Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

A Review of Credit Card Fraud Detection Techniques

Kshitij Pandey[1], Piyush Sachan[1], Shakti[1] and Nikam Gitanjali Ganpatrao[1]
[1]Department of Computer Applications, National Institute of Technology Kurukshetra, Haryana, India

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418024

***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 used state of the art techniques to counter these attacks and highlights its limitations along with proposing a solution for it.**

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

I. Introduction

Nowadays usage of credit cards in developing countries has become an advent. Users use it for purchasing, take care of tabs, and for 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.

Apart from this, the review paper has successive segments. Segment II clarifies and sums up every one of the current strategies that can be utilized in an anomaly detection system. Segment III talks about gaps recognized alongside the experiments done and models being invented to date. Segment IV is a general architecture of the different such strategies based on some parameters or metrics such as Recall, Specificity, Confusion Matrix, Precision, etc. Section

V is the conclusion and finally, section VI is about the future scope of the existing methodologies.

II. 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 model. This cycle can be mechanized yet it is unavoidable 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, misrepresentation types, datasets, and advancement rules. There is no benchmark set for the dataset plus no standard dataset is available. A credit card is naturally a personal property, so converting it to a genuine benchmark for this card is problematic. Subsets of datasets can lead fraud detection frameworks to learn fraud stunts or ordinary conduct somewhat. Fraud analysts used datasets that belong to banks and are created by them. It can't be distributed due to security reasons. A legitimate informational collection is an informational collection that covers different 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] execute 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 Relapse, 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] a sequential model is utilized for fraud detection. This implies that the model was trained in the anticipated result of the CNN model was fed into the training data for the KNN this hence, builds the accuracy rate as the probability of diminishes by the cumulative error rate of both primitive models. We have utilized CNN as the primary phase of the hybrid model. After training the data for 490 cycles specific accuracy has 87.79% and a logarithmic deficiency of 3.90. It is trained by KNN classification, which has a 90.5%. Upon effective hybridization, the resultant model has an accuracy of 98% with a logarithmic loss of 0.647.

In [24] highly skewed credit card fraud data on the presentation of Naïve Bayes, Logistic Regression, and K-Nearest Neighbor. Dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions. A hybrid approach of oversampling and undersampling is completed on skewed data. The 3 procedures apply to the preprocessed and raw data. The work is carried out in Python. The exhibition of the procedures is assessed dependent on Matthews's Correlation Coefficient, Sensitivity, Accuracy, Precision, Specificity, and Balanced Classification Rate. The outcomes derived ideal accuracy for Logistic Regression, KNN and Naïve Bayes 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.

III. GAPS IDENTIFIED

We have found out that the dataset is highly imbalanced. It contains more genuine transactions as compared to fraud 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 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 dataset contains credit card transactions made by European cardholders in September 2013. This dataset 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.

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 validation	Luhn's algorithm Bayes theorem	Validation of the card is a genuine and very low false alarm.	Larger 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 Perceptron), NB, RF, and LR.	RF LR MLP NB	Different ML algorithms can give reasonable results with appropriate pre-processing. SMOTE improves random oversampling. RF was found out to the best.	Computational cost is high in a 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, CNN's accuracy rate increased by 10%.	Will work better if trained over a large balanced dataset.
[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 K-Nearest Neighbor, Logistic Regression, Naïve Bayes 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	KNN shows significant performance for all metrics evaluated except for accuracy in the 10:90 data distribution.

			LR: 54.86% Accuracy	LR accuracy decreased rapidly in 34:66 and 10:90 distribution.
[25]	To detect fraud in real-time and efficiently.	Hybrid approach: SOM and ANN	In this model, we achieved better accuracy precision and cost compared to using SOM or ANN alone.	Do not give the same result when 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.

Multi-Classifier Meta-Learning Approach: The methodology is to make information subsets with the ideal conveyance i.e. 50-50, create classifiers from the subsets by applying ML algorithms (unique on every subset) on them, and coordinates them by learning (meta-learning) from their classification conduct.

Artificial Neural Network (ANN): In this collection of neurons is used to do some tasks. Neurons are the basic building blocks of ANN. It is like a human brain. It is used for pattern searching, classification problem solving, etc. It consists of layers where neurons are there. It can be used for both regression and classification. It uses neurons as the deciding edges and sites between neurons to compute the benefaction of each neuron in the preceding layer in the result and decision at the current neuron. The training of ANN can be supervised i.e. outcome is already noted. It predicts the outcomes and compares them with the existing outcomes and learns from them. Its training can be unsupervised also. There is no information regarding the outcomes. It has to learn by itself. Here to minimize the cost function it uses Gradient Descent. The information is entered into the input layer and then is propagated forward to get our y-hats and then compared with actual values and calculate errors. Then these errors are back-propagated in the opposite direction that allows training the network by adjusting the weights simultaneously.

Genetic Algorithm (GA): It looks for optimum arrangements with a populace of applicant arrangements that are generally spoken to as twofold strings called chromosomes. The essential thought is that the more grounded individuals from the populace have more opportunities to endure and repeat. The strength of an answer is its capacity to take care of the basic issue which is shown by wellness. GA has been utilized in data mining errands primarily for feature selection. It is likewise broadly utilized in mix with a different algorithm for boundary tuning and optimization. It is an evolutionary optimization technique. It is based on Darwinian Theory's concept of "survival of the fittest". It is a search algorithm based on the mechanism of natural selection. It stimulates the process of evolution, which is the basic optimizing process making each coming generation better and better. A Genetic Algorithm helps the neural network in the discarding of the unnecessary and insignificant neurons and thus it makes the training and testing faster. The Genetic Algorithm represents the solution in the form of binary strings called chromosomes. The structure of these chromosomes is called a genome and the instance of a genome is called a phenomenon.

Step 1: Store the valid genes (random parameters) into GENES.

Step 2: Create a random combination of parameters of length of Tar (target combination) from the valid genes variable GENES.

Step 3: Calculate the fitness score upon each of the generated gene by fitness () function. The fitness sore increases as the number of characters in the gene differ from the target set of combination. #Creating initial population

Step 4: Sort the population in increasing order of fitness score.

Step 5: If Fitness score is found as less than or equal to 0, target is found, stop.

Step 6: Else

i. 10% of the current fittest population i.e. fittest (or lowest) fitness scores are taken. #generating new population

ii. The individuals present in 50% of the fittest population are mated to produce off-spring. #Crossover

iii. Generate two random combinations of parameters. #generating two new parents from the current population

iv. Calculate probability by choosing a random number from 0 to 1.

1. If probability < 0.45

Then Take parent1 and produce off-spring

2. Else if probability 0.90

Then Take parent2 and produce off-spring

3. Else

Create a new random gene and insert. #Mutation

Step 7: The newly created generation is then stored into new population.

The output is all the generations with their children and their corresponding fitness score.

Fig. 1. Psuedo code of Genetic Algorithm.

Support Vector Machine (SVM): It is a supervised associated with related learning algorithms that can dissect and perceive designs for classification and regression tasks. Its goal is to find the decision boundary or best line that divides n-dimensional space into classes. It is a parallel classifier. Given a dataset, it isolates them into various classes utilizing a hyperplane. The objective of SVM is to discover this hyperplane. There could be numerous hyperplanes however we are resolved to track down an optimal hyperplane. The focuses nearest to the hyperplane in the various classes are called support vectors and these support vectors are utilized to foresee the classes of the new data point.

Luhn's Algorithm: It is used to allow card esteems that recognize genuine numbers from mistyped or regardless mistaken numbers. Keeping standard computation is used to favor credit card numbers. Observing standard technique is utilized to approve credit card numbers-

Step 1: Invert the request for the digits in the number.

Step 2: Take the primary, third ... also, every other odd digit in the switched digits and total them to shape the partial sum S1.

Step 3: Requiring the second, fourth and each other even digit in the switched digits. Multiply every digit by two and whole the digits if the appropriate response is more prominent than nine to form partial totals for the even digits.

Fig. 2. Psuedo code of Luhn's algorithm.

Naïve Bayes Model: Naive Bayes Model: It is dependent on the Bayes theorem and used to solve classification problems. It is based on some independent assumptions like features or variables that should be independent. But in reality, there is some correlation between them. That's why it is known as Naïve. It works well with high-dimensional training data. The model tries to forecast a class which is known as a result class based on probabilities, and also conditional probabilities of its existence from the training data.

The first step for it is the Bayes theorem for conditional probability, where 'y' is given data point and 'N' is a class:

$$P (N/ y) = P(y/N)/P(y)$$

Also, further advances are finished by making the theory for a data point y = {y1 to yj}, and the current probability of every one of its aspects inside a given class is independent. So the probability of y can be determined as follows:

$$P (N/y) = P (N).\Pi P(yi/N)$$

KNN Classifier: It is a non-linear classifier. It assumes the same things appear or are close to one another. It is adaptable since it can be used for regression, classification, and searching. It may get slow when no of the predictors are increased. This algorithm is used for both classification and regression but it is mostly used for classification and one of the most used algorithms. K-NN is a non-linear classifier. First, calculate the value of k for neighbors to find the category where the new data point is classified. To find which category has the nearest neighbor for new data points used Euclidean distance to classify new data and this is the distance rule. After calculating Euclidean distance find which category has the nearest neighbor to make that new data point in that category this is distance metrics.

Logistic Regression: It is a classification algorithm used to give observations to a different set of classes. It converts its output using the sigmoid function to give a probability value. Here it will be predicted the probability of its happening. It overcomes the limitations of Linear Regression since it allows values less than 0 and greater than 1. The sigmoid function is given as-

$$y = b_0 + b_1 x_1$$

$$p = 1 / (1 + e^{-y}) \text{ [Sigmoid function]}$$

If the value of a sigmoid function is less than 0.5 it is converted into 0 otherwise 1.

Decision Tree: It is supervised. Here the data is split in such a way as to increase the no of certain categories in each of these splits. It can be used to decide which points are at which split. It consists of decision nodes and leaves nodes. Decision nodes are for deciding the travel routes and leaves nodes are the final outputs. For building a decision tree take the following steps that first and foremost to Compute the entropy of each attribute utilizing the dataset in issue then the dataset is partitioned into subsets utilizing the attribute for which entropy is least or gain is most extreme after that to build on a decision tree hub comprise that attribute and ultimately recursion is completed on subsets utilizing resting attributes to make a decision tree.

Random Forest: It is a team of Decision Trees. It takes the average of their outcomes to improve predictive accuracy. It is supervised and can be utilized both for regression and classification. Here random k points are chosen and then a Decision Tree is being built on it. These steps are then repeated. The steps of this algorithm are:

Step 1: Pick aimlessly k data points from the Training set.

Step 2: Build the Decision Tree related to these k data points.

Step 3: Choose the number N tree of trees you need to construct and rehash Steps 1 and 2.

Step 4: For another data point, make every single one of your N tree trees anticipate the classification to which the data point has a place and relegate the new data point to the classification that succeeds the highest vote.

Fig. 3. Psuedo code of Random Forest algorithm.

Multi-layer Perceptron: The Multilayer Perceptron is a feed-forward ANN that contains at any rate 3 layers of hubs: an information layer, yield layer, and concealed layer. Each node uses an activation function. It maps the weighted contributions to the yield of every neuron. Multi-layer Perceptron has more than a single neuron's linear layer. The

least difficult model can be of a three-layer framework which has the principal layer as the info layer, the last layer as the yield layer, and the middle layer as the hidden layer. Input data is taken care of to the input layer and output information is gathered from the yield layer. Hidden layers can be expanded however much we need.

Pipelining: Pipelining alludes to the use of a progression of changes followed by the last classifier. The pipeline is utilized for gathering a few various cycles for cross-validation them together at the same time setting up various boundaries.

Ensemble Bagging Classifier: This method alludes to the mix of the different estimates (base) which are created with a particular learning strategy to enhance one single assessor. Here, a bagging classifier has been utilized in which base classifiers are fitted on each random subset of the dataset (real). At that point collection and joining of every expectation are done.

SOM: The dataset which is employed was unlabeled, therefore an unsupervised algorithm namely Self Organizing Map is used to seek out the outliers within the data, further improving precision using Artificial Neural Network. It is an unsupervised learning algorithm which uses simple heuristic method capable of discovering hidden non-linear structure in high dimensional data. SOMs are more profitable to utilize than other clustering techniques since they do not make suppositions in regards to the circulations of variables nor do they require freedom among variables they're simpler to carry out and are prepared to tackle non-linear issues of high complexity. They viably manage discordant and missing data, little dimensional, and tests of limitless size. To perform unsupervised learning, SOMs apply a competitive learning rule where the output neurons compete among themselves for the chance to represent distinct patterns within the input space.

IV. General Architecture

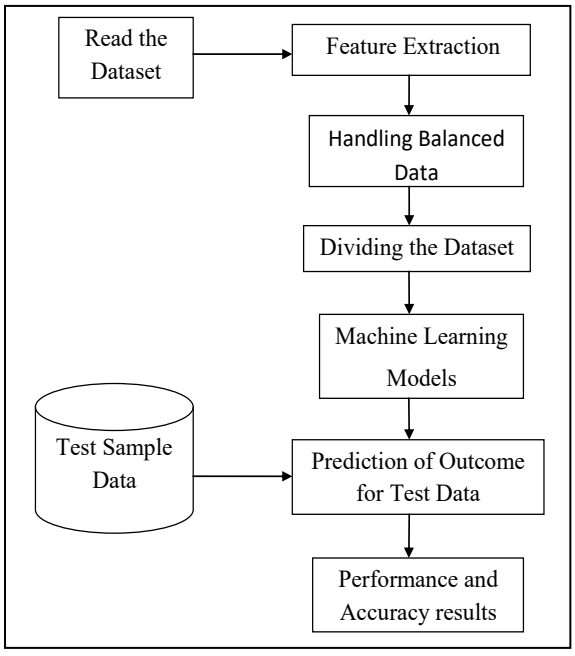

Fig. 4. Salient steps in the fraud detection system of credit card.

In the first step, the dataset will be imported using some methods to read the data values. Some operations will be performed on it to get some essential information regarding the dataset. We will create Matrices of features or Independent variables and a Dependent variable vector from the dataset. These variables will be further used by our model for training and testing purposes.

In the second step, there are some categorical attributes and some numerical attributes in the matrices of features and dependent variable vector but not all are important for finding the result. So we will apply the feature extraction method to find and remove less important attributes. This reduces the dimensionality of our dataset and avoids over-fitting our model.

In the third step, data imbalance means skewed distribution. Here the majority class and minority class concept comes in. A large no of data values is pointed towards one side as compared to the other side. In this block basically, such situations will be handled using some predefined algorithms or methods.

In the fourth step, divide the matrices of features and the dependent variable vector into a training set and test set. The test set will contain less data as compared to the training set.

In the fifth step, the machine learning model will be trained on the training set's matrices of features and performance will be evaluated based on the test set's matrices of features and the dependent variable vector. The model will learn the correlation between independent and dependent variables from the training set and then we will check the correlation on the test set. So better it understands the correlation in the training set, the better it will be predicting the result in the test set. We can choose the parameters of the model by ourselves and also tune them through different algorithms.

In the sixth step, the model will do some predictions and provide outcomes on the test set's matrices of features, and those outcomes will be matched with the test set's dependent variable vector.

In the seventh step, the performance and accuracy of the model will be evaluated through performance metrics like Accuracy, Precision, Specificity, ROC curve, F1-measure, Mathews Correlation Coefficient(MCC), Confusion Matrix(A 2*2 matrix consisting of True Negative(TN), False Negative(FN), True Positive(TP) and False Positive(FP)), etc.

TP- A TP is an outcome where the model predicts accurately the positive class.

TN- A TN is an outcome where the model predicts accurately the negative class.

FP- A FP is an outcome where the model predicts inaccurately the positive class.

FN- A FN is an outcome where the model predicts inaccurately the negative class.

Accuracy- Accuracy is the percentage of accurate predictions for the test data. It can be computed easily by partitioning the number of accurate predictions by the number of total predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision- Precision is the fraction of true positives among all of the positive classes.

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

Recall- Recall is the fractions that were predicted to belong to a class concerning all of the instances that truly belong to the class.

$$\text{Recall} = TP / (TP + FN)$$

V. Conclusion

After doing the comparative study, it has been found that every algorithm contains its focal points and weaknesses. There are many fraud detection techniques available but any of the techniques is not able to detect fraud when it is happening it detects when fraud is already done because a very less number of transactions are fraudulent. So the solution to this problem is that we need a technology that can detect fraud when fraud is happening. So the significant task of the time being is to construct a correct and quick recognizing model that can identify not just fraud occurring on the internet but also tampering with credit cards themselves. The gaps of all the algorithms that they are not assured to provide similar outcomes in every condition. They provide better outcomes with a specific sort of dataset and bad or inadmissible outcomes with another kind. A few strategies like Random Forest and KNN work well in a small set of data and not flexible to a large dataset. Some like SVM and Decision Tree give good outcomes on pre-processed and sampled data while some like ANN have learned from the past and the Genetic Algorithm is fast in detection. Some techniques like Fuzzy systems and Logistic Regression give better results inaccuracies with raw un-sampled data. Dataset is already transformed by the banks due to security issues. It is also highly imbalanced and not easily available in many implementations.

VI. Future scope

The ability of Artificial Neural Networks to learn from the past and achieve a high-performance rate but the limitations that high processing time for the larger neural network. For fast detection, it is advisable to use some optimization techniques like the Genetic Algorithm, Adam optimizer, Stochastic Gradient, etc. For improving the random oversampling SMOTE approach can be utilized, so for successful credit card fraud detection construct a hybrid approach of several algorithms that are utilized in credit card fraud detection for removing their drawbacks and achieve the highest accuracy is advisable.

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, Chinmay Jadhav, Manasi Hodavdekar, 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 Shavige Malleshwara Hills. " Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

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