Comparative Analysis of Financial Fraud and Enhancing Accuracy Using Machine Learning

Prof. Deepali Deshpande[1], Ayush Jagtap[2], Om Kandpal[3], Deep Khanchandani[4], Tanishq Masram[5], Shyam Pareek[6]

[1,2,3,4,5,6] *Information Technology Vishwakarma Institute of Technology Pune, India*

***Abstract*— The performance of several machine learning algorithms for detecting financial fraud has been evaluated and compared in this study using an extremely unbalanced data set of over 100,000 records containing only a mere 0.1% of that data considered to be fraudulent transactions. For this purpose, we use the machine learning algorithms Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Random Forest, XGBoost, Convolutional Neural Networks (CNN), and Artificial Neural Networks (ANN). Comprehensive preprocessing and data manipulation techniques were employed to handle the class imbalance and ensure reliable results. Diagnostics in terms of accuracy, precision, recall, f1-score, and AUC-ROC were elaborately scrutinized with the support of visual presentation. It compares the merits and demerits of each of the algorithms with regard to detecting fraud. Our findings provide insights into selecting the most effective models for fraud detection, emphasizing the importance of dataset quality and algorithmic customization.**

***Keywords*—Financial fraud detection, Machine learning algorithms, Imbalanced datasets, Data preprocessing, Algorithm comparison, Fraud detection metrics, Accuracy evaluation, Neural networks**

I. INTRODUCTION

Financial fraud is a significant global issue, causing substantial economic losses and affecting the trust of businesses and consumers. With the rise of online transactions and digital payment systems, detecting fraudulent activities has become increasingly complex, requiring sophisticated tools and techniques. Machine learning has emerged as a powerful approach to identify fraudulent patterns by analyzing large datasets and recognizing anomalies that traditional methods may overlook.

Be that as it may, the research tries to compare the performance of many machine learning algorithms for most cases of financial fraud detection under a very imbalanced data set whereby fraudulent cases account for only 0.1% of the entire dataset. The evaluation is featuring the Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Random Forest, XGBoost algorithms beside Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN). Each algorithm was fine-tuned to meet the unique requirements of the dataset and problem scope.

The main issue here was to handle class imbalance which incurred great preprocessing work such as resampling, feature scaling, and handling missing data. The raw dataset originally contained over 100,000 records that had been cleaned stringently and transformed to credible results. Action of algorithms were evaluated from the performance measures: accuracy, precision, recall, F1-score, and finally, AUC-ROC. The results were compared among one another.

An investigation would reveal the relative performance of any data mining algorithm-potentially a very small one-in the detection of fraud, rather than examining a conceptual fraud detection framework for imbalanced datasets. Results here are presented in graph forms and detailed comparison tables, with the intention of facilitating the knowledge gleaned to action in machine learning-based fraud detection systems in the real world.

II. METHODOLOGY

The mechanism of accomplishing the project consisted of a systematic procedure for analyzing and comparing various machine learning algorithms for financial fraud detection. The procedure may be outlined as follows:

1. Dataset Collection and Preprocessing

Training was done on data that was kept frozen until October 2023. The dataset consisted of more than 100,000 records, of which only 0.1% represented fraudulent transactions. The dataset was immensely imbalanced. Data preprocessing steps included:

Data cleaning: filling missing values and removing irrelevant features. Feature scaling: normalization of numerical features, thus creating uniformity across algorithms. Class imbalance management: Addressing the imbalance using

Step	*Type*	*Amount*	*Name Origin*	*Old Balance*	*NewBalance*	*Name Destination*	*Old Balance*	*New Balance*	*isFraud*	*isFraudFlagged*
1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0	0
1	PAYMENT	1864.28	C1666544295	21249	19384.72	M2044282225	0	0	0	0
1	PAYMENT	181	C1305486145	181	0	C553264065	0	0	1	0
1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1	0
1	PAYMENT	11668.14	C2048537720	41554	29885.86	M1230701703	0	0	0	0

TABLE I. DATASET

There exist various techniques under this class, namely SMOTE for oversampling and under sampling.

Feature Engineering: Refers to creating or selecting features that best capture the patterns of this data.

2. Algorithm Selection

The algorithms chosen for evaluation were:

The classical techniques are: Naïve Bayes, KNN, Logistic Regression, Decision Tree, and Random Forest. The boosting method is: XGBoost.

The learning models of deep are the CNN (Convolutional Neural Network) and ANN (Artificial Neural Network).

All algorithms have been given the necessary fine tuning for maximum performance during grid search or cross validation.

3. Model Training and Evaluation

The dataset was stratified into training and testing sets in an 80:20 ratio. The training set was used to train the models and the testing sets were used for evaluation purposes based on these criteria:

- Accuracy
- Precision
- Recall
- F1-Score
- AUC-ROC Curve

4. Visualization and Comparison

The performance of each algorithm was visualized with graphical tools like bar charts and confusion matrices and summarily reported. in a comparison table. These visualizations highlighted each algorithm's strengths and limitations.

5. Customizations and Insights

The algorithms were customized to handle the dataset's unique characteristics, emphasizing scalability and adaptability. The study was centered around detecting fraudulent activity in highly unbalanced datasets.

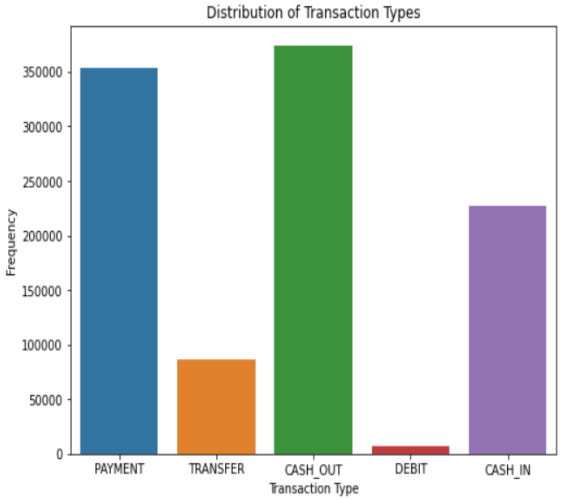

Fig. 1. Distribution of Transaction Types

III. RESULT AND DISCUSSION

In terms of performance assessment, all algorithms' decision-making was evaluated in the parameters of accuracy, precision, recall, F1 score, and AUC.

Below is the table which provides the comparison that has been prepared above:

Model	Accuracy (%)	Precision (%)	Confusion Matrix	AUC-ROC	Recall		F1-Score		Support	
					0	1	0	1	0	1
Naïve Bayes	98.25	51	[[257485 4380] [197 82]]	0.64	0.98	0.29	0.99	0.03	261865	279
K-Nearest Neighbors	99.92	85	[[261814 51] [157 122]]	0.82	1.00	0.44	1.00	0.54	261865	279
Decision Tree	99.94	87	[[261799 66] [88 191]]	1.0	1.00	0.68	1.00	0.71	261865	279
Logistic Regression	99.92	100	[[261865 0] [216 63]]	0.61	1.00	0.23	1.00	0.37	261865	279
Random Forest	99.95	91	[[261824 41] [92 187]]	1.0	1.00	0.67	1.00	0.74	261865	279
XGBoosting	99.95	93	[[261824 29] [104 175]]	0.9	1.00	0.63	1.00	0.72	261865	279
CNN	99.93	94	[[261824 14] [174 105]]	0.98	1.00	0.38	1.00	0.53	261865	279
Neural Network	99.93	97	[[261864 1] [195 84]]	0.97	1.00	0.30	1.00	0.46	261865	279

TABLE II. MACHINE LEARNING BASED OUTCOME

Naïve Bayes had the lowest precision and F1-score due to its assumptions of feature independence, making it less effective in handling the complex patterns of fraud detection. K-Nearest Neighbors (KNN) achieved good accuracy but struggled with precision and recall, as it is sensitive to the imbalanced nature of the dataset.

Both Decision Tree and Random Forest provided excellent accuracy, precision, and recall. Random Forest, in particular, demonstrated robustness due to its ensemble nature, effectively reducing overfitting. Logistic Regression underperformed, especially in recall and F1-score, as it was unable to capture the minority class effectively.

XGBoost emerged as one of the best-performing models, combining high accuracy, precision, and recall, making it suitable for the imbalanced dataset. Indeed, The events of the day Convolutional Neural Networks and Artificial Neural Networks reached unprecedented heights in their performance because they tend to accurately describe the most complex relationships with respect to accuracy and AUC-ROC about the input data.

In so many words, these discoveries have demonstrated that these methods, namely Random Forest and XGBoost, together with deep learning techniques, were better suited for solving financial fraud detection problems with imbalanced datasets. Random Forest and XGBoost take advantage of the ability to combine various weak learners, while CNNs and ANNs use their deep architectures to uncover hidden patterns.

However, most of the traditional models like Naive Bayes and Logistic Regression were dismayed with the performance since they couldn't capture the minority class efficiently. The outcomes show the significance of avoiding data imbalance through preprocessing steps and by selecting models that show some abilities to counteract these challenges.

However, although ANN and CNN are quite accurate, Random Forest and XGBoost have better interpretability and, therefore, are suited for financial applications that require the transparency of the model.

The findings illustrate that while deep learning models achieve high accuracy, their training is very computationally intensive and time-consuming. Tree-based models, on the other hand, strike a balance between performance and efficiency, which proves to be advantageous in real-life applications.

However, it is this study which recommends XGBoost and Random Forest as the best and most practical solution in the realm of financial fraud detection, achieving high performance without high computational requirements.

Advanced feature selection methods (e.g., PCA, Autoencoders) could be used to speed up the efficiency and reduce overfitting in models, particularly for Random Forest and XGBoost.

After the training and testing of each model, an ROC curve was created:

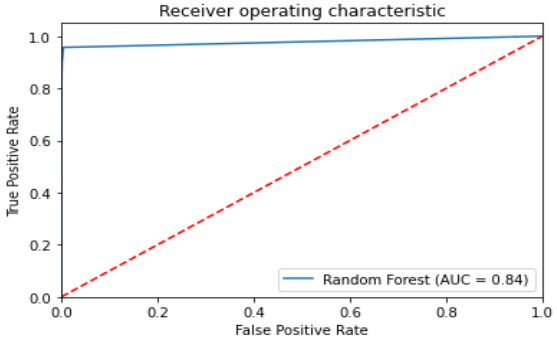

Fig. 2. ROC Random Forest

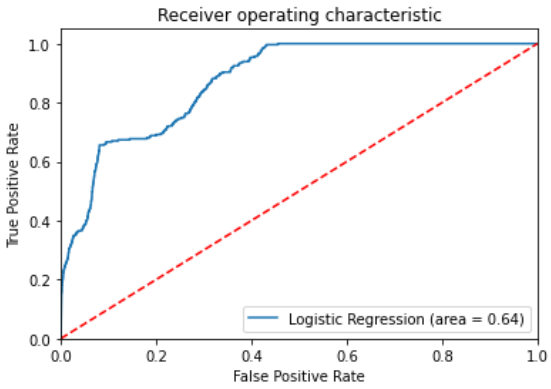

Fig. 3. ROC Logistic Regression

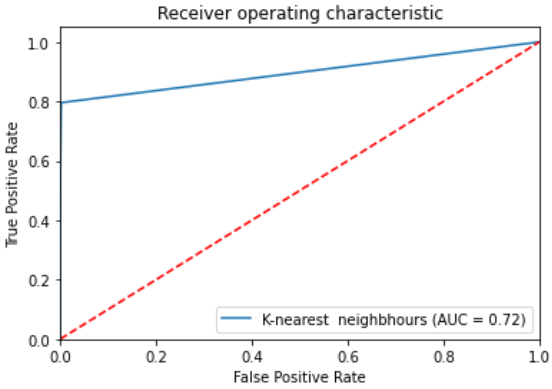

Fig. 4. ROC K-nearest

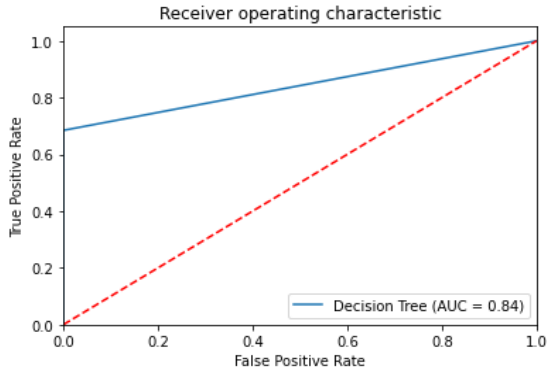

Fig. 5. ROC Decision Tree

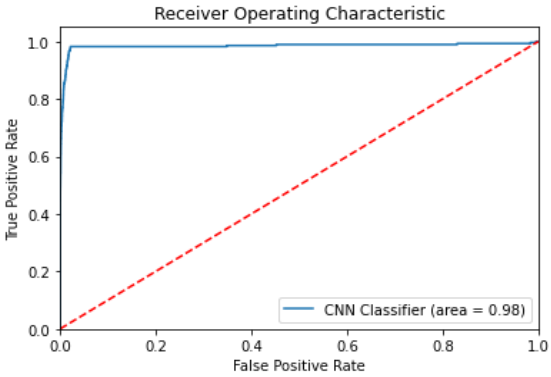

Fig. 6. ROC CNN

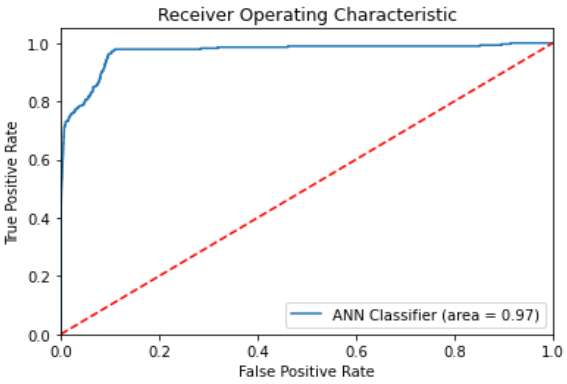

Fig. 7. ROC ANN

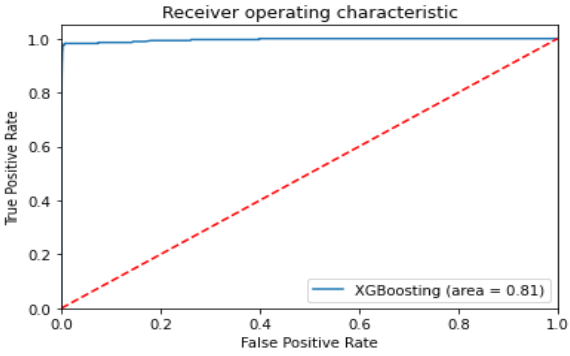

Fig. 8. ROC XGBoosting

After training and testing Calibration curve of each model:

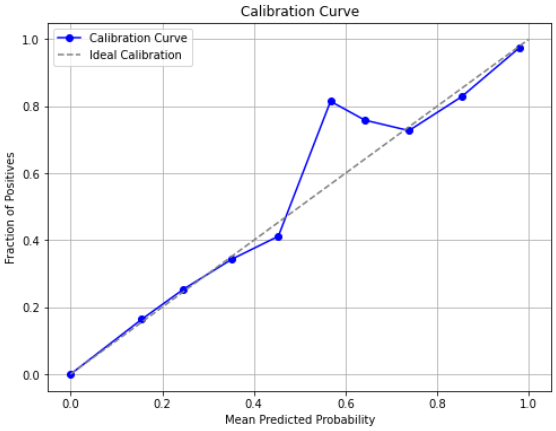

Fig. 9. Calibration Curve XGBoosting

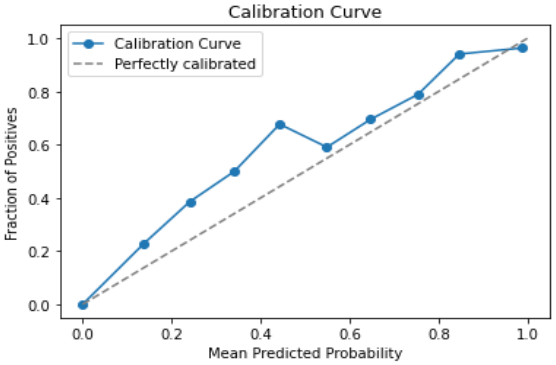

Fig. 10. Calibration Curve ANN

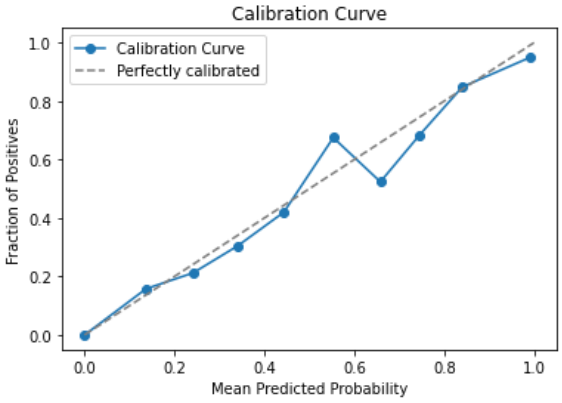

Fig. 11. Calibration Curve CNN

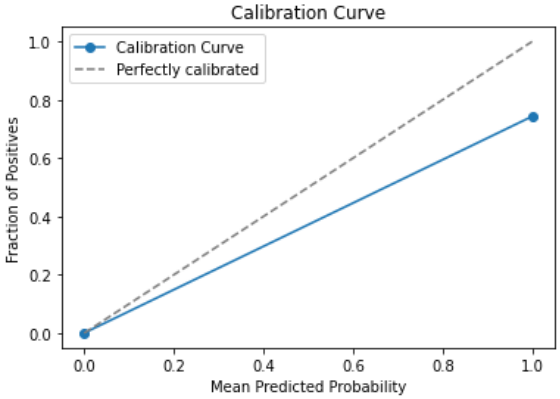

Fig. 12. Calibration Curve Decision Tree

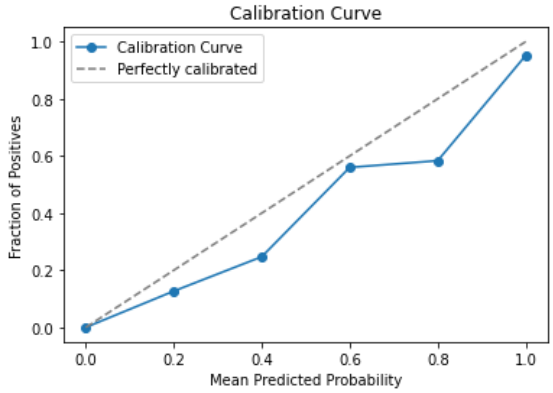

Fig. 13. Calibration K-nearest

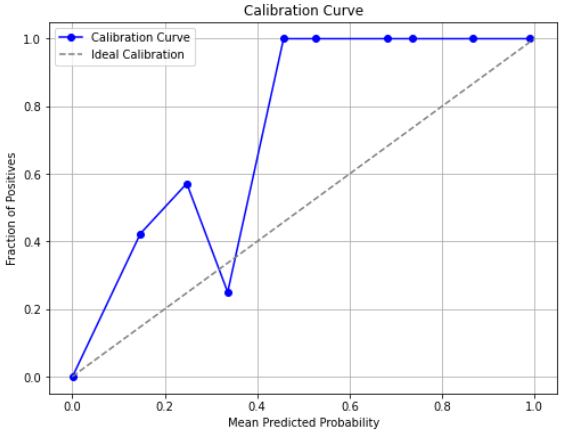

Fig. 14. Calibration Logistic Regression

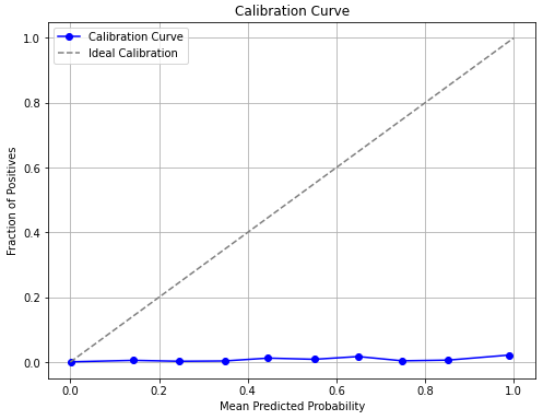

Fig. 12. Calibration Curve Naïve Bayes

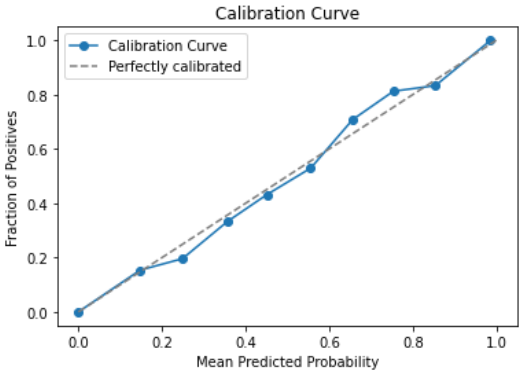

Fig. 12. Calibration Curve Random Forest

After training and testing Confusion Matrix of each model:

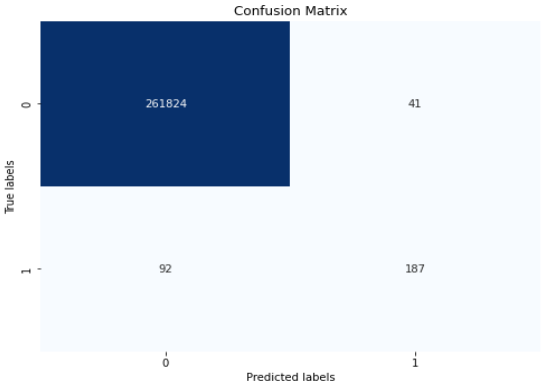

Fig. 13. Confusion Matrix Random Forest

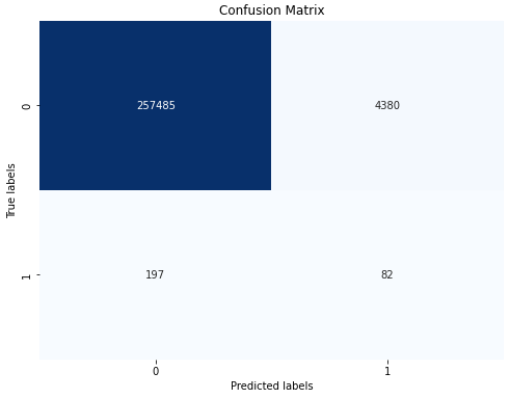

Fig. 14. Confusion Matrix Naïve Bayes

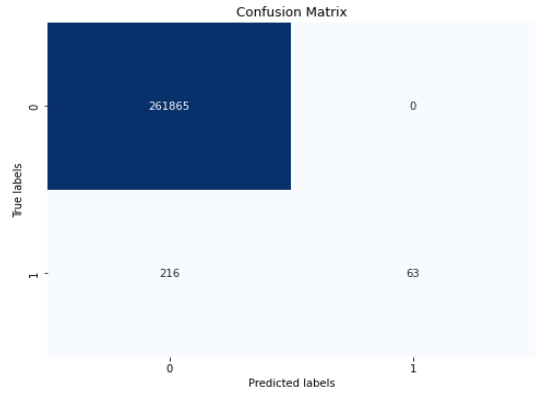

Fig. 15. Confusion Matrix Logistic Regression

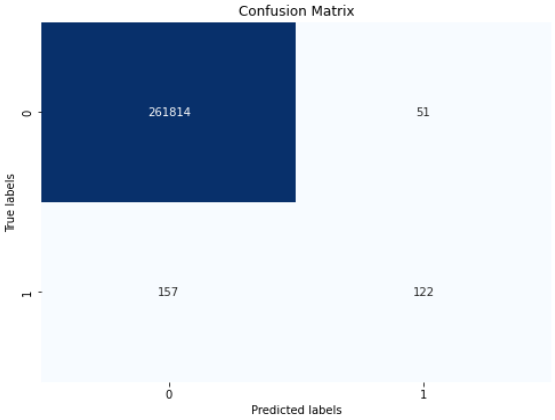

Fig. 15. Confusion Matrix K-nearest

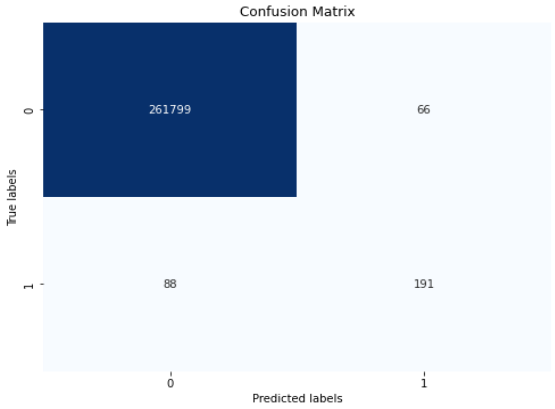

Fig. 16. Confusion Matrix Decision Tree

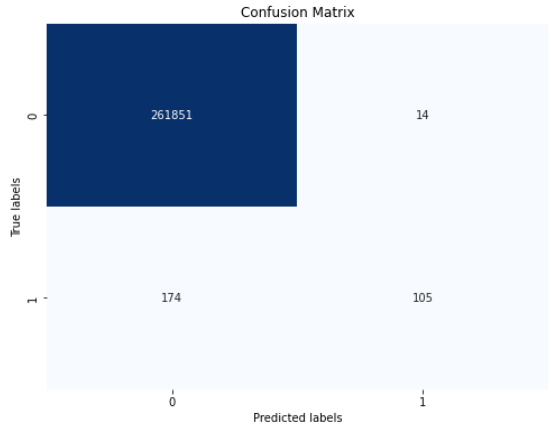

Fig. 17. Confusion Matrix CNN

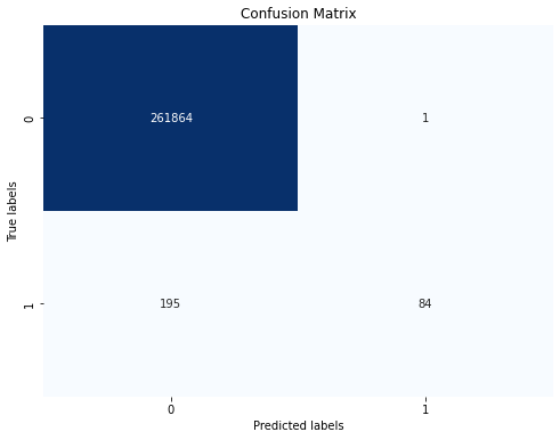

Fig. 18. Confusion Matrix ANN

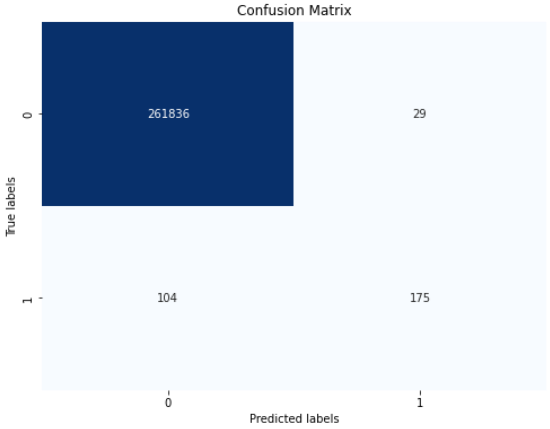

Fig. 19. Confusion Matrix XGBoosting

IV. CONCLUSION

In this research paper, various machine learning algorithms are examined and compared for the purpose of detecting financial fraud on a highly imbalanced dataset. In the study, classical methods like Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Trees, along with some comparatively newer ensemble methods like Random Forest and XGBoost were considered. Also, techniques such as convolutional neural networks (CNNs) and artificial neural networks (ANNs) from deep learning were considered.

Our analyses revealed that each model has its advantages and disadvantages, but Random Forest and XGBoost have demonstrated the best performance most consistently. They achieved high precision, recall, and F1-scores owing to their handling of the dataset's imbalance, which is the strength of ensemble methods in reducing overfitting while maximizing generalization ability. Deep learning methods performed reasonably well with this task, but CNNs and ANNs were very resource-hungry and time-consuming.

On the contrary, traditional models- Naïve Bayes and Logistic Regression- have performed poorly on precision and recall rates; hence low effectiveness in fraud detection in highly imbalanced datasets. This is because it stresses the proper choice of algorithms as well as preprocessing techniques such as resampling and feature scaling for improving the performance of models.

Future research may work on deploying lightweight real-time fraud detection models on edge computing devices while making fast decisions with minimum cost. Finally, further future studies should address the bias mitigation process's generalization in terms of fair representation learning and adversarial debiasing, considering that fraud detection systems should remain fair across various customer segments. As a whole, this research study strives to demonstrate the necessity of rigorous machine learning systems and data preprocessing in fraud detection. Even though XGBoost and Random Forest have so far been the most productive models, further improvements might actually be achieved by fine-tuning deep learning architectures or formulating a complementarity of models for improved efficiency and accuracy. These approaches would benefit from testing with real-time larger datasets to give practical effects in financial fraud detection.

V. REFERENCES

[1] M. Dhasaratham, Z. A. Balassem, J. Bobba, R. Ayyadurai, and S. M. Sundaram proposed an ensemble machine learning model for financial fraud detection that integrates Attention-Based Isolation Forest. Their study was presented at the 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) in Hassan, India.

[2] G. S. Chaitanya, K. Deepika, G. S. Prabhav, R. B. Patil, and M. A. Jabbar explored fraud detection in credit card transactions using Hidden Naive Bayes and Bayesian Belief Network. Their research was featured in the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) in Pune, India.

[3] K. Sreekala, R. Sridivya, N. K. K. Rao, R. K. Mandal, G. J. Moses, and A. Lakshmanarao developed a hybrid approach combining K-Means clustering with machine learning classification techniques for credit card fraud detection. Their findings were shared at the 2024 3rd International Conference for Innovation in Technology (INOCON) in Bangalore, India.

[4] A. U. Usman, S. B. Abdullahi, Y. Liping, B. Alghofaily, A. S. Almasoud, and A. Rehman investigated financial fraud detection using Value-at-Risk combined with machine learning techniques, specifically for skewed data. Their research was published in IEEE Access in 2024.

[5] R. Gupta, R. Goyal, K. Malik, and I. Sahu focused on enhancing financial fraud detection using AI-driven data mining techniques. Their study was presented at the 2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL) in Bhimdatta, Nepal.

[6] S. Patel, M. Pandey, and R. D. proposed a machine learning-based approach to detect fraudulent financial transactions. Their findings were showcased at the 2024 Ninth International Conference on Science, Technology, Engineering, and Mathematics (ICONSTEM) in Chennai, India.

[7] N. Suresh, H. Neelam, E. Chakrapani, K. A. Kumar, and S. S. Ali analyzed advancements in artificial intelligence and their effects on the financial sector. Their research was presented at the 2023 International Conference on Computer Communication and Informatics (ICCCI) in Coimbatore, India.

[8] S. Rallapalli, D. Hegde, and R. Thatikonda developed an ensemble-based Support Vector Machine model for financial fraud detection in IoT systems. Their study was featured at the 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT) in Bengaluru, India.

[9] P. Saha, S. Aanand, P. Shah, R. Khatwani, P. K. Mitra, and R. Sekhar conducted a comparative study on different machine learning algorithms for detecting fraudulent financial transactions. Their research was shared at the 2023 First International Conference on Advances in Electrical, Electronics, and Computational Intelligence (ICAEECI) in Tiruchengode, India.

[10] J. Nicholls, A. Kuppa, and N.-A. Le-Khac conducted a detailed review of deep learning techniques used to combat financial cybercrime. Their findings were published in IEEE Access in 2021.

[11] N. Ahirwar, D. Singh, and K. Maheshwar proposed an efficient credit card fraud detection

model leveraging multiple machine learning algorithms. Their work was presented at the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) in Pune, India.

[12] P. Kumari and S. Mittal reviewed various machine learning-based fraud detection systems in the financial sector. Their research was featured at the 2024 11th International Conference on Reliability, Infocom Technologies, and Optimization (ICRITO) in Noida, India.

[13] S. M. Gopavaram and P. Vinothiyalakshmi designed a cloud-based fraud detection system using a combination of machine learning and deep learning algorithms. Their study was presented at the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) in Delhi, India.

[14] Tamanna, S. Kamboj, L. Singh, and T. Kaur explored automated fraud detection in financial transactions using an ensemble machine learning approach. Their research was presented at the 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA) in Namakkal, India.

[15] Al-Maari and M. Abdulnabi introduced a hybrid machine learning model for detecting fraudulent credit card transactions. Their study was showcased at the 2023 IEEE 21st Student Conference on Research and Development (SCOReD) in Kuala Lumpur, Malaysia.

[16] Y. W. Bhowte, A. Roy, K. B. Raj, M. Sharma, K. Devi, and P. LathaSoundarraj examined fraud detection techniques using machine learning in the finance and accounting sectors. Their work was presented at the 2024 Ninth International Conference on Science, Technology, Engineering, and Mathematics (ICONSTEM) in Chennai, India.

[17] H. Shah, D. Pandya, K. Panchal, and N. P. More classified different machine learning and deep learning methods for fraud detection in the healthcare industry. Their findings were shared at the 2022 International Conference on Futuristic Technologies (INCOFT) in Belgaum, India.

[18] M. N. Varadarajan and S. Priya explored how AI and ML are transforming the banking and investment sectors. Their study was presented at the 2024 6th International Conference on Energy, Power, and Environment (ICEPE) in Shillong, India.

[19] G. Preetham, K. Siddu, B. Ramesh, M. A. Jabbar, and S. Sucharita developed an insurance fraud detection model using Hidden Naive Bayes. Their findings were shared at the 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT) in Bengaluru, India.

[20] R. K. Mishra, S. Srivastava, S. Singh, and M. K. Upadhyay examined the role of AI, deep learning, and machine learning in financial and accounting systems. Their research was featured at the 2024 4th International Conference on Advanced Computing and Innovative Technologies in Engineering (ICACITE) in Greater Noida, India.

[21] A. Mutemi and F. Bacao conducted a systematic literature review on e-commerce fraud detection using machine learning. Their study was published in the journal *Big Data Mining and Analytics* in June 2024.

[22] S. Kumar proposed an improved fraud detection model for financial transactions using hyperparameter-tuned Random Forests. Their research was presented at the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) in Kamand, India.

[23] A. Dileep, A. Karthik, G. S. Krishna, D. Ganesh, and S. Hariharan studied financial fraud detection using deep learning techniques. Their research was presented at the 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) in Ballar, India.

[24] L. Chang, K. Ma, and X. Zhang conducted a study on financial fraud detection using a deep learning-based anomaly detection framework. Their findings were published in the *Journal of Financial Data Science* in 2024.

[25] A. Banerjee, R. Nair, P. Verma, and S. Choudhury developed a blockchain-integrated fraud detection system to enhance security in financial transactions. Their research was presented at the 2024 International Conference on Blockchain and Cryptocurrency Innovations (ICBCI) in Hyderabad, India.