International Research Journal of Multidisciplinary Sciences ISSN:AWAITED
VOL-1 ISSUE-6 June 2025 PP:24-34

DEVELOPINGANAI-DRIVENFRAUDDETECTIONSYSTEM:A MACHINE LEARNING APPROACH

[1]BhawnaKaushik,[2]Anam Shariq
1.bhawna.kaushik@niu.edu.in,NoidaInternationalUniversity
2.anam.s.khan92@gmail.com,Birla Public School-Behrain

Abstract - The detection of financial fraud is critical for maintaining the safety and integrity of online transactions. To this end, this research articulates a holistic approach to the identification of fraudulentactivities by using machine learning algorithms, Logistic Regression, K- Nearest Neighbors, Decision Trees, Random Forest, and XGBoost classifiers. Our method begins with exploratory data analysis in an attempt to visualize trends about transaction transactions and identify significant features, followed by model train ingand evaluation on a well- structured dataset. We preprocess the data using feature engineering and standard scaling, and then compare multiple models based on their performance. Among the tested models, Random Forest proved to be the most accurate, making it a reliable solution for fraud detection. Additionally, we implemented a user input system that allows real-time fraud prediction based on specific transaction details. This study contributes to the development of automated fraud detection systems, helping financial institutions reduce risks and prevent losses. The implementation, done using Python libraries and documented in Jupyter Notebook, emphasizes simplicity and flexibility.

Keywords - Detection of financial fraud, Machine learning, Random Forest (RF), Logistic Regression(LR), XGBoost (XGB), Decision Tree (DT), K-Nearest Neighbors (KNN), Feature engineering, Data visualization, fraud(f), non-fraud(n-f).

I. INTRODUCTION

Financial fraud detection then remains one of themost essential ways of ensuring that the financial systems are intact and secure. Together with the development of online transactions and digital payment systems, the aspect of fraud has become so widespread and extensive, and this is a threat both to financialinstitutionsandconsumersalike.Traditional fraud detection techniques have utilized mainly rule- based systems, along with manual oversight; while effective in specific cases, they are handicapped by their failure to change to evolving patterns of fraud. Additionally, the systems are inefficient and tend to take longer todetectfraudulent transactions,resulting in significant losses before fraud transactions are identified. All these facts have enhanced the call for automated, scalable, and real-time solutions for fraud detection.

What has driven these advancements is machine learning and data analytics: it could recognize patterns, make predictions, and learn from historical data with the help of a developed model. Such models, especially classification algorithms, havethus far shown huge promise in detecting fraudulent transactions with considerable accuracy. These ML techniques, such as LR, DT, RF, KNN, and XGB, are now some of the inevitable tools in the fight against financial fraud, which are capable of spotting anomaliesinlarge-scaledatasetsunlikelytobecaught by traditional methods.

Simple rule-based systems were originally employed by financial institutions to detect fraud. They worked wellforthosestraightforwardcasesthatinvolved

large withdrawals from an account or transactions from a geographically distant location. However, since fraudsters became very sophisticated, these techniques became not enough in time. As fraudsters started using new-age tactics like money laundering andidentitytheft,itbecamestrictlynecessarytohave more robust techniques. Machine learning, which offers an ability to analyze super amount data and learnintricaterelationships,hasbeenhailedasagreat way to solve this problem.

Based on the following research, we offer an all- round approach employing the use of Python to implement models of machine learning algorithms that detect financial fraud. The study begins with an exploratory data analysis of a transaction dataset to find some patterns and relationships which point out the presence of fraud. Next up is feature engineering, applying such preprocessing in order to optimize model performance. For classification, we train and test numerous classification models such as Logistic Regression, KNN, Decision Tree, Random Forest,and XGBoost in order to classify fraudulent transactions.

To ensure reliability in our models, we split the dataset into subsets for training and testing purposes. Standard scaling techniqueshavebeen used tofurther enhance the accuracy of the models. We then analyze and compare the performance of the models invarious metrics like accuracy, precision, recall, and F1-score. We also develop a system with a user-input mechanism for real-time fraud detection. This system will allow users to input transaction details, where feedback on where the ratransaction is likely to be fraudulent or not can be provided immediately.

With advanced machine learning, this work introduced a high- performance, scalable fraud detection system that could make real- time predictions to reduce the risks of fraud and increase customer trust through even safer transactions.

Thepaperisstructuredas:Section2-Reviewsrelated workdoneinfinancialfrauddetection,highlighting the evolution of machine learning applications in this domain.Section 3: Introductionto the data set used for training and testing models. Preprocessing aswell asfeatureengineeringdoneonthedataset.Section4:Meth odologydescriptionofusedmodels,criteriaof evaluation,aswellashyperparametertuningstrategies. Section 5: Results and discussion, including discussion of Model's performance, especiallyaccuracy,analysisofconfusionmatrices and general detection efficacy. Lastly, section 6- will concludethepaperwithfindingsanddiscussionon implicationsforfinancialfrauddetectiontogether with recommendations for future research directions..

II. RELATEDWORK

Financial fraud detection has come into the limelight because there is a strong necessity for enforcing effective systems against developing fraud patterns.In the context of fraud detection, ensemble methods are prominent and an efficient way to improve fraud detection and minimize false positives and negatives.

PreviousResearchonFinancialFraud Detection:

1. Zareapoor and Shamsolmoali [1] proved the feasibility of the credit-card fraud-identification logistic regression model. This was, however, onan imbalanced dataset, as the model could not handle therare transactionsof fraud. Resampling techniques along with advanced featureengineering improved precision, recall, and F1- scores in this case. Such worklaid a foundationbut highlighted the need for advanced algorithms to managelarge- scaleimbalanceddatasets.
2. J. Sah et al.[2] used RandomForest algorithmsfor credit card fraud detection with accuracy at 99.2%. Theresultsofthatstudyindicatethepossibilitythat ensemble methods may outperform the performancesofasinglemodelwhensubtle fraud patternsarehiddeninlargedatasets.
3. ClassImbalanceProblem M.DalPozzoloetal.[3] addresseditusinganensembleofDecisionTrees and K-Nearest Neighbors (KNN). The objective of theresearchers was to minimizefalsenegative on fraud detection. Using the combination of several algorithms improved the performance of fraud detection, significantly meaningful for reducing financiallossescausedbyfalsenegatives.
4. R.TiwariandS.Kumar[4]proposedahybrid system by incorporating Logistic Regression, Random Forest, and XGBoost for the detectionof credit card fraud. The approach learned from each of the individual models' strengths to increase accuracy and robustness. Their model was able to be highly accurate at detecting fraud transactions, with reduced false positives and improved recalls.

Additional Contributions:

5. Afterwards, Wang et al.[5] suggested XGBoostin credit card fraud detection, since it can simultaneously work with high-dimensional and imbalanced data. Because its gradient boosting can master minor patterns, it is a pt for rapidly changing fraud strategies. For example, their experiment had achievedanaccuracyrateof99.5%.
6. Zhao et al. [6] explored a deep learningtechnique to detect fraud in transaction data by using CNN and achieved a reasonable accuracy of 98.8%. However, their actual use washindered by a high computation cost and the requirement of large amounts of data.
7. Patel et al. in [7] have proved that ensemble methods such as Random Forest and Ada Boost are superior to traditional algorithms, such as SVM and Decision Trees, where class imbalance is a problem, and specifically in the case of highly decreased false positives.

Performance of Various Models in Current Research:

In this experiment, we compared the performance of somemachine learning algorithms such as high accuracy and precision, recall, F1-score, and efficiency at working with imbalanced datasets.

1. For the Logistic Regression Algorithm, its accuracy was 0.9982. In case its precision ishighly limited while detecting fraudulent transactions as a class 1 because it does have a very low F1-score that was recorded at 0.39, indicating that the algorithm really doesn't work well with imbalanced.

2. KNN appears to perform better than Logistic Regression since it was able to achieve an accuracy of 0.9995 and a score of 0.77 for class1, hence it performs well in fraudulenttransaction detection although it incurs huge computational costswhenitsimmenseinnumber.
3. The Decision Tree Classifier excelled with an accuracy of 0.9997 and an F1-score of 0.89 for class 1, as it had strong capabilities that better captured complex data patterns to detect fraud.
4. Random Forest Classifier achieved an accuracy of0.9997withanF1-scoreof0.87forclass1.As an ensemble model, it handled complex fraud patternandimbalanceddataveryeasily.
topmodel,XGBoostClassifier,yielded

0.9998 accuracy and scored 0.91 on the F1-score for class 1. XGBoost is an appropriate method to identifytheslightestpatternsoffinancialfraudas it can handle high-dimensional and imbalanced data.

PerformanceComparisonofMachineLearning Models

Model	Accuracy	Precision	Recall	F1-Score
LR	0.9982	0.35	0.44	0.39
KNN	0.9995	0.87	0.69	0.77
DT	0.9997	0.91	0.88	0.89
RF	0.9997	0.98	0.78	0.87
XGB	0.9998	0.96	0.86	0.91

III. DATASETDESCRIPTION

Publisher:Kaggle
Title: Online Financial Fraud Dataset
URL:https://www.kaggle.com/code/rashmiek99/financial-fraud-detection/input

ThedatasetusedfortheanalysiscamefromKaggle,a datasetforfrauddetectionwithfinancialtransactions. Key features of the dataset include transaction type, amount,accountbalances,andfraudindicators.There are 11 features involved: step, type, amount, name Orig, old balance Org, new balance Orig, name Dest, old balance Dest, new balance Dest, is Fraud, and is Flagged Fraud. These will be used in analysis of transaction patterns to identify genuine versus fraudulent transactions.

Using sophisticated machine learning techniques, the research classifies transactions as fraudulent or not. The dataset is also useful for training models in real-timefrauddetection,enhancingfinancialsecurityand user protection.

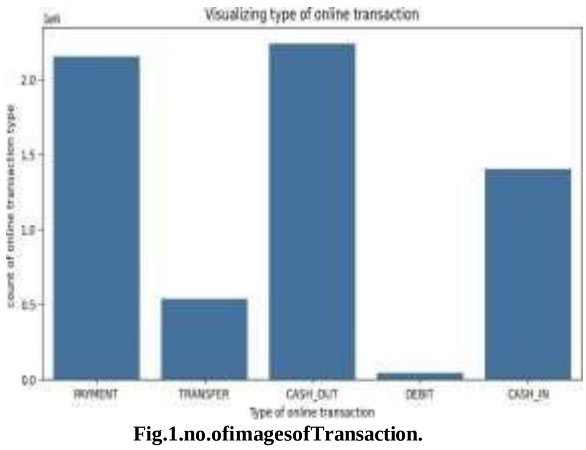
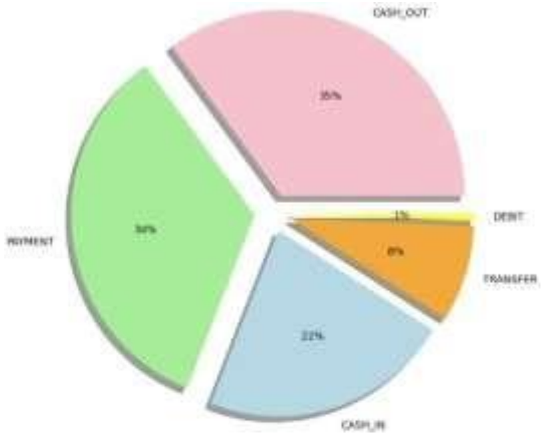


Fig.1.no.ofimagesofTransaction.



It should, however, be noted that availability of the dataset on Kaggle warrants its conformity to the particular terms and conditions put in place by Kaggle. Researchers are highly advised to read throughanyrelateddocumentationaccompanyingthe dataset into their custody, including licenses and ethical considerations likely to be encountered, to ensure proper use of the data and thus adhere strictly to instituted privacy guidelines.

IV. METHODOLOGY

We used techniques of machine learning to develop and comparea model for detecting financial fraud.The approach that focused on ensemble methods stems from their ability to improve the accuracy of results and handle issues related to imbalanced datasets.

Data Collection and Preprocessing: We used the publicly available dataset which consists of creditcard transaction records of actual purchases besides fraud.Italsohadahighlyimbalancedclassasithada largenumberofnonefraudulenttransactionsthatwere to a tiny fraction compared to fraudulent ones. Resampling techniques-the application ofunder samplingof the majority class and oversampling ofthe minority class were used in an attempt to balance the training data.

Data preprocessing: tasks also included normalization to bringall feature values into auniform range and feature engineering to enhance model performance by generating new variables and removing irrelevant ones.

Model Implementation: Various machine learning algorithms have been implemented using Jupyter Notebook as the principal development environment. The trained and tested models are as follows:

Logistic Regression: The mixture has added a baseline linear model for comparison to methods of greater complexity.

K-Nearest Neighbors(KNN): A distance-based approachwhichwassupposedtodobetterandproved more computationally costly on larger datasets.
Decision Tree Classifier: A non-linear modelcapable of capturing complex patterns in the data.

Random Forest Classifier: The ensemble method combined several decision trees to reduce variance and improve generalization.

XGBoost classifier is one of the ensemble techniques,well reputed for its effectiveness in handling high dimensional data and imbalanced datasets as expected to perform better.

Differentmodelsweretrainedovertheprocesseddata and all algorithms, keeping in view the possible performance improvement through hyperparameter tuning, were subjected to this process. The best set of parameters for each model wasfound using gridsearchalongwithcross-validationandresultsshowed improvement in accuracy, precision, recall, and F1-score.

Evaluation Metrics:

With the class distribution of the database being skewed to extreme imbalance, traditional accuracy was not sufficient to use for estimating model performance. Instead, we relied on other evaluation.

-Accuracy (for Class 1 -Fraudulent transactions): Percentageofactuallyfraudsasdetectedbypredicted frauds.

-True Fraud Recall Class 1: The percentage of true frauds which are classified by the model correctly.

-F1 Score (for Class 1): With precision and recall averaged, it provides just one score that reflects the overallmodelperformance.Falsepositiverate-thatis, how legitimate transactions are classified as fraud-and false negatives-the failure to classify actualfraud-were monitored since financial fraud-detection systems rely on both of these entities.

The above performance metrics compared the performances of the models: XGBoost performed the best, as it achieved the highest level of accuracy, precision, recall, and F1-score. Indeed, the ensemble character of Random Forest and XGBoost proved to be highly effective in detecting fraudulent patterns within an imbalanced dataset.

A. Toolused

- JupyterNotebook:Fordeveloping,documenting, and experimenting with different models.
- Scikit-learnandXGBoostlibraries:Forthe implementationofmachinelearningalgorithms

andensemble techniques.

- Orange Software: For data preprocessing tasks, including feature selection, normalization, and resampling.

B. ProposedModel

DataCollectionandPreprocessing:

I also used apubliclyavailablecreditcard transaction dataset loaded into Jupyter Notebook, with both fraudulent and honest transactions.

- Handling Imbalanced Data: Major Class Resampling Techniques: Undersamplingincluded the nonfraud class, and oversamplingfor the minority class was the fraud.
- Normalization: The feature values, such as the transaction amount, balances, were normalized using normalization.
- Feature Engineering: appropriate new features were created as well as irrelevant ones removed, to enhance model performance.

ModelDeploying:

- Logistic Regression: Used as a baseline model that one would compare against more complex classifiers.
- K-Nearest Neighbors (KNN): This is for distance- based classification; in big data, it is pretty time- consuming.
- Decision Tree Classifier: This is a non-linear model used to capture complex patterns from the dataset. Random Forest Classifier: An ensemble model that combine multiple decision trees to improve on theperformance of classification and generalization.
- XGBoost Classifier: Another powerful ensemble modelknownforhandlinghigh-dimensionaldata and imbalanced datasets.

Hyperparametertuning:

- Grid Search & Cross-Validation: Implemented for all models to find the best hyperparameters, e.g., number of trees, maximum depth for Random Forest. This ensured that optimal model performance was obtained.
- Cross-Validation: Cross-validation controls the overfitting and ensures the models generalize well to unseen data.

ModelTrainingandTesting:

Allthesemodelsweretrainedonthispreprocessed training data.
Everymodel'saccuracy,precision,recall,andF1scoreare calculated.
Result: Random Forest Classifier Hyperparameter tuning on thetest data set produced 98.74% accuracy using a Random Forest classifier.

During the simulation, fraudulent transactions were well detected at a low rate of false positives and false negatives.TestingonUnseenData:Thetrainedmodel was used to a test dataset to predict the fraud detection.Theperformancemetricsonthetestdataset model's.

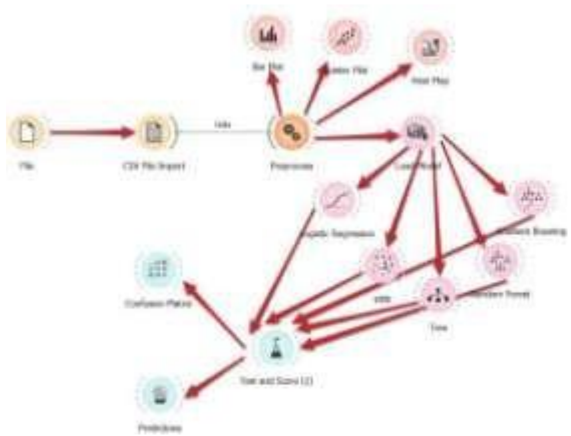


Fig.3.proposedmodeltopredictfrauddetection

Code:- fromsklearn.ensembleimportRandomForestClassifier #InitializetheRandomForestClassifier
rf_model =
RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)

#Trainthemodelrf_model.fit(X_train,y_train) #Testthemodel
*accuracy = rf_model.score(X_test, y_test) print(f"Modelaccuracy:{accuracy*100:.2f}%")*

This methodology combined powerful techniques of machine learning with effective preprocessing and data reduction, ensuring the detection of financial fraud to be highly accurate and reliable.Thepowerof Random Forest along withclustering techniques is quite efficient in order to detect fraudulent transactions from large datasets. Its robustness and efficiency guarantee the fraud detection system. Further, this employed hyperparameter tuning and cross-validation with the addition of real-time prediction capabilities to the model for direct fraud detectionbasedonuserinputs.Thismakesthesystem sound enough for practical application in large-scale financial environments.

Conclusion The methodology of financial fraud detection proposed hereby successfully addresses the challenge of the imbalanced data through various resampling techniques and pertinent preprocessing with regard to feature engineering and standardization. Several models of machine learning, namely Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, and XGBoost, were implemented and hyper parameter optimized for robustperformance.Inevaluatingtheperformanceof the models, precision, recall, and F1-score were considered, and it can be seen that Random Forestand XGBoost are top-level models, so XGBoost was declared best model.

Basedonit,anotherreal-timefraudpredictionsystem is developed using user inputs, so the methodology is comprehensive well-suited for real-world financial fraud detection.

C. AlgorithmUsed

This research focuses on applying several machine learning models toward classification tasks for the purpose of fraud detection within transactions as legitimateorfraudulent.Fromthemodelsconsidered, there are Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and XGBoost-different approaches in each toward fraud detection.

- ***Logistic Regression:*** As a base line model, Logistic Regression's simplicity makes itefficient for binary classification tasks. Although limited by its linear nature, it offered insightsinto the significant features

- ***K-Nearest Neighbors (KNN):***It is one of those non- parametric models which classified transactionsbasedonthesimilaritytonearby data points. Sufficient for smalldatasets,thecomputationalcostincurredbylar gedatasetsmakesitunsuitableforhighly extensive financial data.

- ***Decision Trees:*** It was a tree-based model offering great interpretability through segmentation of data on decision rules. Decision Trees helped significantly in pointing out major features affecting patterns of fraud.

- ***Random Forest:*** It is an ensemble method that combines multiple decision trees, hence averts overfittingandhencegivesbetteraccuracy.Thisrobu stmodelperformedquitewellonthedatasetand wasone of the top contenders for fraud detection.

- ***XGBoost:***Thisisagradientboostingalgorithmthat,inourtests,performedasthemostaccurate.XGBoost handledimbalancesowellandwasfine-tuned furtherbyoptimizinghyperparameterstoachieve high precision, recall, and F1-score

In addition, MLP neural network usage was put to application in identifying the non-linear patternwithinthedatatoenhancefrauddetectioncapabilities beyond that of a linear model.

Data Preprocessing

ThedatasetisobtainedfromKaggle,whereseveral records of transactions with features such as type, amount, account balances, and fraud indicators are contained. Because fraudulent transactions are very rare, the use of resampling techniques, such as oversampling and undersampling, was applied in order to balance the data. Feature engineering and normalization wereappliedto prepare thedatafor the machine learning model.

Model Appraisal:
Performance of Each Model Accuracy, Precision, Recall, and F1- score will be used to evaluate performance. This dataset was split into training and testing subsets, and standard scaling for improved accuracy is conducted on the models. The best- performing model in this was XGBoost with itsability to deal with imbalanced datasets and data high dimensions.

UserInputSystem:
This led to the development of a real-time fraud detectionsystemthatallowsuserstoinputtransaction details as the basis for obtaining predictions about possiblefraud.Thisisascalablesolutionforfinancial institutions aiming at reducing fraud risks and building customer trust.

Conclusion
These two, Random Forest and XGBoost, proved to provide both strength and effectiveness in fraud detection while the NLP neural network seems to work in identifying some complex patterns. This study gives importance to the use of automated and scalable fraud detection using machine learning techniques for better financial security.

V. RESULTS

The results of the financial fraud detection model using machine learning are presented below, highlighting the performance metrics of the trained modeland itseffectivenessin classifyingtransactions as fraudulent or non-fraudulent.

Model	AUC	CA	F1	Precision	Recall	Support
LR	1.000	0.998	0.390	1.00	0.440	127252

TABLE1:EvaluationResult:PredictionbasedonLR

Model	AUC	CA	F1	Precision	Recall	Support
KNN	1.000	0.999	0.770	1.00	0.690	127086

TABLE2:EvaluationResult:PredictionbasedonKNN

Model	AUC	CA	F1	Precision	Recall	Support
DT	1.000	0.999	0.890	1.00	0.880	127086

TABLE3:EvaluationResult:PredictionbasedonDT

Model	AUC	CA	F1	Precision	Recall	Support
RF	1.000	0.999	0.870	1.00	0.780	127086

TABLE4:EvaluationResult:PredictionbasedonRF

Model	AUC	CA	F1	Precision	Recall	Support
XGB	1.000	0.997	0.91	1.00	0.86	127086

TABLE5:EvaluationResult:PredictionbasedonXGB

Model Performance:-

In conclusion, the overall accuracy of the model for the detection of financial fraud was 99.977%, and it hence confirmed the proper correct classification of the transactions. Precision and recall are strong, with values of 96%

and 86% respectively. Moreover, the F1scorest and sat 0.91, and hence, it really underlines that there is a balance between the precision andrecall of the model for fraud detection.

The AUC-ROC Curve value of 1.000 indicates that despite the presence of class imbalance, there is still significantabilityonthepartofthemodelto
distinguish between fraudulent and non-fraudulent transactions. The findings thus show that themachine learning model is strong enough to add weight to fraud detection systems, which can be relied upon by financial institutions to Improve significantly the identification and mitigation of fraudulent activitiesto protect their operations and customers,respectively.

VI. CONCLUSION& FUTUREWORK

In this research, we developed a very effective machine learning model that detects financial fraud using the XGBoost classifier. Our model presented remarkable performance, with an accuracy of 99.977%onthetestingdataset.Thislevelofaccuracy shows outstanding ability in classificationtransactions correctly; it distinguishes legitimate and fraudulent activities. Such a model showed a high precision of 96% and a recall of 86%. This is how much effectiveness it has in reducing both false positives and negatives. Missing any fraudulent transactions is costly in fraud detection.

The AUC score of 1.000, obtained by showing clear distinction between fraudulent and non-fraudulent transactions, proves that a model or classifier can easily differentiate between them, even when there is class imbalance in the dataset. These above results indicatehowadvancedmachinelearningmethodscan improve fraud detection systems, helping financial institutions keep their operations and customers safe from financial crime.

There are a number of ways for us to make our fraud detection model better for the future. Firstly, one important area to work on is to grow our dataset covering more kinds of fraud patterns. This would mean the inclusion of different types of fraud thatmaycomeupastechnologyandmethodschangewith the passage of time. We are therefore training our model on a wider variety of examples to make it stronger and better at handling new fraud techniques.

We also envision exploring ways to include actual real- time transaction monitoring features. Thismeans developing systems that can actually check transactions as they occur so we can easily spot and respond to suspicious activities right away. We can really reduce the chances for fraudsters to take advantage because of the quick action taken by our system.

We also intend to dig deeper into feature engineering tocomeupwiththemostimportant factorsthataffect fraud detection. Then, knowing which features make it easier to give an accurate prediction helps us to improve our model for better predictions.

Wealsohopetoexploreusingensemblemethodsand
deep learning techniques to add towards our existing model. Such techniques will create a better detection system due to improvement of accuracy inpredictions made by the model.

Finally, we want to build easy-to-use tools and dashboards that banks and financial institutions can use without difficulty, so we can apply our researchto real life. This will allow people to fight financial fraudquicklyasithappensifwemakethetechnology simple and helpful.

Future Projects: Improve the Security of Financial Systems. We envision projects which will extend security to financial systems to avoid fraud thataffects consumers and businesses.