Self Calibrating Automated Credit Card Fraud Detection System

Innovative Project Development

Submitted in Fulfillment of the Requirement of Subject MCA-216

To

Department of Computer Applications, National Institute of Technology, Kurukshetra

By: Group 21

Kshitij Pandey - 51910057

Piyush Sachan– 51910076

Shakti - 51910081

Under the Supervision of

Dr. Nikam Gitanjali Ganpatrao
(Assistant Professor, Department of Computer Applications,

National Institute of Technology, Kurukshetra)

DECLARATION

The work on the project, entitled "Self Calibrating Automated Credit Card Fraud Detection System," submitted to the National Institute of Technology, was the original work we did under the direction of Dr. Nikham Gitandjari Gampatlao, Assistant Professor. declare that this is a recording of computer applications, NIT Kurukshetra.

The project work will be presented in part with the requirements of the target MCA-216. The results included in the report have not been presented to other universities or research institutes for the award of diplomas or degrees.

Kshitij Pandey	**Piyush Sachan**	**Shakti**
519100057	**51910076**	**51910081**

This is to prove that the above statement of the candidate is true within my knowledge and beliefs.

Signature

Dr. Nikam Gitanjali Ganpatrao Assistant Professor

Department of Computer Applications NIT Kurukshetra

ACKNOWLEDGMENTS

The satisfaction and well-being associated with the success of a project are incomplete, not to mention those who made it possible.

First of all, I would like to thank my supervisor Dr. Nikam Gitanjali Ganpatrao for always motivating us and giving us this wonderful opportunity to show our creative ideas through the project. Many thanks to her for providing her with valuable guidance at all stages of the project. Without her advice, constructive suggestions, positive and supportive attitude, and continued promotion, this project would not have been possible.

We would also like to thank HOD for providing resources and infrastructure. Everyday tasks were challenging but the encouragement of the people around us throughout the journey of this project made it quite easy for us.

Last but not least we would also like to thank our coordinator, Dr. Sarika Jain, for providing us with the information regarding the project. Despite her busy schedule, she managed to resolve the conflict by providing us with proper guidelines about the project and summing up the process in time. We would also like to acknowledge each of our project partner’s efforts and each other for their wonderful collaboration.

ABSTRACT

Credit card plays a significant standard in the present wealth. It turns into a necessary piece of the family unit, business, and worldwide exercises. Although utilizing credit cards gives might profits when used carefully and dependably, huge credit and monetary effects might be imported by deceitful practices by fraudsters attributable to the ubiquity of electronic asset moves. Financial institutions try to enhance continuously their fraud detection systems, but fraudsters are at the same time hack into the systems with new techniques and tools. Such cheats cause a danger to the protection of humankind, bringing about monetary misfortunes. There is a requirement for planning progressed extortion discovery answers to limit the perils of these fakes. For the detection of deceits, many machine learning algorithms can be utilized. This note paper first discusses the statistics of credit card frauds in the world and primarily in India, then the type of frauds and gives a diagram to analyze the presentation of a few machine learning algorithms by doing a relative report that can be utilized for classifying transactions as misrepresentation or a real one. It also mentions the currently utilized best-in-class strategies to counter these attacks and highlights its limitations along with proposing a solution for it.

Keywords: Machine learning algorithms, Financial Institutions, Deceits, and Fraudsters.

LIST OF FIGURES

LIST OF TABLES

Chapter 1. INTRODUCTION

Nowadays usage of credit cards in developing countries has become an advent. Users use it for purchasing, take care of tabs, and online exchanges. It gives certain points of interest like the simplicity of procurement, keeps client records of loan repayment, the security of buys, and so forth. Be that as it may, the enormous scope utilization of Visas and the absence of viable security frameworks bring about billion-dollar misfortunes to Mastercard extortion. Since credit card organizations are commonly reluctant to declare such realities, it is hard to get an exact estimation of the misfortunes. Notwithstanding, certain information concerning the monetary misfortunes brought about with Visa misrepresentation is openly available. The utilization of Visas without solid security causes billion-dollar monetary misfortunes. In [17] worldwide monetary misfortunes because of Mastercard misrepresentation added up to 16,82,60,69,40,000.00 Indian Rupees in 2017 and are required to persistently increment by 2020, the number is relied upon to arrive at 22,87,75,50,50,000.00 Indian Rupees. In 2019, India has over 29 million credit card users. But Jamtara a well-known city in India has been the cybercrime hub for 5 years now since the advent of technology. 107 Jamtara residents were arrested on cybercrime charges in 2019.

Almost 15, 44,000 Rs. and 163 credit cards were seized from the accused. Illicit usage of a credit card along with its data without the information on the proprietor is alluded to as Mastercard extortion.

Some essential aspects are required for fraud detection solutions for an organization like prediction analysis, outlier model, global profiling, etc. Outlier models can be particularly useful for identifying frauds in developing business sectors where adequate information to make forecasts doesn't yet exist. For example, if transactions are related to purchasing only but if any transaction related to selling takes place, an outlier model would flag that card. In predictive analytics, these happen to predict models dependent on billions of payment exchange of credit card and purchaser profiles. Different countries are dealing with different and new types of fraud due to the diversity of transactions taking place through credit cards. For finding a solution they are working upon real and synthetic datasets and creating algorithms to solve this issue. Unavailability of datasets due to security reasons is a big issue for them they are facing right now. The information on the planet is insignificant without human understanding. Many organizations are nowadays hiring fraud analysts for human intervention and explanation of technical terms and providing better solutions.

In [2] two types of frauds are Application and Behavioral Fraud are discussed. Application fraud takes place when scammers apply for novel cards from banks or giving organizations utilizing bogus or other data. Numerous applications might be presented by one client with one bunch of client subtleties (called duplication fraud) or an alternate client with indistinguishable subtleties (called identity fraud). Behavioral fraud, then again, has four chief sorts: Taken/Lost card, Mail burglary, Fake Card, and the Cardholder Not Present misrepresentation. Taken/Lost Card Fraud happens when fraudsters take a Mastercard or gain admittance to a lost card. Mail Burglary Fraud happens when the fraudster gets a Visa via the post office or individual data from the bank before arriving at the real cardholder. In both Counterfeit and Cardholder Not Present fakes, Mastercard subtleties are gotten without the information on cardholders. In the previous, distant exchanges can be led utilizing card subtleties through the mail, telephone, or the web. Counterfeit Cards are inferior to card data. Fraudulent exchanges can be recognized either by grouping approach or by identifying distant exchanges from typical exchanges. For the grouping approach, the principal model is prepared for preparing information. Features are extracted and changed from crude information while offering it to train the model.

1.1 PROBLEM STATEMENT

To build a Hybrid model using Random Forest, K-Nearest Neighbor, and Artificial Neural Network algorithm that can detect and prevent precisely and accurately fraudulent transactions while they are in transit by Majority Voting approach.

1.2 ORGANISATION OF REPORT

The report starts with the motivation behind Fraud Detection in Chapter 1. It further introduces the Problem statement.

This Chapter is followed by Chapter-2 Background. In this, background research is done on the three techniques which we have used. All that is there to be learned and study about these techniques is written in this chapter.

In Chapter-3 Literature Review, we go on further to throw light on the widely spread literature on similar problems after extensive reading and learning sessions from available sources and saturating them to provide a wider perspective on similar problems.

Chapter-4 Project Objective provides an understanding of the hurdles which are faced while developing our system and the gaps present in the already developed solutions. It throws light on what our system aims to achieve. It also mentions all the tools, techniques, methodologies, and datasets that can be used.

Chapter-5 Proposed Solution relates the techniques that are used in our proposed system to achieve the aforementioned goals and filling the gaps identified. It provides a better understanding of the working of our model by depicting it in the form of a flowchart and providing its architecture. It also describes the algorithm for the working of the three techniques.

In Chapter-6 Experimental Results, we have shown and discussed the various results graphs produced by our system and have shown the outputs generated on various inputs in the form offer a better understanding.

In Chapter-7 Conclusion and Future Scope, we provide an understanding of the overall conclusion of the proposed solution i.e. the combination of the 3 techniques, namely Random Forest, K-NN, and ANN, and have discussed the future scope in this problem.

1.3 CONTRIBUTION OF THE TEAM MEMBERS

- **Kshitij Pandey:** Implemented the code for the Random Forest algorithm, collected the required dataset, explored it and preprocessed it, and worked on the report.
- **Piyush Sachan:** Implemented the code for K- Nearest Neighbors algorithm and also worked on the report.
- **Shakti:** Implemented the code for ANN(Artificial Neural Network), executed the Majority Voting approach, and worked on the report.

Chapter 2. BACKGROUND

2.1 RANDOM FOREST ALGORITHM(RF)

RF is a standard ML technique that has a spot with the supervised learning technique. It might be utilized for both Classification and Regression issues in ML. It confides in the possibility of ensemble learning, which is an association of uniting various classifiers to deal with an intricate dispatch and to recover the presentation of the model.

As the name proposes, "RF is a classifier that contains various decision trees on different subsets of the given dataset and takes the normal to improve the prescient exactness of that dataset." Instead of relying upon one decision tree, the RF takes the expectation from each tree and ward on the majority votes of conjectures, and it predicts the last yield.

The more evident total trees in the forest provoke huge accuracy and anticipate the dispatch of overfitting.

2.1.1 Application of Random Forest

- Banking: The banking area generally utilizes Random Forest for distinguishing proof of advance danger.
- Medicine: With the assistance of Random Forest, sickness patterns and dangers of the illness can be distinguished.
- Land Use: We can distinguish the territories of comparative land use by Random Forest.
- Marketing: Marketing patterns can be distinguished utilizing Random Forest.

2.1.2 Benefits of Random Forest

- RF is fit for performing both Classification and Regression errands.
- It is fit for taking care of huge datasets with high dimensionality.
- It upgrades the exactness of the model and forestalls the overfitting issue.

2.1.3 Why utilizing Random Forest

- Taking less time when contrasted with other algorithms.

- It keeps up accuracy when a huge segment of information is absent.
- It predicts output for little or enormous datasets proficiently.

2.2 K-NEAREST NEIGHBOR ALGORITHM(KNN)

KNN algorithm is a sort of supervised Machine Learning algorithm which can be utilized for both classifications just as regression prescient issues. Notwithstanding, it is predominantly utilized for the prediction of classification problem issues in the organization. The accompanying 2 resources would characterize KNN well −

- Lazy learning: This is the K-nearest neighbors algorithm since it does not have a particular preparing stage and in classification it using all the data for training.
- Non-parametric learning: K-nearest neighbors is furthermore a non-parametric learning technique since it doesn't acknowledge whatever regarding the essential information.

2.2.1 Pros

- KNN is an extremely straightforward algorithm to comprehend and decipher.
- KNN is valuable for nonlinear data because there is no presumption about data in KNN.
- KNN is an adaptable algorithm as we can utilize it for regression just as classification.
- KNN has generally huge accuracy yet there are vastly improved Supervised learning models, than K-nearest neighbors.

2.2.2 Cons

- It is an estimated somewhat costly technique since it reserves all the training data.
- High memory cache needed when conflicts with other supervised learning algorithms.
- If there is an occurrence of large N when delayed in prediction
- KNN is touchy to the amount of data just as immaterial features.

2.2.3 Uses of K-Nearest Neighbors

Coming up next are a portion of the zones in which K-nearest neighbors can be enforced effectively −

- Bank Sector: K-nearest neighbors can be utilized in a financial framework to anticipate climate a particular is good for credit endorsement? Does that particular have the attributes like the defaulters one?
- Ascertaining Credit Ratings: K-nearest neighbors can be utilized to track down a people's credit record by contrasting and the people having comparative qualities.
- Governmental issues: K-nearest neighbors assists that we can arrange a possible citizen into different classes like "Will Vote", "Won't Vote".
- Different regions where K-nearest neighbors can be utilized are Speech Recognition, Handwriting Detection, Image Recognition, and Video Recognition.

2.3 ARTIFICIAL NEURAL NETWORK (ANN)

An information processing technique is known as ANN. It works like how the human cerebrum measures data. Artificial Neural Network joins endless related planning units that collaborate to deal with data. They moreover make huge outcomes from it.

We can affix Neural Network not only for classification. It can similarly affix for relapse of reliable target attributes. Neural networks find remarkable applications in data mining utilized in areas.

A neural organization may accommodate the going with 3 layers:

- Input layer: The enterprise of the info units tends to the rough data that can deal with in the organization.
- Hidden layer: To choose the enterprise of each hidden unit. The activities of the input units and the heaps on the relationship between the covered up and the input units. There may be at any rate one hidden layer.
- Output layer – The direction of the output units confides in the enterprise of the hidden units and the heaps between the output and the hidden units. At every neuron, each input has a connected weight which changes the strength of each input. The neuron incorporates all of the input sources and figures an output to be passed on. The output is settled through non-linear activation work. The

activation function is normally a determined limit that changes the output to a number that is between 0 &s1. There can be other activation functions, which may be significant are inspected in the accompanying sub-territory. Neural processing requires different neurons, to be related together into a "neural network". Neurons are planned in layers. Artificial Neural Network comes in numerous constructions like Recurrent NN, Associative NN, etc multi-layer-feed-forward neural organization. "Figure 1" portrays a basic multi-layer feed-forward neural network.

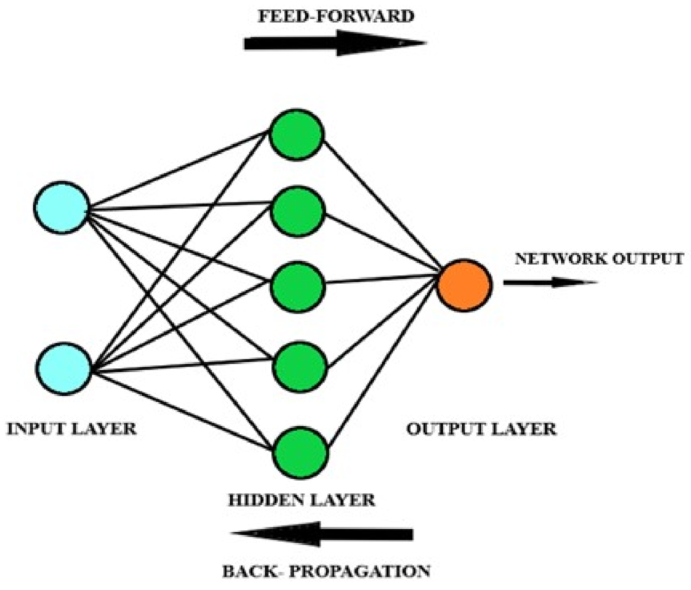


Figure 1. Multi-layer feed-forward neural network

It comprises a Hidden Layer, Input layer, and output layer, hidden layer relies upon the difficulty we will address, it very well may be no or more than one hidden layer. The quantity of neurons in input layer compares to the total input attributes in the training dataset and the total neurons in the output layer is rely upon the sort of issue you will tackle, in credit card fraud detection case we have two output one is fraudulent and the other one is legitimate i.e., 0 and 1 separately.

2.3.1 Feed-forward neural network

In this, as we can find in "fig. 2", there is no input circle. Every one of the neurons in each layer is associated with one another without making any circle and the connection between these neurons has loaded. The association between every neuron doesn't play out any count yet is utilized to store the loads. These loads are instated for certain irregular qualities and changes at each cycle in the training process. The straightforward neuron in each layer are frequently called perceptron is the easiest neuron network.

2.3.2 Backpropagation

Because of the worth that the hub has terminated, we acquire the last output. At that point, utilizing the blunder capacities, we figure the disparities between the anticipated output and coming about output and change loads of the neural network through an interaction known as backpropagation.

2.3.3 Choosing a right activation function

Since we have seen so numerous activation functions, we need some rationale/heuristics to know which activation functions, need to be utilized in which prospects. Fortunate or unfortunate – there is no general guideline.

- Anyway, commit to the properties of the difficulties we could determine on a superior decision for the quiet and disagreeable intermingling of the network.
- Sigmoid capacities and their blends for the most part work superior on account of classifiers.
- Sigmoids and tan h quantities are in some cases kept away from because of the vanishing gradient problem.
- ReLU work is an overall actuation work and is utilized as a rule nowadays.
- If we know-how an instance of dead neurons in our organizations the flawed ReLU work is the most ideal decision.
- Always recognize that ReLU space should just be utilized in the hidden layers.
- As a generic guideline, you can open with utilizing ReLU space and after move over to other activation functions on the off chance that ReLU does not furnish with ideal outcomes.

A feed-forward perceptron works by sending the contribution to the neurons and ship off the output neuron after preparing.

All the perceptrons in the Neural Network have two capacities for example Input and

Activation Function. As the name recommends, the input function work gathers all the input and performs summation work on the input, and afterward moves the outcome to the actuation function. An input function play out some procedure on the outcome after summation and afterward move to the following level

The consequence of this summation work is then passed to the activation function. It scales the estimation of S's inappropriate reach. Basic activation functions are sigmoid actuation work which chips away at threshold value, if the estimation of S surpassed the threshold value, the hub pass output. There is two actuation function which is usually utilized in Neural Networks, Sigmoid and Hyperbolic Tangent Activation Function. It relies upon the training dataset on which we will train the network that which actuation function is acceptable.

2.3.4 Threshold function

This function outputs a value of 0 when the total of the weights of inputs is less than a certain threshold. It outputs a value of 1 in the case where the weighted sum of inputs is equal to or greater than a certain threshold value.

$$f(x) = \{0, \sum w_i < \Theta 1, \sum w_i \geq \Theta\}$$

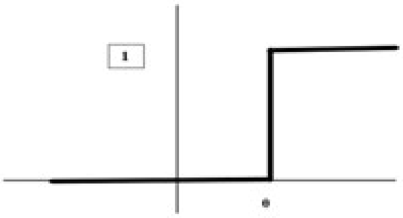


Figure 2. Threshold Function

2.3.5 Sigmoid function

The following activation function that we will take a gander at is the Sigmoid capacity. It is quite possibly the most generally utilized non-linear activation function. Sigmoid changes the qualities between the reach 0 and 1. Here is the numerical articulation for sigmoid-

$$f(x) = 1/(1+e^{-x})$$

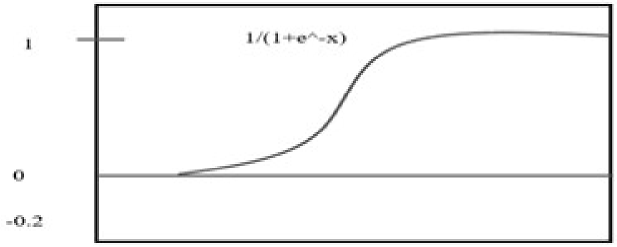


Figure 3. Sigmoid function

2.3.6 Hyperbolic Tangent Activation function

The tan h work is fundamentally the same as the sigmoid capacity. The lone distinction is that it is symmetric around the root. The scope of qualities for this situation is from - 1 to 1. In this manner, the inputs or contributions to the following layers won't generally be of a similar sign. The tan h work is characterized as-

$$\text{Tan h}(x) = (2*\text{sigmoid}(2x))-1$$

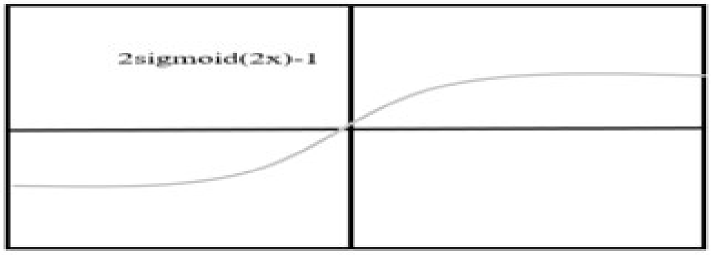


Figure 4. Hyperbolic Tangent Activation function

2.3.7 ReLU function

The ReLU work is another non-linear activation function that has acquired fame in the deep learning area. ReLU represents Rectified Linear Unit. The principal benefit of utilizing the ReLU work over other activation functions is that it doesn't actuate every one of the neurons simultaneously. This implies that the neurons may be deactivated if the output of the direct change is under 0. The plot underneath will assist you with understanding this better-

$$f(x) = \max(0,x)$$

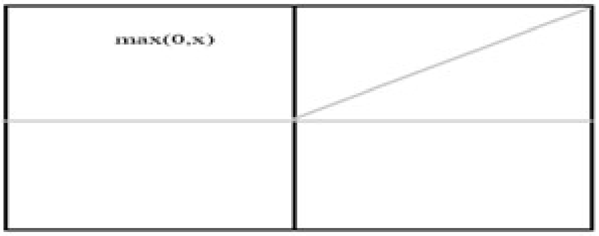


Figure 5. ReLU function

2.3.8 Specifics for Training a Neural Network

Step 1. Prefer a Network Architecture (for example Availability Pattern between the neurons).

Step 2. Irregular Initialization of the Weights

Step 3. Execute Forward Propagation to gain the beginning forecast for any x(i).

Step 4. Execute the Cost Function for the ANN.

Step 5. Execute back-propagation to process halfway subsidiaries

Step 6. Use gradient checking to affirm that your back-propagation works. At that point debilitate gradient checking.

2.3.9 Pros

- Having adaptation to fault-tolerant: Corruption of at least one cell of ANN does not keep it from generating output. This component assembles the network's fault-tolerant.
- Gradual defilement: An industry eases back after some time and goes through relative deprivation. The organization issue doesn't promptly consume right away.
- Ability to make ML: ANN learns possibility and settles on extract by assertion on comparative occasions.

2.3.10 Cons

- Hardware reliance: ANN requires processors with equal preparing potential, as per their design. Therefore, the acknowledgment of the hardware is reliant.
- Unexplained conduct of the network: This is the main concern of Artificial Neural Networks. At the point when Artificial Neural Network outcome an examining

layout, it doesn't provide some vision with regards to why and how. This diminishes expectations in the network.

- Determination of non-fraudulent organization system: There is no fact guideline for deciding the construction of fake neural networks. The fitting network system is proficient through experience and observations.

Chapter 3. LITERATURE REVIEW

3.1 RELATED WORK

A decent comprehension of fraud discovery advancements can help recognize credit card extortion.

The work in [1] formulated a Multi-Classifier approach and our outcomes show that the methodology can essentially decrease misfortune because of ill-conceived exchanges. One impediment of our methodology is the need of running fundamental trials to decide the ideal dissemination dependent on a characterized cost algorithm. This cycle can be mechanized yet it is inevitable since the ideal dispersion is profoundly reliant on the cost model and the learning calculation.

In [2] we look to survey the condition of craftsmanship in Visa extortion location strategies, misrepresentations, and a few credits of client profile or conduct. We accept that the commitment of qualities is a basic factor that should be thought of. Additionally, an appropriate dataset should have the option to mirror this present reality of credit cards.

We proposed a framework in [5] that assists with identifying extortion in charge card exchanges utilizing a Choice Tree with a mix of Luhn's calculations. It is utilized to approve card digit numbers and we utilized the location coordinating guideline to check if the charging and transportation address coordinated.

The work presented in [7] executes Naïve Bayes and K Nearest Neighbor to recognize charge card extortion additionally examines the dataset. A sort of extortion identification strategy is utilized in business banks to distinguish misrepresentation by checking the conduct of various cardholders. Information mining assumed a significant function in the discovery of Mastercard misrepresentation in online exchanges. Data analysis and data cleaning is an extremely difficult task to take care of. The algorithm sets aside a ton of effort to emphasize the enormous dataset. In any case, fraud detection of credit cards just utilizing one specific calculation won't be a proficient practice. As each algorithm contain their focal points and drawbacks and to do credit card fraud detection viably, we should join algorithms with the goal that we would have the option to exploit every algorithm and complete our work adequately.

The work presented in [8] checks the exhibition of various calculations like Choice Trees, Strategic relapses, Irregular Backwoods, and SVM classifiers on profoundly slanted information. Results show the precision of the Choice Tree, Strategic Relapse, Arbitrary Woodland, and SVM classifiers. Banks have faced some troubles in distinguishing misrepresentation in Visas. The Random Forest will work better in the case of large training data. Its speed will suffer during the testing phase. SVM needed more pre-execution to provide better outputs. It deals with the data imbalance issue.

In [9] distinctive AI calculations for banks in identifying cheats in charge cards were utilized. The exhibition of extortion recognition in Visa exchanges is enormously influenced by the testing approach on the informational collection, choice of factors, and location methods utilized. In this, we adjusted the dataset by doing oversampling because the dataset is extremely imbalanced.

The work in [11] initially presents the best information mining calculation called "AI calculation" which is utilized to distinguish Mastercard misrepresentation. We additionally resolve an issue for true information of exploration on Mastercard misrepresentation that isn't effectively accessible by utilizing "Ada Boost and greater part vote technique". By utilizing this technique dataset is freely accessible.

[14] In this, the fundamental commitment of work is the improvement of an extortion location framework that utilizes a profound learning design along with a serious component designing cycle dependent on homogeneity-situated conduct examination. We direct a near report on the China Bank dataset to evaluate the viability of the calculated model. There are still a few constraints to our work. For instance, we didn't evaluate the estimation cost of our proposed HOBA highlight designing system, which develops a lot bigger variable set than RFM. In this way, in future work, we need to complete further exploration from two viewpoints. The first one spotlight on investigating the estimation interest of a continuous fraud detection framework and the other one is to investigate the use of further developed AI strategies and potential blends of profound learning techniques and conventional data mining strategies in fraud detection.

[17] This work proposes an advanced light angle boosting machine (O Light GBM) way to deal with identifying charge card extortion. In the offered solution a Bayesian-based hyper-parameter calculation is cleverly coordinated to improve the boundaries of a Light GBM. It utilizes two true open exchanges dataset comprising deceitful exchanges and

authentic ones to exhibit the adequacy of our proposed O Light GBM for recognizing Visa misrepresentation exchanges. Credit card fraud location has ended up being a test mostly because of 2 issues that it presents both the profiles of ordinary and false practices change and datasets are exceptionally slanted. The exhibition of extortion location is influenced by the factors utilized and the strategies we used to identify misrepresentation. Credit card fraud location has ended up being a test mostly because of 2 issues that it presents both the profiles of ordinary and false practices change and datasets are exceptionally slanted. The exhibition of extortion location is influenced by the factors utilized and the strategies we used to identify misrepresentation.

[18] Compares the performance of LR, K-NN, RF, NB, Pipelining, Ensemble learning, Multilayer Perceptron, and Ada Boost on the card's fraud data and analyzed them and found that Pipelining was the best among them. Dataset is quite imbalanced. Sampling techniques can be used to solve this issue.

In [21] an average accuracy-based model has been built by combining three different algorithms namely KNN, ANN, and Decision Trees which works quite better for all kinds and sizes of datasets. The model provides substantial improvement in accuracy thus reduce the disadvantages of each existing algorithm. It gave an average accuracy of 95.66%.

In [22] the sequential model is used to detect fraud. This means that the model is trained with the expected results of the CNN model and is entered into the KNN training data. Therefore, the cumulative error rate of both primitive models reduces the probability and builds the degree of accuracy. We used CNN as the main phase of the hybrid model. After training on 490 data cycles, the specific accuracy is 87.79% and the logarithmic deficit is 3.90. 90.5% KNN classification is taught. Inefficient hybridization, the accuracy of the obtained model is 98% and the loss of logs is 0.647.

In [24] Lots of biased credit card fraud for Bayes' naive presentations, logistic regression, and K-close neighbors. The MasterCard exchange data set is from a European cardholder and supports 284,807 exchanges. The hybrid approach to oversampling and undersampling ends with distorted data. The three steps refer to pre-processed raw data. The work is done in Python. The performance of the procedure is evaluated according to the Matthews correlation coefficient, sensitivity, accuracy, precision, specificity, and balanced degree of classification. The results with ideal accuracy for logistic regression, KNN, and naive Bayesian classifiers are 54.86%, 97.69%, and 97.92% respectively separately. The similar outcomes derive that K-Nearest Neighbor performs better compared to Logistic Regression and Naïve Bayes strategies.

Table 1. Summary of studies (comparison) of existing credit card fraud detection technologies

Sr./Ref. No	Objective	Methodology	Achievement	Limitations
[1]	Side-side programming work Scalable learning	Multi-Classifier approach	Scalable to larger amounts of data. Non-uniform cost per error.	Running fundamental experiments to know the distribution dependent on a cost model.
[2]	Comparison of different Machine learning algorithms	ANN Genetic Algorithm SVM	ANN: Ability to learn from the past. GA: Fast in detection. SVM: It can be robust even when the training sample has some bias.	ANN: High Processing time for the larger neural network. Genetic Algorithm: High computational cost. SVM: Processing large datasets is a tedious job for it.
[5]	Card number	Luhn's	Validation of the	Larger

	validation	algorithm Bayes theorem	card is a genuine and very low false alarm.	processing time.
[7]	Enhanced accuracy and flexibility.	Naive Bayes model KNN classifier	Naive Bayes: 95% precision KNN: 90% precision	It takes an excruciating time to loop over the big dataset.
[8]	Collating ML models and compares the performances.	Logistic Regression SVM Decision Tree Random forest	RF: 98.6% (accuracy).	SVM deals with data imbalance and needed more pre-execution. In RF speed during testing and application will suffer.
[9]	Comparing these three algorithms.	Logistic Regression Decision Tree Random Forest	Random Forest classifier is the best algorithm with an accuracy of 95.5%.	Imbalance dataset Feature selection problem
[10]	To study the Random Forest.	Random Forest	RF obtains a good result in a small set of data.	Imbalance data and RF itself should be improved.
[13]	Compare the performances of MLP(Multilayer	RF LR	Different ML algorithms can give reasonable	Computational cost is high in a

	Perceptron), NB, RF, and LR.	MLP NB	results with appropriate pre-processing. SMOTE improves random oversampling. RF was found out to the best.	neural network. Skewed distribution of data.
[18]	Using a novel approach to detect fraud.	Pipelining, Ensemble learning	Pipelining works best as compared to another algorithm.	Dataset is highly imbalanced.
[21]	Creating a model that works quite better for all kinds and sizes of the dataset.	Hybrid approach: KNN, Decision Trees, and ANN	Average Accuracy: 95.66%	No optimization technique was used to improve the speed of ANN.
[22]	Comparative study of the outcomes of Convoluted Neural Network and K-Nearest Neighbors and the hybrid model of both.	Serialized approach: KNN and CNN	KNN: 90.66% Accuracy CNN: 88.12% Accuracy Hybrid: 98% Accuracy Upon hybridization,	Will work better if trained over a large balanced dataset.

			CNN’s accuracy rate increased by 10%.	
[23]	Solving the credit card data imbalance problem.	A SMOTE-based oversampling data-point approach with SVM, LR, DT, and RF.	The ability to recognize positive classes improved. Random Forest and Decision Trees produced the best performance.	After oversampling, the ability to recognize negative classes degraded especially in SVM.
[24]	Comparative performance of Logistic Regression, Naïve Bayes, K-Nearest Neighbor models in two set distribution of imbalanced credit card fraud data.	NB, KNN, and LR	Naïve Bayes: 97.92% Accuracy K-Nearest Neighbors: 97.69% Accuracy LR: 54.86% Accuracy	KNN shows significant performance for all metrics evaluated except for accuracy in the 10:90 data distribution. LR accuracy decreased rapidly in the 34:66 and 10:90 distribution.
[25]	To detect fraud in real-time and	Hybrid approach: SOM	In this model, we achieved	Do not give the same result when

	efficiently.	and ANN	better accuracy precision and cost compared to using SOM or ANN alone.	applied to a different dataset.
[26]	Examining the variety of models can be utilized to find several credit card frauds with changeable degrees of accuracy.	ANN, SVM, Genetic Algorithm, and RF.	Artificial Neural Network obtained a high-performance rate. A Genetic Algorithm detects the fittest solution.	Simulated Annealing is a time taking process.

Chapter 4. PROJECT OBJECTIVES

4.1 GAPS IDENTIFIED

We have found out that the dataset is highly imbalanced. It contains more genuine transactions as compared to fraudulent ones. Dataset is not easily available. Most of the attributes are already transformed by the banks due to security reasons. Supervised ML algorithms only detect old fraud patterns but it is slow and accurate in the case of historical data whereas Unsupervised ML algorithms only detect novel fraud patterns. It is fast but requires an extensively large dataset for learning in real-time. No model is perfect or generally used to detect fraud. Each one has its advantages and disadvantages. Table I shows the machine learning strategies that can be utilized to achieve the detection of credit card fraud. It is fundamental to make a point that the basic variation of this table is presented in the given references. To best remember comparisons in credit card fraud detection we have been summarizing the objective, methodology, achievement, and limitation of different techniques utilized for credit card fraud detection.

In [1] Chase Manhattan Bank data set is utilized that consists of half a million transactions from 10/95 to 9/96, about 20% of which are fraudulent.

In [10] China's commercial organization dataset was used. Its consists of transactions from November 2016 to January 2017. It has 62 features and 30,000,000 no. of transactions. Total transactions of 82,000 are named as fraud with a proportion of 0.27% and the dataset just imbalance issue ought to be contemplated.

In [7, 8, 9, 13, 18, 21, 23, 24] the dataset being utilized here is from the origin Kaggle. The dataframe comprises credit card transactions formed by European cardholders in September 2013. This dataframe displays transactions that arise in 2 days, where we have a total of 492 frauds out of 284,807 transactions. The dataset is broadly imbalanced, from the total transactions only 0.172% is of positive class (fraud). It comprises only mathematical input variables which are the outcome of a PCA transformation. Tragically, because of confidentiality matters, we can't provide the original features and more background information about the data. Features V1, V2 … V28 are the principal components gather with PCA, some features which have not been transformed with PCA are 'Amount' and 'Time'. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

4.2 OBJECTIVES

- To built a Hybrid model consisting of three different algorithms ANN(Artificial Neural Network), RF (Random Forest), KNN(K-Nearest Neighbors), and with a Majority Voting approach.
- Improve the recall, f-score, precision, and AROC curve of the Hybrid model.
- For a particular transaction, to get the opinion of 3 algorithms regarding whether is legitimate or fraudulent by implementing the Majority Voting approach.
- Do more exploration of the dataset to understand the problem deeply.
- Individually apply each algorithm on the test set to see how they perform concerning the Hybrid model.
- Building a model that works adequately well for wholly kinds and sizes of the dataset.

Chapter 5. THE PROPOSED SOLUTION

5.1 TOOLS AND TECHNIQUES

5.1.1 Techniques

5.1.1.1 For Training

- Random Forest (RF)
- Artificial Neural Network (ANN)
- K-Nearest Neighbor (KNN)

5.1.1.2 For Testing

- Random Forest (RF)
- Artificial Neural Network (ANN)
- K-Nearest Neighbor (KNN)
- Hybrid Model(RF, K-NN and ANN)

5.1.2 Methodologies

5.1.2.1 ANN Supervised Learning Laws

- Relu function: To instantiate weights.
- Sigmoidal function: To Converge ANN.
- Delta Rule: To calculate the LMS value between calculated output and desired output.
- Multi-layer feed-forward backpropagation algorithm: Algorithm used in neural network
- Random Forest
- KNN

5.1.3 Languages and Libraries

- Python: It is an interpreted, high-level, general-purpose programming language. It works in many ways like object-oriented way, procedural way, or functional way. It is developed by Guido Van Rossum. Python grows day by day. It is mainly used for machine learning but it is also used for web development, Desktop applications, and many more because it supports many machine learning libraries. Python is utilized to hold big data and perform complicated mathematics.
- Pandas library: It is a python library that stands for 'Data Analysis tool'. It is easy to use and high-performance tool. It takes the data file in (like CSV files, like the file or SQL Database) then it creates the dataset with rows and columns that is looking very similar to the EXCEL file.
- Numpy library: It stands for 'numeric python'. It is a python library that is used for working with arrays. It has a large collection of a high mathematical functions that is helpful for multidimensional arrays and matrices. Python uses lists that are similar to arrays but NumPy provides array object that is faster than python lists. In NumPy array object is known as the array.
- Matplotlib library: It is a visualization library in python that works with NumPy array and scipy functions to plot various types of graphs.
- Keras library: It is an open-source neural network library written in python. They can work on the peak of tensor flow which in turn is a great library used in the production of deep learning models. However, the latter is complex to use while Keras is easy to work with and serves to be a great high-level neural network API.
    - Keras.model: Keras model could be built to be either Sequential or functional.
        - Sequential: This allows us to create the model layer by layer. It has a limitation that it could not deal with in cases of multiple outputs, sharing of layers.

    - Keras.layers:
        - Dense: Performs activity: Output = Activation (point (input, bit) + predisposition), where triggering is an activating function for each component passed as an activation argument, the kernel is created from layers The loading matrix is and the slope is the vector for

predisposition created by the layer (which can be used if use_bias is True).

- Scikit learns library: Scikit-learn is a free programming AI library for the Python programming language. It gives different various classification, regression, and clustering techniques.
    - sklearn.preprocessing: This package provides a mean to change raw feature vectors into a representation which suits the downstream estimators in a better way. It accommodates certain probable utility functions and generator classes to achieve a representation that is more suitable for downstream estimators. The standardization of the dataset provides a lot of benefits to learning algorithms like a neural network.
        - StandardScaler: This function transforms our dataset in such a way that it has a mean value of zero and one value for standard deviation. This is what is known as the standardization/normalization of data frames.
    - sklearn.model_selection
        - train_test_split: this function is used to split our dataset into training data and test data.
- Seaborn library: Seaborn is a Python library that is used for creating measurable designs in Python. It expands on top of matplotlib and coordinates intimately with pandas' data structures. It assists you with investigating and comprehend your data. Its plotting functions work on data edges and exhibits consist of entire datasets and inside play out the vital semantic mapping and analytical gathering to deliver educational plots. Its dataset-oriented, demonstrative API allows you to zero in on what the various components of your plots mean, as opposed to on the subtleties of how to draw them.
- Tensorflow library: It is a free library that is used for mathematical computation and immense scope ML. TensorFlow packages together a vast number of ML and deep learning (on the other hand called neural networking) models and algorithms and generate them profitable via a common metaphor It utilizes Python to give a beneficial front-end API for creating applications with the framework while

executing those applications in admirable C++.

- Imblearn library: Imbalanced-learn is an open-source, MIT-authorized library depending on scikit-learn and furnishes tools when managing classification with imbalanced classes.

5.1.4 Platform

Google Collaboratory: A Collaboratory is a short form of Colab, which is a thing from Google Research. Colab licenses anybody to form and execute self-self-assured python code through the program and is especially fitting to ML, schooling, and information investigation. Even more really, Colab is a cloud-based encouraged Jupyter notebook organization that requires no game plan to use, while giving free permission to figuring resources including CPUs, GPUs, and TPUs.

5.1.5 Dataset

Dataset is taken from kaggle.com. This data set has been used in 15 of the publications mentioned in Chapter – 3. It contains exchanges made by Mastercards in September 2013 by European cardholders for more than 2 days, where we have 492 cheats out of 284,807 exchanges. The dataset is exceptionally uneven, the positive class (fakes) represents 0.172%, everything being equal. It has altogether 31 highlights out of which 28 are just mathematical input factors that are the consequence of a PCA (Principal Component Analysis) change. These 28 highlights correspond to attributes of a customer like a name, age, occupation, location, account balance, type of card, etc. For security purposes, they have been PCA transformed. The other three features are time, amount, and class.

5.2 FLOWCHART

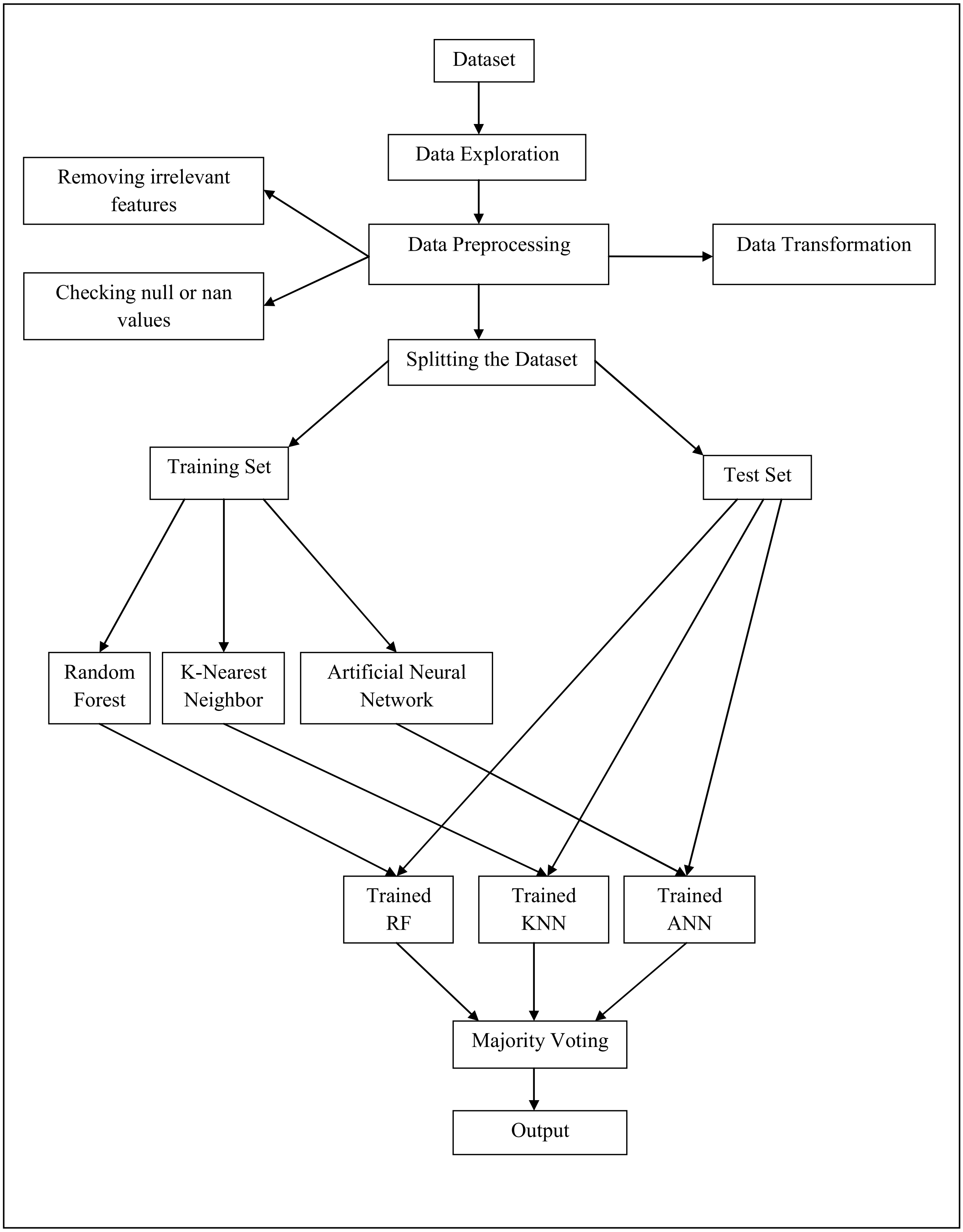


Figure 6. Flowchart Description

5.3 LIBRARY FUNCTIONS USED IN IMPLEMENTATION

- dataset.shape: It gives all the rows and all the columns of the dataset.
- dataset.columns: for getting the number of columns.
- dataset.describe: It describes all the transaction details of both fraud and non-fraud.
- head(): This function is used to return rows from the dataset. By default it returns 5 rows but if you want any number of rows only pass the value as a parameter.
- drop()**:** This function is used to remove rows and columns that have values either null or nan values.
- Iloc**:** iloc in pandas is utilized to pick rows and columns by number, in the enact that they emerge in the dataset.
- info()**:** It displays the summary of the dataset.
- classifier.fit(): fit(X_train, Y_train) sort out how the classifier direct on the training set by saving track of what amount of time it requires to prepare the classifier with the time module, save the training model, the grade, the training score and the training time into a word reference.
- fit_transform(): It transforms the value of that column that is not in the range of other columns. It simply returns values of those columns in the same range.
- classifier.evaluate(): This function is used to figure out the conduct of our trained model against the test data and returns a scalar consisting of the loss value and the value of the metric.

5.4 ALGORITHMS

- Artificial Neural Network (ANN)
- K-Nearest Neighbor (KNN)
- Random Forest (RF)

5.4.1 Artificial Neural Network(ANN)

Step1: Randomly instate the weights to little numbers near 0.

Step2: Input the main perception of the dataset in the input layer, each element in one input node (neuron).(No of units=29 is being utilized).

Step3: One hidden layer is being used with 29 neurons.

Step4: Forward-Propagations from left to right. Spread the activations until getting the anticipated outcome y.

Step5: Compare the anticipated outcome to the real outcome. Measure the created mistake.

Step6: Back-Propagation from option to left, the blunder is back-propagated.

Step7: Update the weights as per the amount they are liable for mistake.

Step8: Reiterate stages 1 to 5 and update the weights after every perception.

Step9: When the entire training set went through ANN, that makes an epoch. Retry more epochs.

5.4.2 K-Nearest Neighbors(KNN)

Step-1: Select the number K of the neighbours. Here K=5 being used.

Step-2: For K number of neighbours evaluate the Euclidean distance(p=2).

Step-3: Take the K closest neighbors according to the determined Euclidean distance.

Step-4: Among these k neighbors, tally the quantity of the daily points in every class.

Step-5: Assign the new daily points to that class for which the quantity of the neighbor is greatest.

Step-6: Our model is prepared.

5.4.3 Random Forest Algorithm

```
Step-1: Select  from the training set irregular K data points.
Step-2: Construct the decision tree related with the choice data points (Subsets).
Step-3: For decision trees that you need to construct take the number N . Here N=10 is being utilized.
Step-4: Reiterate Step 1 & 2.
Step-5: For new data points. discover the forecasts of every decision tree, and allocate the new data points to the class that successes the dominant part casts a majority votes.
```

5.5 EVALUATION METRICS TO BE USED

The dataset is imbalanced and the number of the fraudulent class label is less than the number of legitimate class labels.

1) Confusion Matrix: It is a N x N framework applied for determining the exhibition of a characterization model, where N is the number of target classes. The grid diverse the genuine objective aspects and those await by the AI model. This gives us an all-encompassing perspective on how well our order model is performing and what sorts of blunders it is making. It comprises 4 parts-

True Positive (TP): A TP is an outcome where the model predicts accurately the positive class.

True Negative (TN): A TN is an outcome where the model predicts accurately the negative class.

False Positive (FP)**:** A FP is an outcome where the model predicts inaccurately the positive class.

False Negative (FN): A FN is an outcome where the model predicts inaccurately the negative class.

2) Precision: Precision is the fraction of true positives among all of the positive classes.

$$\text{Precision} = TP / (TP + FP)$$

3) Recall: Recall is the fractions that were predicted to belong to a class concerning all of

the instances that truly belong to the class.

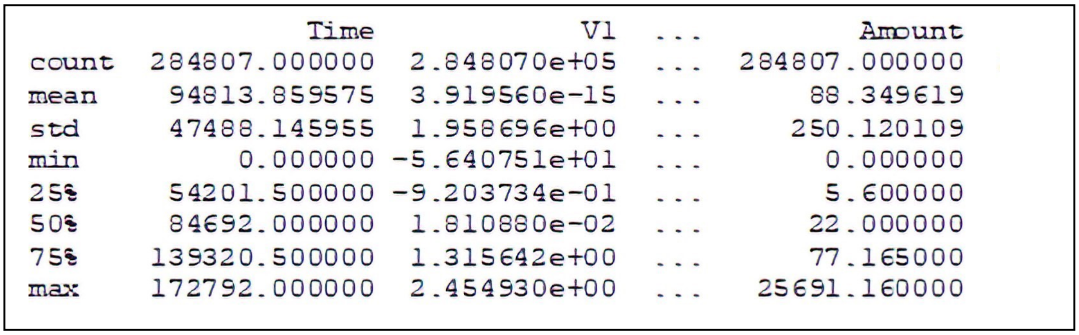
Recall= TP / (TP +FN)

4) F1-score: 2*Precision *Recall / (Precision + Recall)

5) Area under ROC curve*:* It is better for skewed data. It is used to test model performance and also to check the model's capability to differentiate between target classes. These models have an AUC value near to 1is good and value near to 0 is worst.

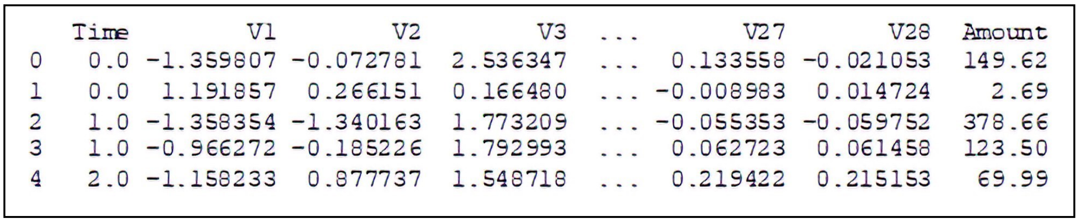
Chapter 6. RESULTS

Figure 7 describes the dataset description followed by figure 8 displays the dataset of 5 rows.



```
                Time            V1  ...         Amount
count  284807.000000  2.848070e+05  ...  284807.000000
mean    94813.859575  3.919560e-15  ...      88.349619
std     47488.145955  1.958696e+00  ...     250.120109
min         0.000000 -5.640751e+01  ...       0.000000
25%     54201.500000 -9.203734e-01  ...       5.600000
50%     84692.000000  1.810880e-02  ...      22.000000
75%    139320.500000  1.315642e+00  ...      77.165000
max    172792.000000  2.454930e+00  ...   25691.160000
```

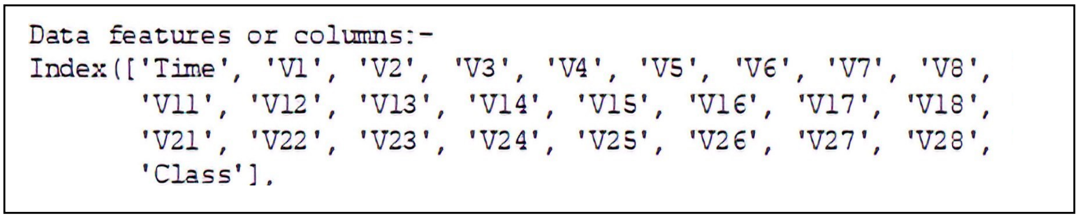
Figure 7. Dataset Description



```
   Time        V1        V2        V3  ...       V27       V28  Amount
0   0.0 -1.359807 -0.072781  2.536347  ...  0.133558 -0.021053  149.62
1   0.0  1.191857  0.266151  0.166480  ... -0.008983  0.014724    2.69
2   1.0 -1.358354 -1.340163  1.773209  ... -0.055353 -0.059752  378.66
3   1.0 -0.966272 -0.185226  1.792993  ...  0.062723  0.061458  123.50
4   2.0 -1.158233  0.877737  1.548718  ...  0.219422  0.215153   69.99
```

Figure 8. Dataset

In figure 9 displays all the columns of the dataset followed by figure 10 histogram that displays all the columns value chart.



```
Data features or columns:-
Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8',
       'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18',
       'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28',
       'Class'],
```

Figure 9. Columns

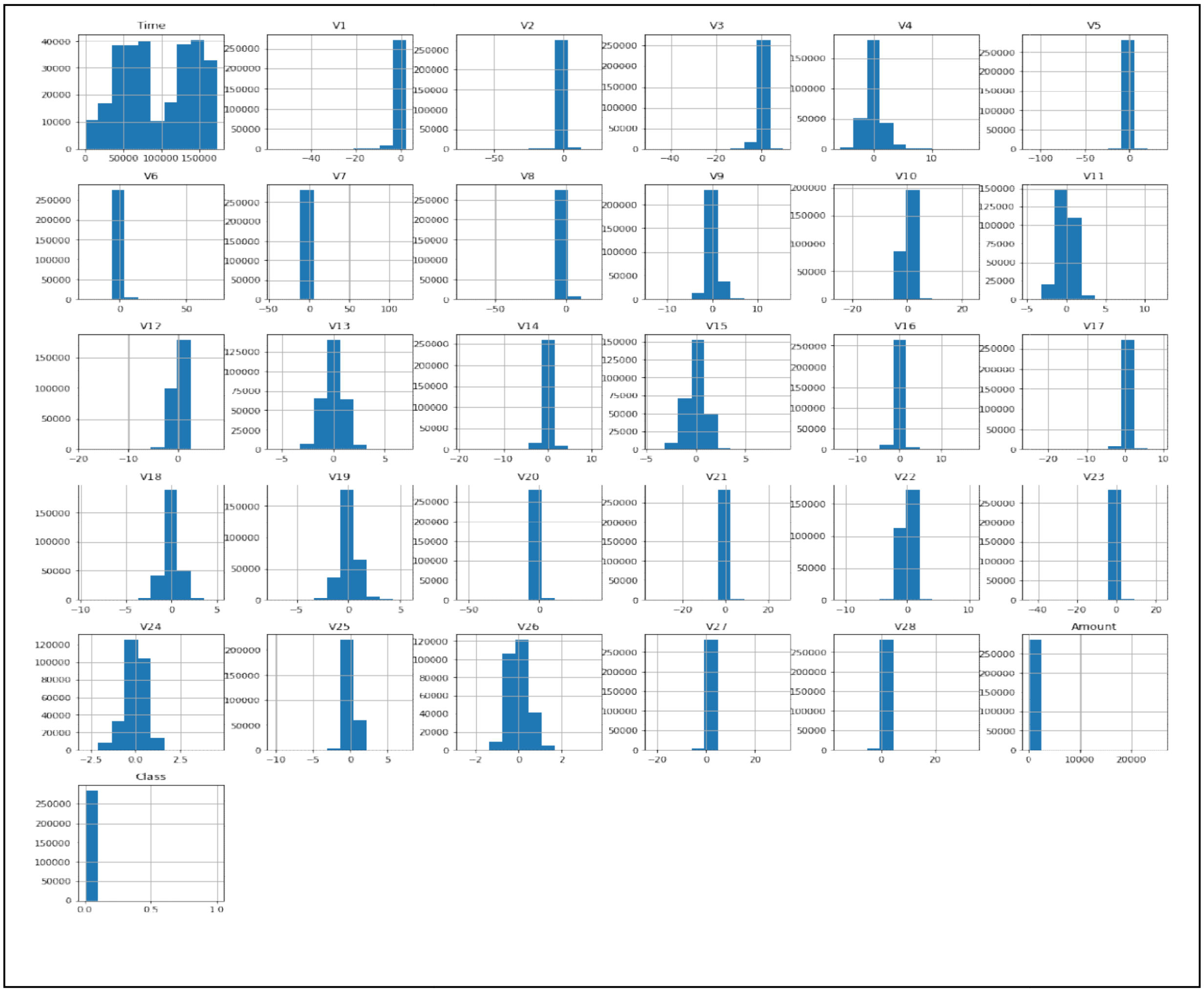


Figure 10. Histograms

In figure 11 there is the correlation matrix of all the columns of the dataset in figure 12 displays the Imbalanced dataset with all fraud and non-fraud transactions.

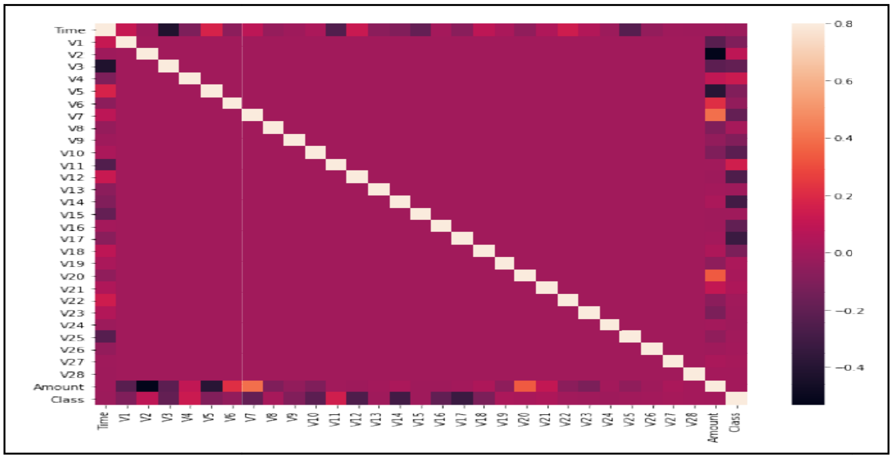
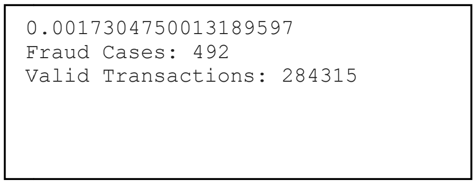


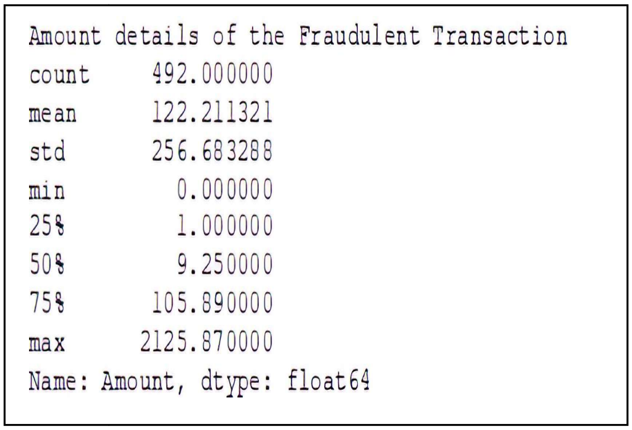
Figure 11. Correlation Matrix



```
0.0017304750013189597
Fraud Cases: 492
Valid Transactions: 284315
```

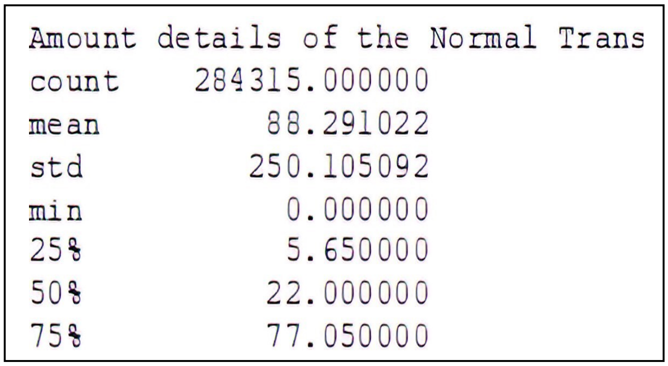
Figure 12. Imbalance Dataset

Figures 13 and 14 display all the amount details of both fraudulent and legitimate transactions.



```
Amount details of the Fraudulent Transaction
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

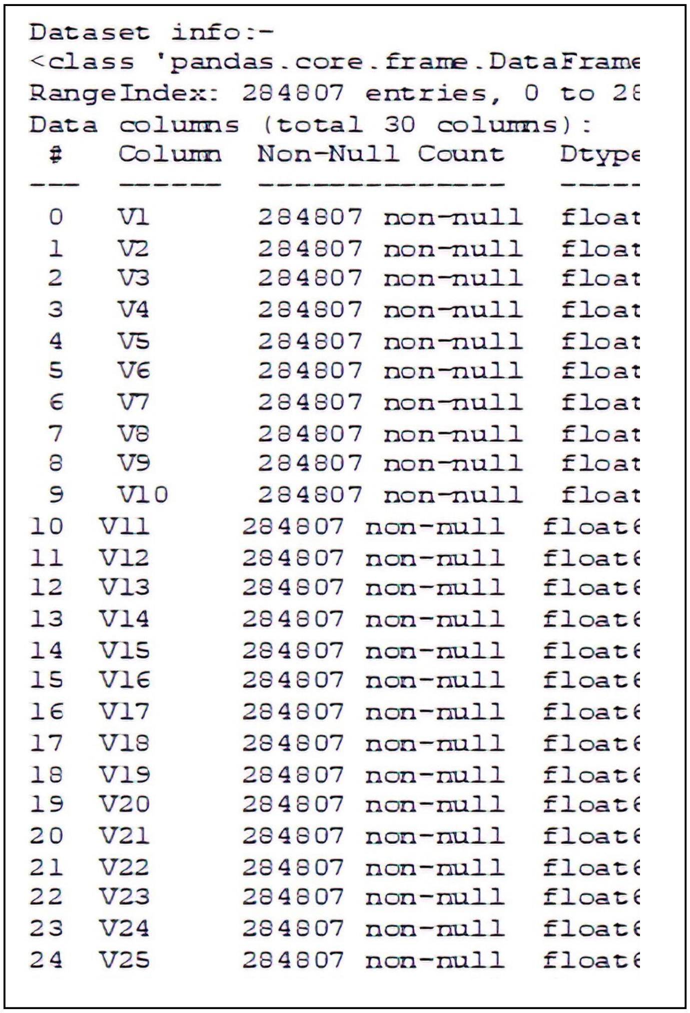
Figure 13. Amount Details of Fraudulent Transactions



```
Amount details of the Normal Trans
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
```

Figure 14. Amount Details of Legitimate Transactions

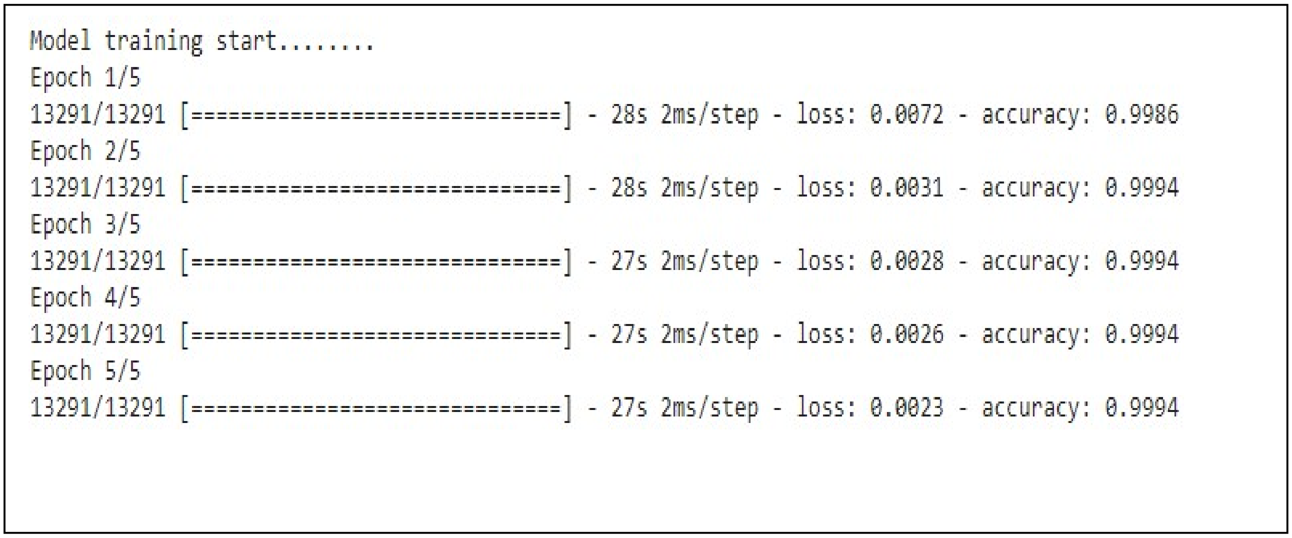
In figure 15 describe all the information of the dataset in null and non-null values.



```
Dataset info:-
<class 'pandas.core.frame.DataFrame
RangeIndex: 284807 entries, 0 to 28
Data columns (total 30 columns):
 #    Column   Non-Null Count    Dtype
---   ------   --------------    -----
 0    V1       284807 non-null   float
 1    V2       284807 non-null   float
 2    V3       284807 non-null   float
 3    V4       284807 non-null   float
 4    V5       284807 non-null   float
 5    V6       284807 non-null   float
 6    V7       284807 non-null   float
 7    V8       284807 non-null   float
 8    V9       284807 non-null   float
 9    V10      284807 non-null   float
10   V11      284807 non-null   float6
11   V12      284807 non-null   float6
12   V13      284807 non-null   float6
13   V14      284807 non-null   float6
14   V15      284807 non-null   float6
15   V16      284807 non-null   float6
16   V17      284807 non-null   float6
17   V18      284807 non-null   float6
18   V19      284807 non-null   float6
19   V20      284807 non-null   float6
20   V21      284807 non-null   float6
21   V22      284807 non-null   float6
22   V23      284807 non-null   float6
23   V24      284807 non-null   float6
24   V25      284807 non-null   float6
```

Figure 15. Dataset Information

Figure 16 shows the training of neural networks. Each line corresponds to each round of forwarding and backward propagation.



```
Model training start........
Epoch 1/5
13291/13291 [==============================] - 28s 2ms/step - loss: 0.0072 - accuracy: 0.9986
Epoch 2/5
13291/13291 [==============================] - 28s 2ms/step - loss: 0.0031 - accuracy: 0.9994
Epoch 3/5
13291/13291 [==============================] - 27s 2ms/step - loss: 0.0028 - accuracy: 0.9994
Epoch 4/5
13291/13291 [==============================] - 27s 2ms/step - loss: 0.0026 - accuracy: 0.9994
Epoch 5/5
13291/13291 [==============================] - 27s 2ms/step - loss: 0.0023 - accuracy: 0.9994
```

Figure 16. Training of Neural Network

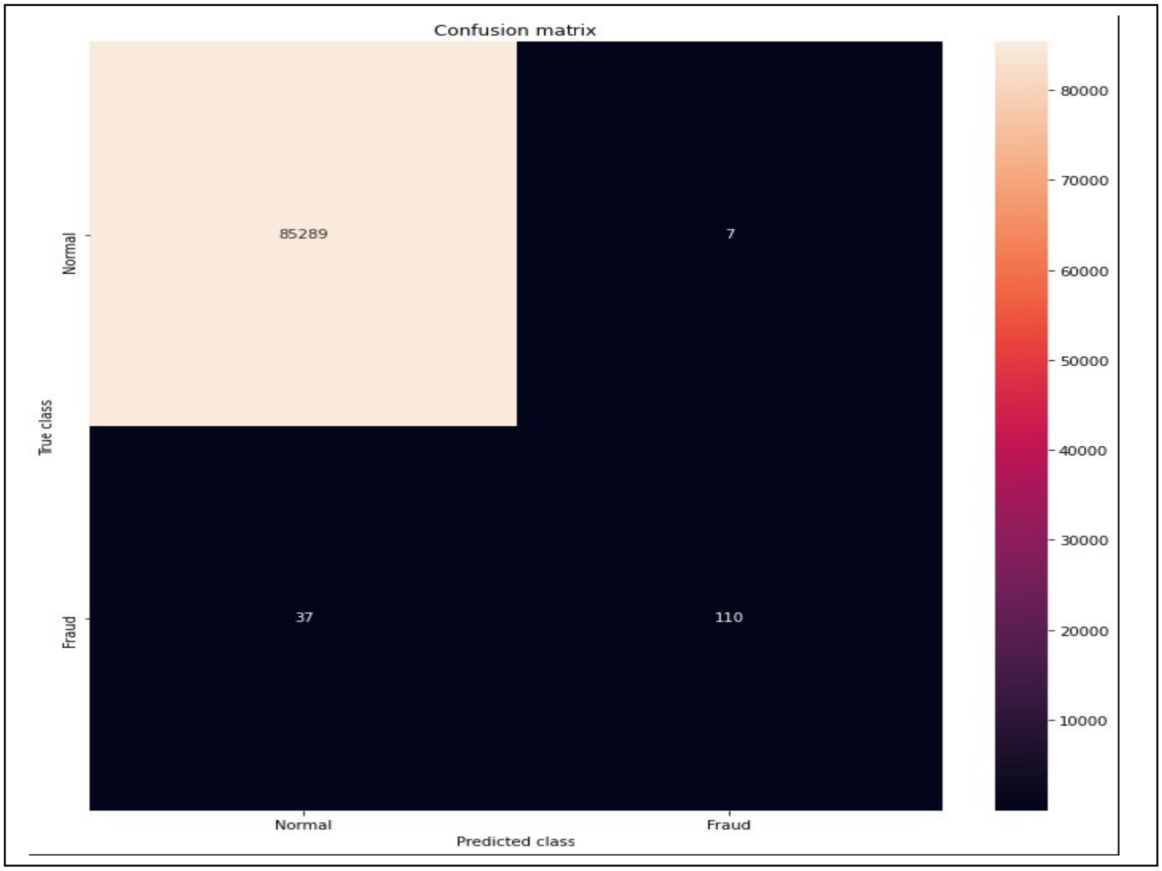


Fig 17. Confusion Matrix of Random Forest for test set

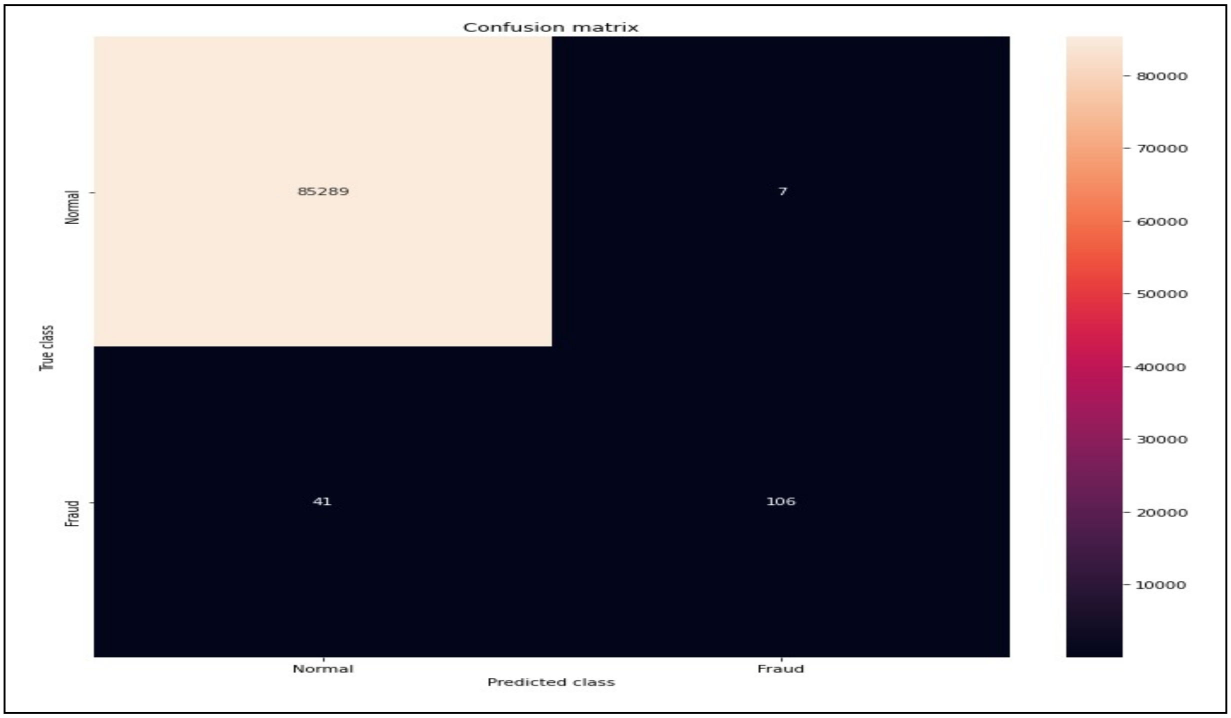


Fig 18. Confusion Matrix of K-NN algorithm for test set

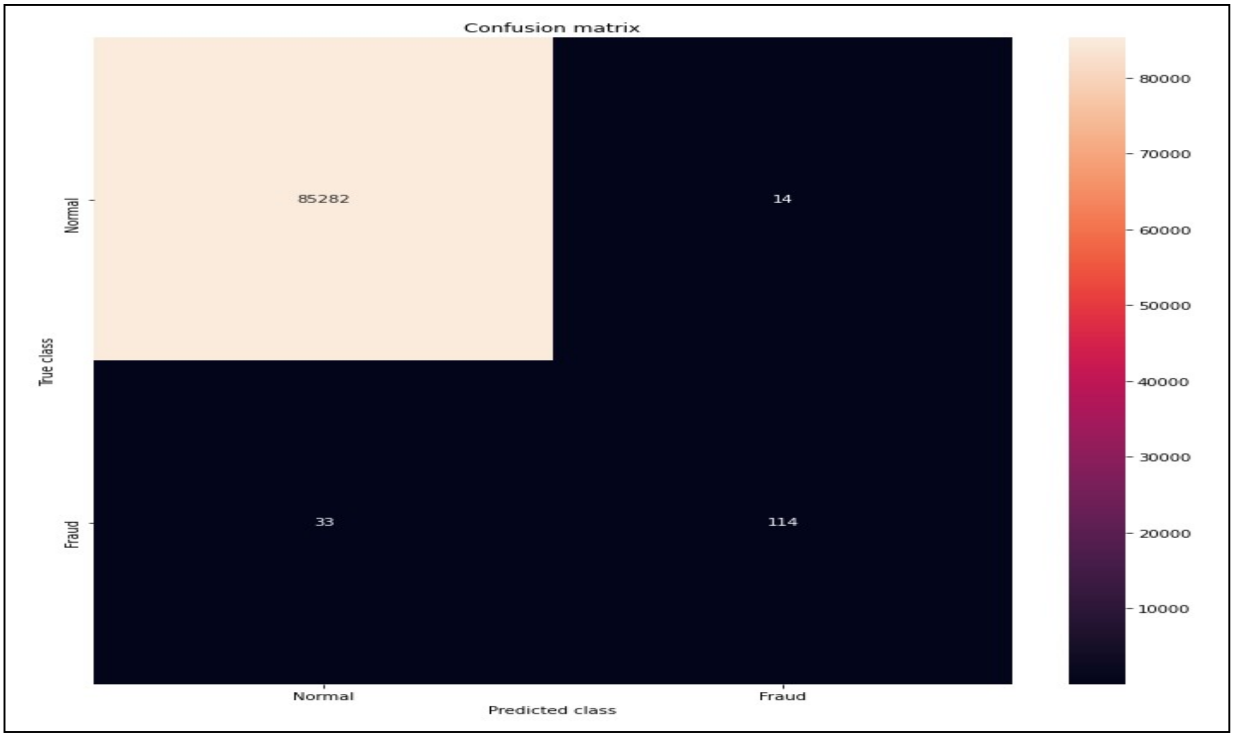


Fig 19. Confusion Matrix of ANN for test set

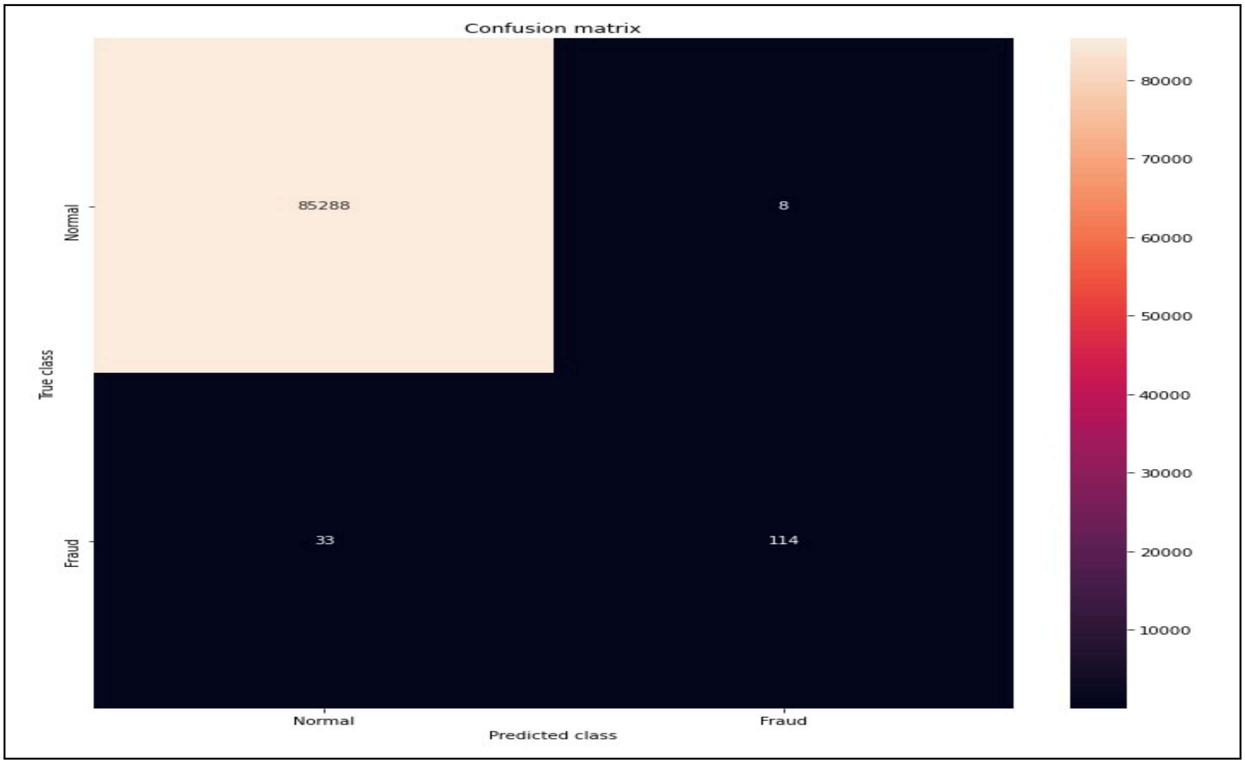


Fig 20. Confusion Matrix of Hybrid Model for test set

This chapter encapsulates the main attainments of this research work. It discusses the future research work that could be done towards the achievement of the eventual goal in the field of the development of an effective and affordable credit card fraud detection system. The organization would curb the growing menace of fraudulent acts.

Chapter 7. DISCUSSION

Test Set: 85443 Transactions

Non-Fraudulent: 85296 Transactions

Fraudulent: 147 Transactions

Table 1. Classification Report for 1 Class i.e (147) Fraudulent Transactions

Algorithms	Precision	Recall	F1-score
Random Forest (RF)	0.94	0.75	0.83
K-Nearest Neighbor (KNN)	0.94	0.72	0.82
Artificial Neural Network (ANN)	0.89	0.78	0.83
Hybrid Model	0.95	0.79	0.85

Table 3. AROC Score of Algorithms

Algorithms	AROC Score
Random Forest	0.8741086262867555
KNN	0.8605031841098847
ANN	0.8876730348864362
Hybrid Model	0.8945167895520617

Table 4. Classification Report for 0 Class i.e (85296) Non-Fraudulent Transactions

Algorithms	Precision	Recall	F1-score
Random Forest (RF)	1.0	1.0	1.0
K-Nearest Neighbor (KNN)	1.0	1.0	1.0
Artificial Neural Network (ANN)	1.0	1.0	1.0
Hybrid Model	1.0	1.0	1.0

The above table shows that by applying the Hybrid model using Majority Voting, every metric improves in the case of Precision, Recall, and F-1 score for class 1 transactions(fraud).

In the case of class 0 transactions(legitimate), every metric for every model has a value of 1.0. The reason behind this is the dataset imbalance problem where the model was trained moreover legitimate ones as compared to fraudulent ones. So every model is much more capable of detecting legitimate transactions as compared to fraud transactions. In such a case accuracy is not a good metric for evaluation.

In ANN, for avoiding underfitting, overfitting, and improving the training speed we have taken batch size = 15 and Epoch = 5.

AROC Score of Hybrid model is highest i.e 0.8945167895520617.

By default, K-NN has parameter n_neighbors = 5 which is a good parameter value for K-NN working.

Random Forest has n_estimators = 10 by default i.e no of Decision Trees in the forest which is a good parameter value for Random Forest efficient working.

The Hybrid model can detect any transaction as fraudulent or legitimate with great accuracy with the Majority Voting approach(Taking opinions of 3 algorithms for prediction).

Chapter 8. CONCLUSION AND FUTURE SCOPE

8.1 OVERALL CONCLUSION

In this development, we build a Hybrid model by combining three different algorithms Random Forest, Artificial Neural Network(ANN), and K-Nearest Neighbor (KNN), and utilized the individual's results of each model in the Majority Voting concept to predict whether the given transaction is fraudulent or legitimate.

The Precision, Recall, F-1 score, and AROC value of the Hybrid approach respectively were 0.95, 0.79, 0.85, and 0.8945167895520617.

Overall we can conclude that the model that we built is more powerful and affirm to manage adequately well for wholly sizes and varieties of the dataset with accurate results since for a particular transaction we get to know the opinions of three different algorithms.

8.2 FUTURE SCOPE

There are many fraud detection techniques accessible present-day but any technique can not recognize all cheats viably when they are done, they generally identify it after the misrepresentation has been submitted. This shows up because not many exchanges out of the absolute exchanges are deceitful. So we need an innovation that can identify the fraudulent exchange when it happens so it very well may be halted without further ado and that too at least expense. So the significant errand of present-day is to make accurate, precise, and fast detecting fraud detection system for credit card frauds that can identify not just fake occurring over the web like phishing and website cloning yet additionally messing with the credit card itself for example it flags an alert when the altered Mastercard is being utilized. This proposed model that we implemented deals with the Data Imbalancing problem which can be solved by applying Undersampling or Oversampling techniques. It can be further extended by adding more algorithms along with taking care of the computation cost and training time. In the case of an even number of models, the majority voting for the fraudulent or legitimate transaction may get the same. For this, it is recommended to go with the powerful model which has a high F-1 score. Mathews Correlation Coefficient(MCC) like metrics can be used in the future since accuracy is not a good metric to measure a model in the case of the Data Imbalance problem. Optimization Techniques, Hyper-Parameter, or Tuning algorithms like for example Genetic Algorithm can be added with the models to reduce its computation cost and training time thus achieving more better results.

REFERENCES

[1] P. Chan and S. Stolfo, “Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection,” KDD, 1998.

[2] SamanehSorournejad, Z. Zojaji, R. E. Atani, and A. H. Monadjemi, “A survey of credit card fraud detection techniques: Data and technique-oriented perspective,” arXiv [cs.CR], 2012.

[3] Rohilla, Anju, and Ipshita Bansal. "Credit Card Frauds: An Indian Perspective." Volume 2, 2015: 591-597.

[4] Maes, Sam, et al. "Credit card fraud detection using Bayesian and neural networks." Proceedings of the 1st international naiso congress on neuro-fuzzy technologies. 2002.

[5] Save, Prajal, et al. "A novel idea for credit card fraud detection using a decision tree." International Journal of Computer Applications 161.13 (2017).

[6] F. Braun, O. Caelen, E. N. Smirnov, S. Kelk, and B. Lebichot, “Improving card fraud detection through suspicious pattern discovery,” in Advances in Artificial Intelligence: From Theory to Practice, Cham: Springer International Publishing, 2017, pp. 181–190.

[7] Kiran, Sai, et al. "Credit card fraud detection using Naïve Bayes model-based and KNN classifier." International Journal of Advanced Research, Ideas, and Innovations in Technology 4.3 (2018).

[8] Campus, Kattankulathur. "Credit card fraud detection using machine learning models and collating machine learning models." International Journal of Pure and Applied Mathematics 118.20 (2018): 825-838.

[9] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine learning for credit card fraud detection system." Int. J. Appl. Eng. Res. 13.24 (2018): 16819-16824.

[10] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, “Random forest for credit card fraud detection,” in 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, pp. 1–6.

[11] Divakar, Kavya, and K. Chitharanjan. "Performance evaluation of credit card fraud transactions using boosting algorithms." Int. J. Electron. Commun.Comput. Eng. IJECCE 10.6 (2019): 262-270.

[12] U. Porwal and S. Mukund, “Credit card fraud detection in E-commerce,” in 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2019, pp. 280–287.

[13] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit card fraud detection - machine learning methods,” in 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, pp. 1–5.

[14] 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. (Ny), 2019.

[15] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, “Real-time credit card fraud detection using machine learning,” in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 488–493.

[16] CHILAKA, UL, GA CHUKWUDEBE, and A. BASHIRU. "A Review of Credit Card Fraud Detection Techniques in Electronic Finance and Banking."

[17] A. A. Taha and S. J. Malebary, “An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine,” IEEE Access, vol. 8, pp. 25579–25587, 2020.

[18] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit card fraud detection using pipelining and ensemble learning,” Procedia Comput. Sci., vol. 173, pp. 104–112, 2020.

[19] SG, Kruthika, and S. R. Manjunatha. "A survey on SMOTE Deep: Novel link-based classifier for fraud detection."

[20] M.Mohanapriya, and M. Kalaimani.“Credit Card Fraud Detection”.IARJSET.2020.7903.

[21] P. Tiwari, S. Mehta, N. Sakhuja, I. Gupta, and A. K. Singh, “Hybrid method in identifying the fraud detection in the credit card,” in Evolutionary Computing and Mobile Sustainable Networks, Singapore: Springer Singapore, 2021, pp. 27–35.

[22] A. M. Nancy, G. S. Kumar, S. Veena, N. A. S. Vinoth, and M. Bandyopadhyay, “Fraud detection in credit card transaction using hybrid model,” in 1ST INTERNATIONAL CONFERENCE ON MATHEMATICAL TECHNIQUES AND APPLICATIONS: ICMTA2020, 2020.

[23] N. Mqadi, N. Naicker, and T. Adeliyi, “A SMOTE based Oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection,” Int. J. Comput. Digit.Syst., vol. 10, no. 1, pp. 277–286, 2021.

[24] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” in 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1–9.

[25] Harsh Harwani, Jenil Jain, ChinmayJadhav, ManasiHodavdekar, Ed., Credit Card Fraud Detection Technique using Hybrid Approach: An Amalgamation of Self Organizing Maps and Neural Networks, vol. 07, no. 2020. International Research Journal of Engineering and Technology (IRJET), 2020

[26] N. Shirodkar, P. Mandrekar, R. S. Mandrekar, R. Sakhalkar, K. M. Chaman Kumar, and S. Aswale, “Credit card fraud detection techniques – A survey,” in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE),2020, pp. 1–7.

[27] X. Kewei, B. Peng, Y. Jiang, and T. Lu, “A hybrid deep learning model for online fraud detection,” in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 431–434.

[28] B. J. Kaur and R. Kumar, “A hybrid approach for credit card fraud detection using naive Bayes and voting classifier,” in Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2019), Cham: Springer International Publishing, 2020, pp. 731–740.

[29] T. K. Behera and S. Panigrahi, “Credit card fraud detection: A hybrid approach using fuzzy clustering & neural network,” in 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015, pp. 494–499.

[30] Suma, V., and ShavigeMalleshwara Hills. &quot; Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics.&quot; Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

[31] Chandy, Abraham. &quot; Smart resource usage prediction using cloud computing for massive data processing systems.&quot; J InfTechnol 1, no. 02 (2019): 108-118.

ANNEXURE I

Installation Instructions:

The system must have installed Python version-2 or above with essential Data Science libraries given below:

Numeric Computing libraries – Numpy and Pandas.

Plotting libraries – Seaborn and Matplotlib.

Statistics and Machine learning libraries – Scipy, Scikit-learn, Imblearn, Tensorflow, and Statsmodels with an Integrated Development Environment (IDE) like Jupyter or Pycharm or Google Collaboratory(an online platform).

The selection of appropriate tools and techniques is essential for the execution of the "Self Calibrating Automated Credit Card Fraud Detection System"(Hybrid approach).

To install Imblearn library:- !pip install imblearn

To install TensorFlow library:- !pip install tensorflow

ANNEXURE II

Contribution of the team members:

- **Kshitij Pandey**: Implemented the code for the Random Forest algorithm, collected the required dataset, explored it and preprocessed it, and worked on the report.
- **Piyush Sachan**: Implemented the code for K- Nearest Neighbors algorithm and also worked on the report.
- **Shakti**: Implemented the code for ANN(Artificial Neural Network), executed the Majority Voting approach, and worked on the report.

Together we all have read 31 papers which comprise survey papers, implementation papers, etc related to Credit Card Fraud Detection to understand the problem, how these problems are handled through machine learning and artificial intelligence, what are the existing methodologies, and the current state of the art techniques.

ANNEXURE III

Publications if any:

Publication: Review paper

Paper accepted: ICCMC 2021 conference

Paper id: ICCMC298

Conference Name: 5th International Conference on Computing Methodologies and Communication (ICCMC 2021)

Indexing: Scopus

Paper accepted and presented at ICCMC 2021. Published on IEEE Xplore.