ENSEMBLING LOGISTIC REGRESSION(LR) WITH SMOTE(SYNTHEIC MINORITY OVER SAMPLING TECHNIC FOR CREDIT CARD FRAUD DETECTION

1st Kandi Sneha
dept.computer science Engineering
B V Raju Institute of Technology
Narasapur, Medak District, Telangana state,502313,India
kandisneha2005@gmail.com

2nd Khaja mohammad fazil afzal shareef
dept.computer science Engineering
B V Raju Institute of Technology
Narasapur, Medak District, Telangana state,502313,India
ayazshareef123@gmail.com

3rd Marumanula Bhanu Sri
dept.computer science Engineering
B V Raju Institute of Technology
Narasapur, Medak District, Telangana state,502313,India
Bhanusri753@gmail.com

4th Kolli Manish Moses
dept.computer science Engineering
B V Raju Institute of Technology
Narasapur, Medak District, Telangana state,502313,India
manishmoses138@gmail.com

5th G.Geetha
Assistant professor
dept.computer science Engineering
B V Raju Institute of Technology
Narasapur, Medak District, Telangana state,502313,India
geethareddy0412@gmail.com

***Abstract*—Due to the exponential rise in electronic transactions, credit card theft is becoming a major concern for both consumers and financial institutions. The goal of this research is to create a sophisticated fraud detection system that leverages machine learning techniques. A large dataset that includes transaction details, user activities, and historical patterns is used by the suggested system. The model seeks to precisely identify and categorize fraudulent behaviors in real-time by utilizing supervised learning approaches, such as Random Forest, Support Vector Machines, and Neural Networks. In order to assess the model's effectiveness in reducing false positives and false negatives, it is benchmarked against current fraud detection techniques. The results of this study support continuing efforts to strengthen financial systems against fraudulent activity, so improving**

***Index Terms*—Machine learning,Random Forest,Support Vector Machines (SVM),Real-time detection,False positives,False negatives,Financial systems,User activities, Historical patterns**

I. Introduction

The financial environment has benefited greatly from the proliferation of credit cards and the rise of computerized transactions. Nonetheless, the growing dependence on electronic payment systems has also led to the emergence of a persistent threat: credit card fraud. Fraudulent activities present serious difficulties for both individuals and financial organizations, such as identity theft and unlawful transactions. Sophisticated fraud detection techniques are necessary to meet these challenges.This research explores the field of credit card fraud detection with the goal of creating and improving techniques that make use of cutting-edge technologies, particularly machine learning algorithms. The goal is to develop a strong, flexible system that can quickly detect and stop fraudulent activity. Through the examination of complex patterns, transaction histories, and user habits, this study aims to improve

II. Literature Survey

A fraud prevention system and several authorizations are built into the credit card payment process, which is an electronic means of payment that is dependable and safe. The problem of unbalanced data has been tackled in this work, and the author has suggested an ensemble strategy called ESMOTE-GAN that is based on SMOTE and GAN. so,The GAN model is taking training on the initial samples produced by SMOTE to provide a synthetic dataset. The model's performance has increased by 3.2 percent in fraud detection rate with zero false alarm rate, outperforming the other models used for comparison. However, this strategy has a limitation in that it causes the model to become overfit, and the reason for this is the similarity of the generated samples.[1] Almarshad, Gashgari, and Alzahrani in their study have proposed another data sampling approach, generative adversarial network (GAN).In the technique minority data is given as input to the GAN which generates more similar data and hence results in data balancing. The model has attained a 0.989 AUC value which is highest among all the compared models. Here, fine-tuning on the produced dataset is performed to match the same as setup data to the original data, but the artificially produced data may impact the precision of the model on real-world data.[2] Another study suggests TimeGAN ADV-O, MIMO ADV-O, and GAN GAN-based models for data sampling. In this study, the fraud is seen as a time series, and the suggested model accurately predicts the frauds behavior, allowing for the detection of anomalies.[3] Another method of data sampling, called VAE-GAN, was developed by Ding et al.'s research. In this method, the model creates fictitious data while it is trained on minority data. When synthesized datasets are implemented on logistic regression, decision tree, random forest, neural network, and XGBoost models, this model performs better than other data sampling approaches like GAN, VAE, and SMOTE.[4] where the SMOTE-ENN technique for data sampling has been proposed by the authors. This model is a stacking ensemble model that is tested on unseen data using the stratified 10-fold cross-validation technique. It is trained on a dataset sampled using the SMOTE-ENN approach, with LSTM along GRU acting as a base learner and MLP as a meta learner. The model has obtained 100 percentage sensitivity and 99.7 percentage specificity,outperforming comparable models [5] Habibpour has suggested a method in a different study to identify uncertainties in the CCFD models prediction and for these three approaches to uncertainty quantification. Dropout, Ensemble, and Ensemble in Monte Carlo Confusion matrices, which have two distinct confusion matrices—certain and uncertain categories—are used to compare Monte Carlo Dropout. In this case, the DNN models performance has improved as a result of both MCD and ensemble models. [6] In order to address this issue, Esenogho et al.'s research suggests a data sampling technique called Long Short-Term Together with SMOTE-Edited Nearest Neighbor (SMOTE-ENN), long short-term memory (LSTM). In this case, the dataset is oversampled using SMOTE (Synthetic Minority Oversampling strategy), and sampling is done by ENN. This strategy also prevents the model from being overfit. By contrasting this methods performance with the models on the original dataset , the method's efficacy is verified. Additionally, the model's performance on unobserved data is assessed using a stratified 10-fold cross-validation technique. The suggested model fared better than the other models, with 0.996 sensitivity and 0.998 specificity, indicating the model's correctness.[7] Another data sampling method is presented in the research, and CVAE is shown to beat SMOTE and Random Oversampling when it comes to handling unbalanced data. methods. The FID analysis test on the MNIST and Fashion MNIST datasets is used to assess this study.[8] Based on confusion matrices, sensitivity, and specificity parameters, the proposed model outperformed the other two models by 3 percent using a European dataset. Since the proposed model is supervised, it may perform differently in real-world scenarios where the environment is unsupervised. Overall, this approach outperformed all other compared models.[9] An improved Deep Belief Network (DBN) with Homogeneity Oriented Behavior Analysis (HOBA) has been proposed in one of these studies. This is where the HOBA method is used. to examine client behavior from the past, which is utilized for feature engineering. This technique rejects high-risk authorized transactions and unauthorized transactions at the model's authorization stage, and the alert management module shares the alert with the monitoring department. Real data is used for online testing and offline training of the model. The method includes two predetermined restrictions on the false-positive rate: first, the model is satisfactory if the value is less than 1 percent, and a maximum of 3 percent is acceptable. The HOBADBN model has an F1-score of 0.577 and 58.33 percent, outperforming other models.[10] Three phases make up the ensemble-LSTM model, which is another LSTM-based research model. These phases are base classifier training, middle vector creation and training of voting classifiers. In terms of accuracy, the suggested model has surpassed other models such as LSTM, GRU, ensemble GRU, and Ensemble model; nevertheless, in terms of efficiency, Ensemble GRU has surpassed all of the models combined with the proposed model. Alarfaj et al. present the Convolutional Neural Network (CNN), a well-liked deep learning model.[11] In their research, Misra et al. have supported feature extraction technique, suggesting that a 2-layer Autoencoder be used to extract lower The new dataset and its dimension features are utilized to train three different classification models, including multilayer perceptron (MLP), K-Nearest Neighbor (KNN), and logistic regression (LR). The suggested method has performed better than the most recent data sampling approaches.[12] Another study suggests using the GAN-based models MIMO ADV-O and TimeGAN ADV-O for data sampling. In this research, the fraud is viewed as a time series and the suggested model effectively predicts the fraudster's behavior, allowing the anomaly to be identified.[13] In their study, Li, Liu, and Jiang proposed the Deep Neural Network (DNN) and Full Center Loss (FCL) models, which have outperformed every model that was compared. AUC-PR and F1-score, whose values are 0.805 and 0.879 on the European dataset and 0.813 and 0.825 on the private dataset, respectively, are used to assess the model's performance. By addressing the significance of the loss function in feature extraction and its effect on the model's performance, the work has produced a distinctive addition.[14] A neural network-based model called the Feed-forward Neural Network model has been proposed in another study; it has proven to be more effective than the Hybrid Ensemble Learning model.Run a model with all the parameters set to 3.8 , 2.1 and 5.5 percent, respectively, for recall transaction, recall (card), and cost reduction rate. The document contains the methods for all of the neural network-based CCFD models that have been suggested.[15]

III. PROBLEM STATEMENT

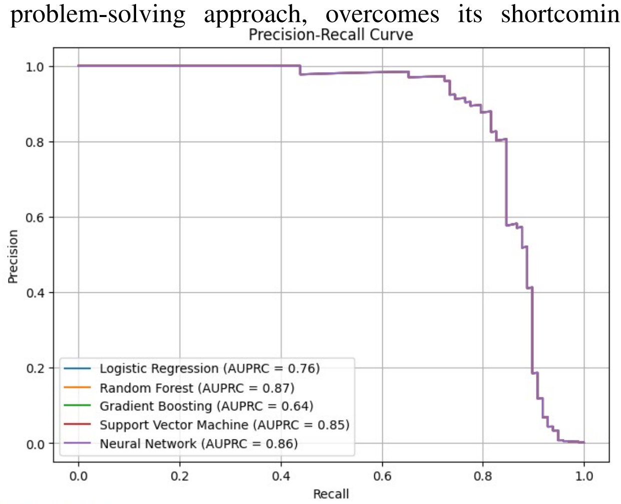
mainly there is not all the doubtful transactions are considered fraudulent. It is commonly called as false positive (FP) which means that the case was not fraud although it was flagged as being scam. This process of state ture for each transaction that outliers from the cardholders normal routine brings doubt about different activities mainly, the expenses related with explore an unknown no. of false positives are high.

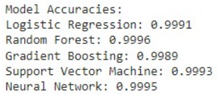
IV. OBJECTIVES

Implement a system that continuously monitors transactions in real-time, enabling swift identification and response to suspicious activities to enhance fraud prevention. Minimize false positives to a level where legitimate transactions are rarely incorrectly identified as fraudulent, thereby enhancing the system's precision and reducing unnecessary disruptions for cardholders.

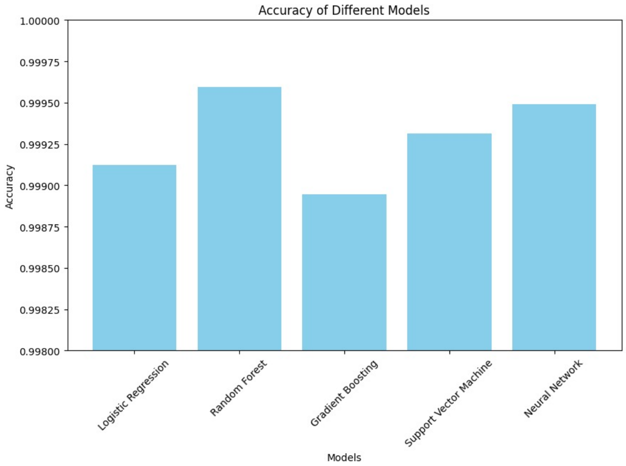
V. EXISTING WORK

Since credit card fraud detection systems are a well-studied area, there are various algorithms and strategies for creating credit card fraud detection programs. Card fraud detection systems include Critical Cost Tree , Vector Assist Machine (SVM), Random Forest, and more. Follow the whale herd optimization algorithm to find available values. Use the BP network to correct the incorrect value card fraud This scheme, which uses a whaling algorithm and a problem-solving approach, overcomes its shortcomings.





```
Model Accuracies:
Logistic Regression: 0.9991
Random Forest: 0.9996
Gradient Boosting: 0.9989
Support Vector Machine: 0.9993
Neural Network: 0.9995
```



VI. PROPOSED WORK

A. Data preprocessing

Open the dataset with the credit card transaction information. To standardize the 'Amount' column, use StandardScaler to perform feature scaling. Divide the data into the target variable (y) and features (X). Utilizing SimpleImputer and a mean technique, impute missing values in the features.

B. Handling class Imbalance

Open the dataset with the credit card transaction information. To standardize the 'Amount' column, use StandardScaler to perform feature scaling. Divide the data into the target variable (y) and features (X). Utilizing SimpleImputer and a mean technique, impute missing values in the features.

C. Model training

To correct the class imbalance in the target variable (fraudulent and non-fraudulent transactions), use the Synthetic Minority Over-sampling Technique, or smote To balance the distribution of classes, Smote creates samples for the fraudulent transactions.

D. Training Models

Apply the resampled training data to a Logistic Regression model for training. Because of its simplicity, interpretability, and efficiency in binary classification tasks, logistic regression is selected as the classifier.

VII. RESULT

A. Evaluation of Model Performance

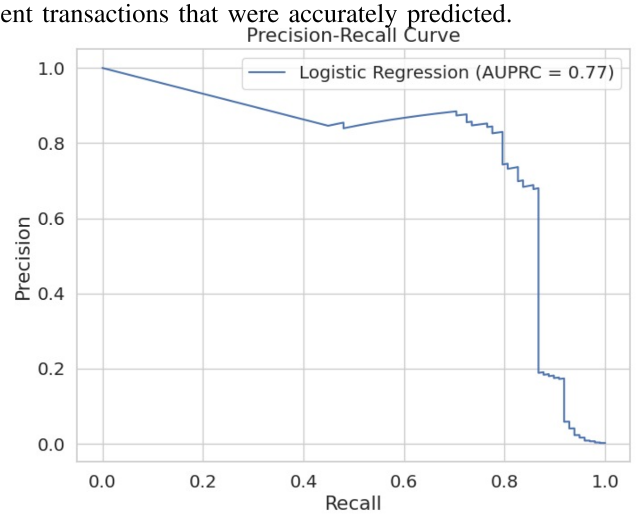
The code trains many ml models for the detection of credit card fraud, including random forest, gradient boosting, support vector machine, neural network, and Logistic Regression

B. Each model is assessed based on the subsequent metrics

The trade between recall and precision is summed up by the (AUPRC).

C. Precision

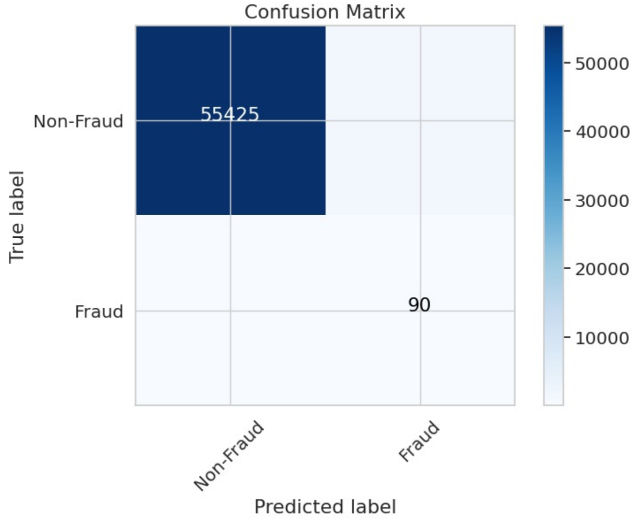
Shows plpercentage of all projected positive cases fraudulent transactions that were accurately predicted.



For unbalanced datasets, the F1-score a harmonised of precision and recall is helpful.

D. Confusion Matrix

It gives atrue positive, false positive, true negative, and false negative predictions in outcome



and the goal of reducing false positives and false negatives, these measures aid in evaluating the models accuracy in identifying fraudulent transactions.

E. Visual Representations

The code creates visual representations of the Precision-Recall Curve and Confusion Matrix for each model. The trade between recall and precision at different thresholds is represented visually by the Precision-Recall Curve. A graphical representation of the model's predictions is given by the confusion matrix, which shows true positive, false positive, true negative, and false negative values.

F. Model Comparison

Stakeholders can determine which algorithm for credit card fraud detection is the more effective by comparing and contrasting the performance data of several models Better performers are models with reduced false positive and false negative rates, as well as higher accuracy, F1-score, AUPRC, and F1-score.

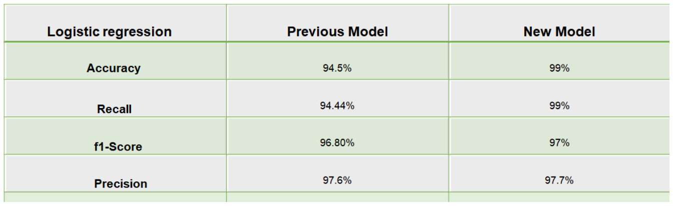
G. Model Choice and Implementation

Stakeholders can choose the best model for real-time fraud detection in a production setting based on the evaluation results

VIII. Discussion

The code trains and tests several machine learning models, exhibiting a strong methodology for credit card fraud detection. Stakeholders can compare performance measures like as AUPRC, precision, and F1-score to determine which algorithm is best suited for implementation. Visual perception of model performance is aided by graphical representations, which promotes well-informed decision-making. All things considered, the code provides a thorough framework for creating precise fraud detection systems, which is essential for guaranteeing financial security.

IX. TABLE



Logistic regression	Previous Model	New Model
Accuracy	94.5%	99%
Recall	94.44%	99%
f1-Score	96.80%	97%
Precision	97.6%	97.7%

X. Conclusion

The code offers a methodical way to compare and evaluate models, making it possible to choose the best model for the fraud detection task. The accuracy, F1 score, and recall percentages of Logistic Regression can be used to evaluate its performance and provide important insights into how well it detects fraudulent transactions. Additional refinement, testing, and optimization of hyperparameters can be required to enhance model performance and fulfill particular application needs. With the freedom to add new features, models, and evaluation methods in later iterations, the code acts as a basis for more sophisticated fraud detection systems.

References

[1] F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. Al- Hadhrami, "Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection," IEEE Access, vol. 11, pp. 89694–89710, 2023, doi: 10.1109/ACCESS.2023.3306621.

[2] F. A. Almarshad, G. A. Gashgari, and A. I. A. Alzahrani, "Generative Adversarial Networks-Based Novel Approach for Fraud Detection for the European Cardholders 2013 Dataset," IEEE Access, vol. 11, pp. 107348–107368, 2023, doi: 10.1109/ACCESS.2023.3320072.

[3] D. Lunghi, G. M. Paldino, O. Caelen, and G. Bontempi, “An adversary model of fraudsters’ behaviour to improve oversampling in credit card fraud detection,” IEEE Access, pp. 1–1, 2023, doi: 10.1109/ACCESS.2023.3337635.

[4] Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, “Credit Card Fraud Detection Based on Improved Variational Autoencoder Generative Adversarial Network,” IEEE Access, vol. 11, pp. 83680–83691, 2023, doi: 10.1109/ACCESS.2023.3302339.

[5] I. D. Mienye and Y. Sun, “A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection,” IEEE Access, vol. 11, pp. 30628–30638, 2023, doi: 10.1109/ACCESS.2023.3262020.

[6] M. Habibpour et al., “Uncertainty-aware credit card fraud detection using deep learning,” Eng Appl Artif Intell, vol. 123, p. 106248, Aug. 2023, doi: 10.1016/j.engappai.2023.106248.

[7] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, “A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection,” IEEE Access, vol. 10, pp. 16400–16407, 2022, doi: 10.1109/ACCESS.2022.3148298.

[8] V. A. Fajardo et al., “On oversampling imbalanced data with deep conditional generative models,” Expert Syst Appl, vol. 169, p. 114463, May 2021, doi: 10.1016/j.eswa.2020.114463

[9] E. N. Osegi and E. F. Jumbo, “Comparative analysis of credit card fraud detection in Simulated Annealing trained Artificial NeuralNetwork and Hierarchical Temporal Memory,” Machine Learning with Applications, vol. 6, p. 100080, Dec. 2021, doi: 10.1016/j.mlwa.2021.100080

[10] X. Zhang, Y. Han, W. Xu, and Q. Wang, “HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture,” Inf Sci (N Y), vol. 557, pp. 302–316, May 2021, doi: 10.1016/j.ins.2019.05.023.

[11] J. Forough and S. Momtazi, “Ensemble of deep sequential models for credit card fraud detection,” Appl Soft Comput, vol. 99, p. 106883, Feb. 2021, doi: 10.1016/j.asoc.2020.106883.

[12] S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, “An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction,” Procedia Comput Sci, vol. 167, pp. 254–262, 2020, doi: 10.1016/j.procs.2020.03.219.

[13] H. Tingfei, C. Guangquan, and H. Kuihua, “Using Variational AutoEncoding in Credit Card Fraud Detection,” IEEE Access, vol. 8, pp. 149841–149853, 2020, doi: 10.1109/ACCESS.2020.3015600.

[14] Z. Li, G. Liu, and C. Jiang, “Deep Representation Learning With Full Center Loss for Credit Card Fraud Detection,” IEEE Trans Comput Soc Syst, vol. 7, no. 2, pp. 569–579, Apr. 2020, doi: 10.1109/TCSS.2020.2970805.

[15] E. Kim et al., “Champion-challenger analysis for credit card frauddetection: Hybrid ensemble and deep learning,” Expert Syst Appl, vol. 128, pp. 214–224, Aug. 2019, doi: 10.1016/j.eswa.2019.03.042.