd International Conference on Computer Science and Engineering*, 2018. doi: 10.1109/UBMK.2018.8566574.

[13] S. Sagadevan, N. Malim, and O. S. Yee, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," *J. Telecommun. Electron. Comput. Eng.*, vol. 10, pp. 23–27, 2018.

[14] K. Yau, K. P. Chow, S. M. Yiu, and C. F. Chan, "Detecting anomalous behavior of PLC using semi-supervised machine learning," in *2017 IEEE Conference on Communications and Network Security (CNS)*, IEEE, Oct. 2017, pp. 580–585. doi: 10.1109/CNS.2017.8228713.

[15] N. Khare and S. Y. Sait, "Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models," *Int. J. Pure Appl. Math.*, vol. 118, no. 20, pp. 825–838, 2018.

[16] D. Zhang, B. Bhandari, and D. Black, "Credit Card Fraud Detection Using Weighted Support Vector Machine," *Appl. Math.*, vol. 11, no. 12, pp. 1275–1291, 2020, doi: 10.4236/am.2020.1112087.