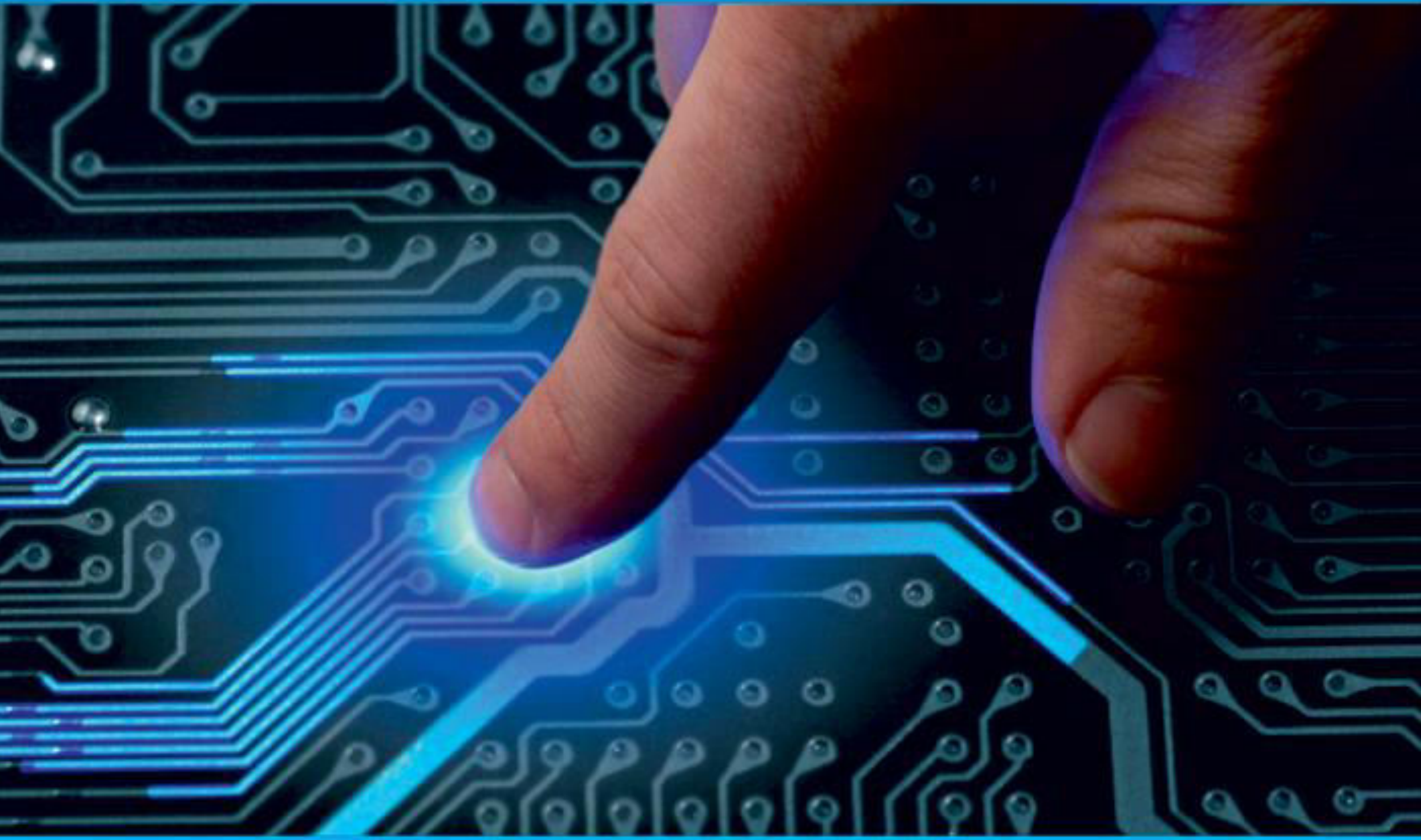
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

Efficient Model to detect Credit Card Fraud Detection through Hidden Markov Model

Devyani Jadhav, Prashant Kalokhe, Savita Mahamuni, Sonal Kamble , Prof. Pooja Oza

Department of Computer Engineering, JSCOE Hadpsar, Pune, India

ABSTRACT: The Fraudsters have been successful in evading the financial institutions' security protocols by employing a variety of approaches. Banks and other credit services have been extremely successful in developing solid security mechanisms that may considerably minimize credit card fraud, but not totally eliminate it. This is owing to the fact that criminals' strategies are exceedingly elaborate and intricate, resulting in a large number of variables that must be understood in order to detect them. As a result, in order to propose a solution for this problem, this research study employs deep learning methodologies to accomplish the goal. The methodology implements Linear Clustering and Entropy Estimation along with Hidden Markov Model and Decision Tree for achieving the Credit card fraud detection accurately. The experimental results for error rate evaluation have provided promising levels of accuracy which have been stipulated in the later sections of this research work.

KEYWORDS: Hidden Markov Model, Linear Clustering, Entropy Estimation and Decision Tree.

I. INTRODUCTION

A credit card is a compact, thin plastic or fiber card that carries information about the individual, such as a photo or signature, and allows the person identified on it to charge products and services to his connected account, which is deducted regularly. ATMs, swiping machines, retail readers, banks, and Internet transactions all read card information these days.card's physical security as well as the privacy of the credit card number.

The number of credit card transactions is rapidly increasing, which has resulted in a significant increase in fraudulent incidents. In credit card transactions, 'fraud' refers to the unlawful and unwelcome use of an account by someone who is not the account's owner.

To stop this misuse, necessary preventative steps should be adopted, and the behavior of such fraudulent acts may be analyzed to decrease it and guard against future occurrences. In other words, credit card fraud occurs when a person uses another person's credit card for personal gain while the owner and card issuing authorities are ignorant of the transaction. To identify fraud, a variety of data mining and statistical tools are employed. Artificial intelligence and pattern matching are used in several fraud detection strategies. It is critical to detect fraud using effective and secure ways.

Credit card fraud is on the rise, and as a result, financial losses are skyrocketing. As new technology emerges, the Internet or online transactions are expanding in popularity. Credit cards account for the majority of these transactions. To decrease these losses, fraud protection or detection must be implemented. As technology develops at a quick pace, several forms of scams emerge. Fraud detection is tracking the behaviors of large groups of people to predict, detect, or avert unacceptable behavior such as fraud, intrusion, or default.

This is a very significant subject that necessitates the attention of groups such as machine learning and data science, where the answer may be automated. This problem is especially difficult to solve from the standpoint of learning since it is characterized by many aspects such as class imbalance. The number of genuine transactions considerably outnumbers the number of fraudulent transactions. Furthermore, transaction patterns frequently modify their statistical features over time.

The number of viable machine learning applications for combating criminal activity is continually expanding, and it is impossible to include all deserving instances here. So, while various machine algorithms are utilized to identify fraud, hybrid algorithms are increasingly being employed due to their superior performance. However, in this research, we use machine learning algorithms to detect credit card fraud based on time and transaction amount.

The second section of this research article focuses on detailing current field research. In Section 3 of the recommended methodology, the implementation is detailed. The outcome is detailed in Section 4, which is titled "Results and Discussions. Finally, part 5 concludes this research project, as well as the scope of future developments.Eachch card has a unique card number, which is extremely significant; the card's security is mostly dependent on the

II. LITERATURE SURVEY

A. Benchaji describes there is a vast growth in the field of credit card transactions, as there is credit card fraud identification is widely increasing. Credit card fraud identification system includes major challenges such as the number of fraudulent transactions is increasing and data sets are highly imbalanced. [1]Thus the proposed paper implements a credit card fraud detection system using K-means clustering and the genetic algorithm. Thus the proposed paper implements an effective framework for credit fraud detection.

K.Hafiz specifies to reduce financial loss in credit card companies in Canada the system is developed called adopted fraud monitoring solutions called has Predictive Analytics Technologies (PAT). [2] Criteria, features, and capabilities are the relevant evaluation scorecard where the authors focused. Thus the proposed paper also described five credit card predictive analytics.

I.Benchaji has there is a large scale growth in the field of financial fraud crimes due to there is a huge loss of amount in the finance industry. [3] The main aim of the proposed paper is to detect frauds where legal transactions are used as a huge dataset. Thus the developing fraud detection is known to be done on imbalanced datasets is a distinct and major challenge of fraud detection. They have used K-means clustering and the genetic algorithm to gain accurate fraud detection.

F.Ghobadi expresses fraud is growing in bank transactions due to the electronic payment systems and rapid increase in e-business. [4] Thus the researchers of the proposed paper use machine learning algorithm like Artificial Neural Networks (ANN) to develop the credit card fraud detection (CCFD) model. [4] ANN technique is used for credit fraud detection due to the imbalanced dataset. With Artificial Neural Networks the proposed model also uses a Cost-Sensitive Neural Network (CSNN) for cost-saving and effective results.

F.El hlouli narrates their growth in services of digitalization of banking and growth in mobile banking applications. [5] Thus the author of the presented methodology uses Extreme Learning Machine (ELM), artificial neural network classifiers, and Multilayer Perceptron (MLP) it is utilized on the dataset containing fraud being done with credit card. The accuracy rate of the proposed approach is evaluated on basis of precision, classification time, accuracy, and recall and this achieves 97.84% and 95.46% by ELM classifiers.

A.A.Khine explains due to the real-time industrial applications data mining and data streaming is known as a very hot topic to research and continue the data is generating. In the modern business organizations environment is rapidly changing in the term of information systems and many applications. [6] Two main methods of the data mining technique are classification and clustering. Thus with this classification algorithm decision, tree learning method and data stream classification learning methods are implemented in the proposed technique.

K.Modi describes nowadays financial transactions becoming more popular because mobile wallets, online transactions, and transactions on credit cards are cashless transactions. [7] As there is a vast growth in the field of cashless transactions there is also growth in fraudulent transactions. Credit card companies and banks are using fuzzy clustering approach, hidden Markov model; rule-based mining and the neural network are some data mining technique used to detect the fraud behavior.

S.Khatri explains credit card transactions are nowadays commonly used and usage of electronic payment techniques such as debit cards and credit cards. But this technique also came with a set of problems. [8] Thus to solve these problems the proposed paper came with a solution by implementing machine learning algorithms. Thus they have used Logistic Regression and Random Forest models with a combination of kNN, Naive Bayes, and Decision Tree by using an imbalanced dataset.

A.Thennakoon describes in huge financial losses credit card fraud detections plays a major part. A huge amount of these transactions is of online transactions have grown rapidly. There has been a huge amount of demand from banks and financial institutions to develop the framework for detection of credit card fraud. [9] The dataset collected from the site has to go under three stages as Data Cleaning, Data Integration, and Data Transformation, and then the machine learning algorithm is realized.

Y.Lucas states to detect credit card frauds data mining techniques is been is used on large scale. They have to face many challenges regarding dataset shift. [10] The dataset is divided into four parts such as Sundays, school holidays, working days, and Saturdays. With data mining technique machine learning algorithm is used such as Random Forest classifier and agglomerative clustering algorithm. Thus the proposed model is known as an effective model for detecting fraud.

S. S. Harshini Padmanabhuni narrates if the payments are not done in time in a credit account it's because of fraud committed on credit cards. [11] Thus to detect fraud detection the proposed paper has used machine learning as algorithms such as Linear Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor algorithm, Random Forest, Neural Network, and Probabilistic Neural Network. Probabilistic Neural Network is used for classification problems and pattern recognition.

V.Jain proposes in making online payments credit cards are very commonly used. By using credit cards many fraud cases are reported in recent times. Artificial Intelligence (AI) technique and machine learning (ML) technique is used for detecting and preventing credit fraud detection. Machine learning algorithms are used on data set of credit card frauds.
[12] Machine learning algorithm such as Random Forest with the combination of Decision Tree and XGBOOST algorithms is implemented.

C.Wang narrates huge development in the field of Internet finance and credit fraud detection is also simultaneously increasing. [13] The proposed paper implements credit fraud detection by using a whale algorithm optimized BP neural network. This algorithm is implemented on the Matlab platforms. Thus the model implements successfully the WOA-BP algorithm with high accuracy detection.

III. PROPOSED METHODOLOGY

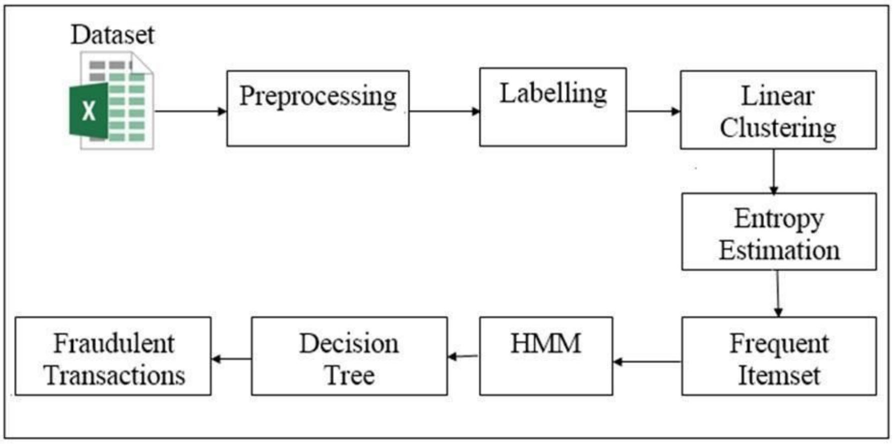


Figure 1: Credit card fraud detection System Overview

The presented approach for the purpose of enabling an effective approach for fraudulent transaction detection on credit cards has been given in the figure 1 above. The steps required for the purpose of enabling this approach have been stipulated below.

Step 1: Data Collection, Preprocessing and Labeling – The prescribed technique for the detection of fraudulent transactions have been implemented through the use of Java programming language. A credit card transactional Dataset is provided as an input to this system for the purpose of detection of fraud. The Dataset containing the credit card transactions is provided by the URL - https://www.kaggle.com/mlg-ulb/creditcardfraud

The dataset is in the workbook format with attributes such as, time taken for the transaction along with the amount of the transaction and the label for the fraudulent transaction detection. There are several other attributes that have been obfuscated to preserve the privacy of the customers. This dataset is provided to the system with the help of the JXL library. This allows the interfacing of the workbook file into the java code for the purpose of achieving the input in the form of a double dimension list.

The Dataset provided as an input to the java code is stored in the form of a double dimension list. This is read by the system and preprocessed to eliminate the irrelevant and inconsistent data from the dataset. Once the dataset is preprocessed and labeled, it is provided to the next step for the clustering of the data.

Step 2: Linear Clustering – The preprocessed and labeled dataset is provided as an input to this step of the procedure for the purpose of achieving the clustering of the data. The dataset that has been previously preprocessed hasbeen used for the clustering procedure through the use of Linear Clustering. The linear Clustering provides a very well balanced approach to improve the clusters based upon various implementations.

The Linear Clustering module first extracts the minimum and maximum amount from the input preprocessed list. Once the values of the minimum and maximum amount are achieved, it is provided for the evaluation of the boundaries of the clusters. This is done by the extraction of the mid amount by subtracting the minimum amount from the maximum amount. The resultant amount is divided by 2 and then again added to the minimum amount to get the middle amount. This amount is also referred to as the average amount which is provided to the cluster formation procedure.

There are two clusters that are achieved through the linear clustering approach. These two clusters are populated based on the mid amount achieved previously. The clusters are based on the amount of the transaction. The amount is subjected to the comparison with the mid amount. If the amount is less than the mid amount, the value is added to the cluster 1, otherwise it is added to the cluster 2. This is done to all of the entries in the dataset. The achieved clusters are provided to the next step for the purpose of entropy estimation.

This process of Linear Clustering can be depicted using the below mentioned algorithm 1.

ALGORITHM 1: Linear Clustering

```
//Input : $MIN_{AMT,}$ $MAX_{AMT}$, Preprocessed List $P_L$
//Output: $L_C$
linearCluster($MIN_{AMT,}$ $MAX_{AMT}$, $P_L$) 1: Start
2: $P_L$ =∅ , List1= ∅ , List2=∅
3:        T= ($MAX_{AMT}$ - $MIN_{AMT}$ )/ 2
4:        Mid= $MIN_{AMT}$ +T
5:    *for* i=0 to Size of $P_L$ 6:              ROW = $P_{L[i]}$
7:          $R_A$→$ROW_{[AMT]}$
8:                         *if* ($R_A$< Mid), *then*
9:                                  List1= List1+ ROW
10:                *else*
11:                        List2= List2+ ROW 12:
13: *end for*
14:      $P_L$ = $P_L$ + List1
12:      $P_L$ = $P_L$ + List2
14: return $P_L$
15: **Stop**
```

Step 3: Entropy Estimation & Frequent itemset – The information gain values of the cluster entries need to be evaluated. The clusters achieved in the previous step are used as an input to this step of the procedure. The entropy is evaluated for the data in the clusters through the use of Shannon information gain given in the equation 1 given below.

퐻 = $\frac{횤}{}$ 퐥 표 퐠$\frac{횤횥}{}$ 퐥 표 퐠^{횥}____(1)

Where

c c c

a= matched amount count (Frequent itemset) c= total unique amount count
b= c-a
E = Entropy Gain factor

The amounts are extracted for all the entries in the cluster and the number of similar transactions with the same amount is counted which is referred to as the frequent item set count. This count is then subjected to the calculation of the information gain through the Shannon information gain equation given above. The achieved values of the entropy are then stored in the form of a list and provided to the next step for the purpose of further processing.

Step 4: Hidden Markov Model – The Hidden Markov Model is utilized to achieve the proper realization of the fraudulent transactions. The entropy list achieved in the previous step is utilized in this step as an input. The entries in the entropy list are utilized to achieve the time of the transaction for that particular transaction. Once the time is extracted, it is subjected to comparison with other transactions. If the time matches with another transaction, then it is counted and the process is repeated for every entry in the data.

Two temporary lists are created, one of which contains the entries with only one matching time and the other list contains entries with more than one matching time. These lists are used to create the time lists and the duplicates are eliminated. Two double dimensional arrays are used to capture the relevant data related to these entries and then provided to the HMM probability value generation.

The probability values are measured by the HMM module for both the arrays and the resultant values are then compared with one another. If the probability scores are more than the other list, then the values are added. This procedure is enacted to eliminate the similar values and achieve a HMM list with unique values.

Step 5: Decision Tree – This is final step of the procedure where the output from the Hidden Markov Model given above is provided as an input. The Decision Tree approach is a highly suitable approach for the classification of the output precisely. The use of the if-then rules facilitates the segregation of the output to achieve the effective output required for the analysis of fradulent transactions.

The output from the HMM returns a prpobability value for the presence of phishing contents or not. The values are provided as an input which subject it to certain rules for the detection. The probability vlaues are effectivey classified

The presented approach for the detection of credit card based on the proabaility scores. The only relevant scores are between the range of 1 and 99.90 based on the amount. This output of the detected fraudulent transactions is provided to the user thorguh the Graphical User Interfcae.

V. RESULTS AND DISCUSSIONS

Fraud through the use of Hidden Markov Model and Decision Tree has been implemented using the Java Programming language. The NetBeans IDE has been used to develop the presented approach. The development approach is implemented through the use of a machine powered by an Intel Core i5 processor with 500GB of storage and 4GB of RAM.

The fraudulent transaction detection is done through the use of an extensive dataset consisting of transactions which is given as an input to this system. The credit card fraud detection system needs to be evaluated for its accuracy to understand the efficacy of the approach. The evaluation is performed through the use of experimentation as described below.

Performance Evaluation through Root Mean Square Approach

The error achieved by the presented methodology is one of the most effective approaches to achieve the credit card fraud detection technique's performance metrics. Intensive experimentation and evaluation of the HMM detection methodology for its usefulness in fraudulent transaction detection is used to obtain the error.

The RMSE, or Root Mean Square Error, has been used to calculate the detection error. The error is determined in order to assess whether or not there is fraud in the input credit card transactions. The RMSE methodology is particularly accurate when it comes to calculating the error. The evaluation could suggest that the Hidden Markov Model that was utilized for the identification was properly implemented or not. The error is calculated between two entities that are inextricably linked. Our methodology for measuring error uses two variables: inaccurate detection of credit card fraud and accurate detection of credit card fraud. The following equation 2 is used to calculate the RMSE value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^{2}}{n}} \_\ (2)$$

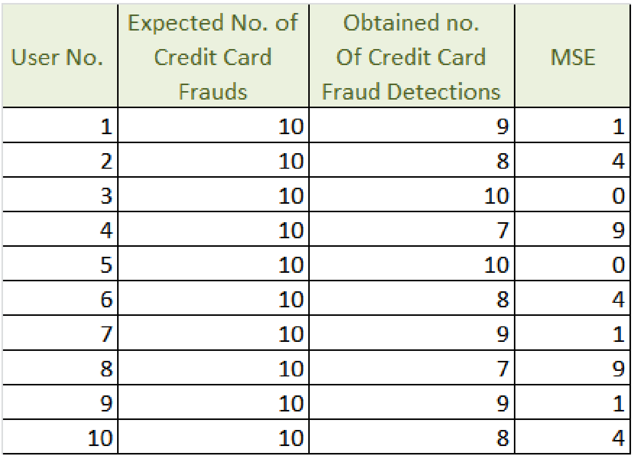
Where,

$\sum$ - Summation

$(x_1 - x_2)^2$ - Differences Squared for the summation in between the expected No. of credit card fraud detections and the obtained No of credit card fraud detections

n - Number of Trail.

The extensive experimentation has been performed for a number of trails of input data to the system. The results of the experimentation are used to make the table displayed by table 1 below.



User No.	Expected No. of Credit Card Frauds	Obtained no. Of Credit Card Fraud Detections	MSE
1	10	9	1
2	10	8	4
3	10	10	0
4	10	7	9
5	10	10	0
6	10	8	4
7	10	9	1
8	10	7	9
9	10	9	1
10	10	8	4

Table 1: Mean Square Error measurement

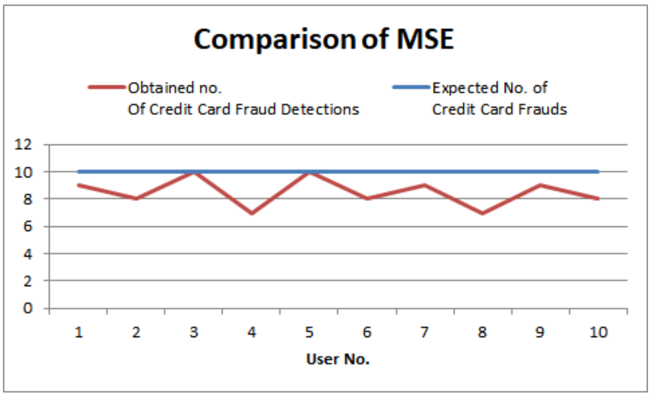


Figure 3: Comparison of MSE in between Expected No of credit card frauds identified V/s Obtained No of credit card frauds identified

The accomplished outcomes for the RMSE have been used to plot the graph given in the figure 3 above. The values of MSE and RMSE reached by the system are 3.3 and 1.81 respectively. These values are expected as the Hidden Markov Model has been accurately applied to achieve highly precise results. The low error rate can be accredited to the implemented Decision Tree approach that can expressively reduce the errors by elimination of the improper results.

VI. CONCLUSION AND FUTURE SCOPE

The proposed system for detection of credit card fraud has been elaborated in detail in this research paper. There is a need for a technique that can accurately detect credit card fraud by examining the several parameters and making a well-informed choice. These credit card transactions that are fraudulent in nature, pose a significant risk to both the credit card holder and the financial institution that provides credit services. Therefore this methodology takes the input data set containing credit card transactions as an input. The data set attributes are effectively extracted and processed to remove the redundant and incomplete data. This preprocess data set is then subjected to the process of labeling through which most of the attributes are converted into appropriate integer format. The preprocessed and labeled data set is then provided for clustering through the use of linear clustering module. The clusters achieved through the Linear clustering are segregated and provided to the entropy estimation for evaluation of the information gain value. Through the information gain values of frequent itemset is generated and provided to the hidden Markov model. The hidden Markov model performs the hidden layer estimations and achieve probability values that are effectively e classified through the use of the decision tree classification protocol. Thorough experimentation has been performed to extract the error achieved by the prescribed system which has been acceptable levels of precision.

The future research direction can be focused on deploying this approach on the cloud platform for easier integration and implementation and it can be improved to work on real time transactions.

REFERENCES

[1] I. Benchaji, S. Douzi and B. ElOuahidi, "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection," 2018 2nd Cyber Security in Networking Conference (CSNet), 2018, pp. 1-5, doi: 10.1109/CSNET.2018.8602972.

[2] K. T. Hafiz, S. Aghili and P. Zavarsky, "The use of predictive analytics technology to detect credit card fraud in Canada," 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), 2016, pp. 1-6, doi: 10.1109/CISTI.2016.7521522.

[3]I. Benchaji, S. Douzi and B. ElOuahidi, "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection," 2018 2nd Cyber Security in Networking Conference (CSNet), 2018, pp. 1-5, doi: 10.1109/CSNET.2018.8602972.

[4] F. Ghobadi and M. Rohani, "Cost sensitive modeling of credit card fraud using neural network strategy," 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), 2016, pp. 1-5, doi: 10.1109/ICSPIS.2016.7869880.

[5]F. Z. El hlouli, J. Riffi, M. A. Mahraz, A. El Yahyaouy and H. Tairi, "Credit Card Fraud Detection Based on Multilayer Perceptron and Extreme Learning Machine Architectures," 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020, pp. 1-5, doi: 10.1109/ISCV49265.2020.9204185.

[6] A. A. Khine and H. W. Khin, "Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree," 2020 IEEE Conference on Computer Applications (ICCA), 2020, 10.1109/1CCA49400.2020.9022843. PP. 1-4, doi:

[7] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," 2017 International Conference on Intelligent Computing and Control (12C2), 2017, pp. 1-5, doi: 10.1109/12C2.2017.8321781.

[8] S. Khatri, A. Arora and A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," 2020 10th International Conference on Cloud Computing, Data Engineering (Confluence), 2020, pp. 10.1109/Confluence47617.2020.9057851. Science & 680-683, doi:

[9] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 488-493, doi: 10.1109/CONFLUENCE.2019.8776942.

[10] Y. Lucas et al., "Dataset Shift Quantification for Credit Card Fraud Detection," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 97-100, doi: 10.1109/AIKE

2019.00024.

[11] S. S. H. Padmanabhuni, A. S. Kandukuri, D. Prusti and S. K. Rath, "Detecting Default Payment Fraud in Credit Cards," 2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), 2019, pp. 15-153, doi: 10.1109/ICISGT44072.2019.00018.

[12] V. Jain, M. Agrawal and A. Kumar, "Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 86-88, doi: 10.1109/ICRITO48877.2020.9197762.

[13] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai and S. Pan, "Credit Card Fraud Detection Based on Whale Algorithm Optimized BP Neural Network." 2018 13th International Conference on Computer Science & Education (ICCSE), 2018, pp. 1-4. doi: 10.1109/ICCSE.2018.8468855.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING