# An implementation of decision tree algorithm augmented with regression analysis for fraud detection in credit card

Hammed, Mudasiru[1] and Soyemi, Jumoke[2]
[1&2]Department of Computer Science, Federal Polytechnic, Ilaro, Ogun State
(mudasiru.hammed@federalpolyilaro.edu.ng; jumoke.soyemi@federalpolyilaro.edu.ng)
Corresponding Author : jumoke.soyemi@federalpolyilaro.edu.ng

***Abstract* - Credit card fraud is a common crime committed with the aid of credit card to defraud people of their funds during transaction. The intension of fraudsters most times is to obtain goods without payment or to get funds from unauthorized accounts. Of recent, the use of credit card to complete transactions is on the rise because of the introduction of online shopping and banking. Although, there are several fraud detection systems proposed by previous studies, yet, credit card fraud is still on the increase. Some techniques used in detecting credit card fraud have been compromised due to improvement in technology. Many studies that used decision tree to build detector system did not use regression analysis. This study presented a decision tree algorithm augmented with regression analysis to build a very strong fraud detector system. The system covers all areas in terms of monitoring and reporting fraud in credit cards. The analysis of result here shows that this technique is 81.6% accurate with 18.4% misclassification error and the system successfully verified all the injected intrusions used for the purpose of testing.**

**Key words: credit card fraud, machine learning, data mining techniques, decision tree algorithm, regression analysis.**

## I. INTRODUCTION

The increase in the use of credit card is facilitated by technological changes especially the internet [1]. Credit card is produced in form of plastic issued out by financial institutions such as banks for the purpose of making transactions very easy and convenient without necessarily holding cash, which makes it a good alternative for cash payment. Also, it could be used for transactions based on electronic devices such as card swapping machine, computer with internet facility and others [2]. This credit card has recently become very attractive to fraudsters with changes in their activities during the last few decades as a result of increase in technological development [3]. Credit card fraud takes place whenever an unauthorized user gain access to credit card or information contained in the card without permission from the card owner [4]. There are different types of fraud including credit card frauds, telecommunication frauds, computer intrusions, Bankruptcy fraud, Theft fraud/counterfeit fraud, Application fraud and Behavioral fraud [5]. However, there is the urgent need prevent businesses and financial institutions from credit card frauds experienced often nowadays. Credit card fraud detection systems have been proposed and implemented by several studies. The detection systems are classified into two general categories: misuse detection and anomaly detection. The fraud detection systems for both misuse and anomaly detections include; Neural Networks, Artificial Immune system, Support Vector Machine and Genetic Algorithm. This study proposed Decision Tree algorithm augmented with regression analysis to implement an effective system for credit card fraud detection. The Decision Tree technique according to [6], outperformed such as Support Vector Machine approach in detecting credit card frauds.

## II. RELATED WORK

There are different methods proposed by previous studies for all classes of fraud detection systems. Some of them are presented in this work.

Study [6], proposed an optimization technique and evolutionary search based on the genetic and natural selection for credit cards fraud detection system. The system proves accurate to find out the fraudulent transactions and minimizing the number of false alerts in credit cards. Although Genetic algorithm is appropriate for detecting or predicting the fraud in a very short span of time after the transactions has been made. However,


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

there is need for extensive knowledge tools to set up and operate the system and this may render the system useless. Even though, to select a rule with highest prediction is a difficult task.

Study [7], proposed Hidden Markov Model which is fast in fraud detection. It maintains a log for transactions which reduces tedious work of employee. The implementation of this approach, however, was not flexible because customer's information was treated as a fraud when detected whereas it was not. This study predicted some customers as legitimate when they are fraudulent. The accuracy of Hidden Markov Model is low and not scalable to large size data sets.

An improved competitive learning neural network for network intrusion and fraud detection was proposed by [8], the system proposed two clustering algorithms for fraud detection and network intrusion detection. the system however was not applied to credit card fraud detection.

Study by [9] proposed a method that used a support vector machine (SVM) trained with data from Questionnaire Responded Transaction data of users (QRT). The QRT models were used to predict new transactions. A personalized approach for credit card fraud detection that employs both SVM and ANN was proposed by [10]. The systems tried to prevent fraud for users even without any transaction data. However, these systems are not fully automated and depended on the user's expertise level.

Study [11], worked on Support Vector Machine for detecting fraudulent credit card transactions through Decision Support System. This method delivered an optimal solution because an optimal hyper-plane that separated legitimate and fraudulent credit cards was appropriately located. The system training sample were however biased and could not also handle large dataset.

Study [4] looked at decision tree and support vector machine to detect fraud in credit cards. The approach demonstrated that decision tree and support vector machine reduce the bank's risk in detecting credit card fraud and decision tree approach outperforms support vector machine approach

In this study, the proposed system used decision tree algorithm augmented with regression analysis to build a very strong fraud detector system which is meant to cover monitoring and reporting fraud in credit cards. With Decision tree, large datasets are separated into many simple ones to resolve the sub problems, this make the system to be very easier and flexible

## III. MATERIALS AND METHODS

The fraud detection system checks card details which includes credit card number, card names, the sex of card's owner, card type, expiry date and the amount of money to validate. The system used information to determine whether the transaction is genuine or not. The implementation of decision tree algorithm to detect fraud transaction in credit cards establish the relationship among input training set and identify the spending history of the owner. The system stores data of different amount spent in each transaction and categorizes them in ranges of low, medium or high values and find any variance in the transaction based on the amount spent.

The system pseudo-code for detecting fraud;

*Step 1: The user input the details of the credit card.*

*Step 2: The user specifies the amount of transaction.*

*Step 3: The Design application model verify the pattern of transaction against the previous transaction record, for regularity in pattern of transaction.*

*Step 4: IF the transaction matches the previous transaction THEN activate the user.*

*Step 5: ELSE, alert the bank for fraudulent activities.*

*Step 6: STOP transaction.*


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 1 is the architecture of the credit card where a user inserts his/her card to make transaction before the user will be authenticated, the user request will first go to the bank (issuer of the card). The bank will also send the user request to the credit card management (the unit managing the credit card in financial institutions). This credit card manager is the one using decision tree algorithm to confirm the authenticity of each credit card. It confirms the details on the credit card whether it corresponds to the one issued for the user and whether the user is the right owner of the card. After which the confirmation whether the user should be activated or deactivated will be sent to the bank and the bank will now send the result of the request to the user. The figure 2 shows the flow of information in the system.

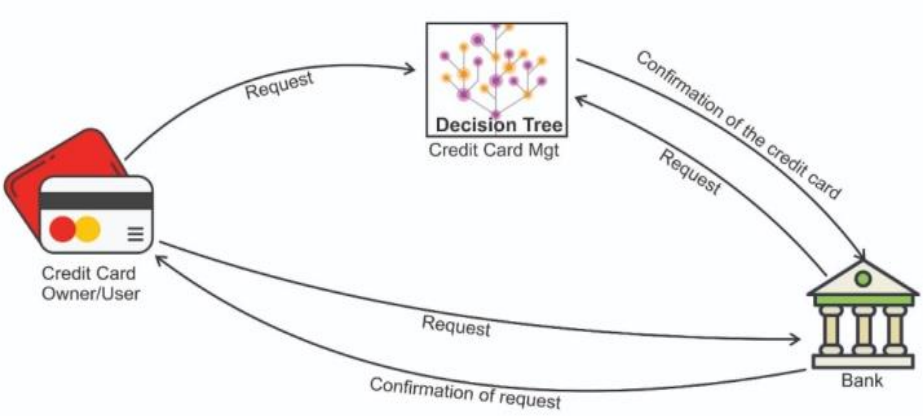
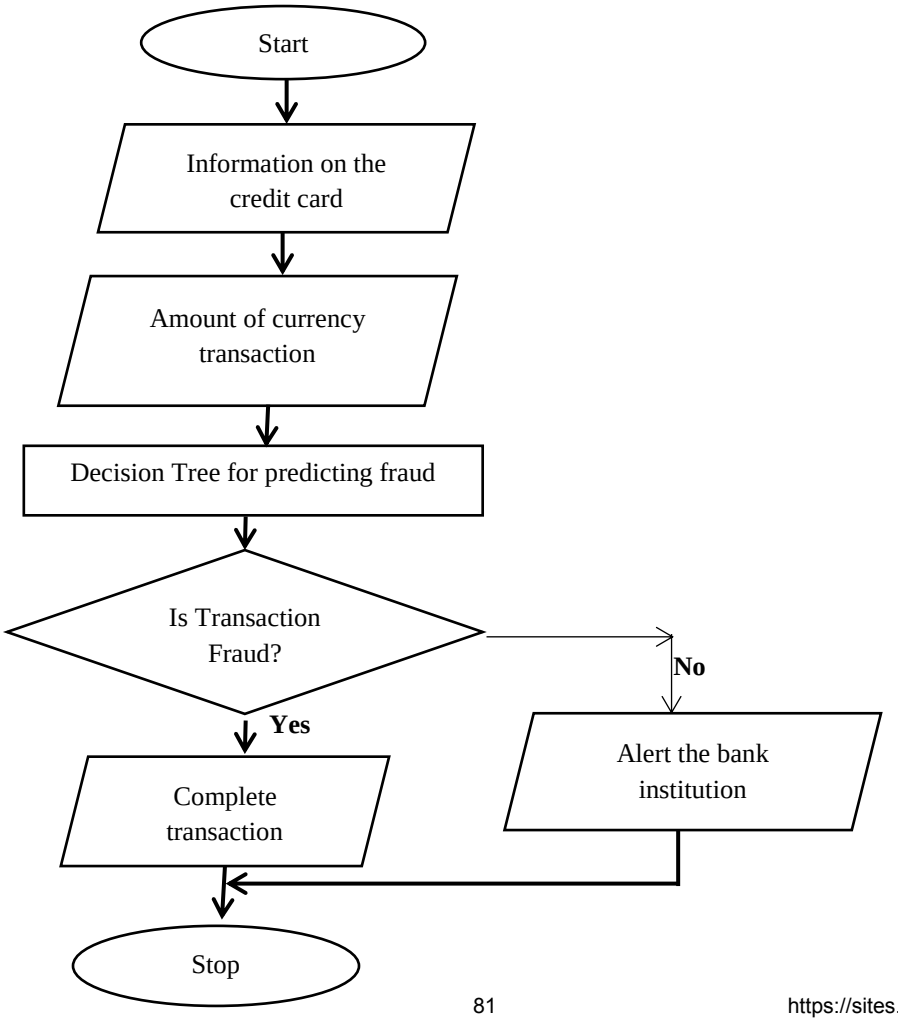



**Fig 1:** Architecture for credit card detection system.





https://sites.google.com/site/ijcsis/
ISSN 1947-5500

**Fig 2:** Diagram depicts flow of information in the system

The data used in the study consist of 17 features of dataset which were refined for training phase before the modelling phase. Behavioral prediction of a certain dependent variable was done using multiple regression analysis as it is shown in equation 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e \quad (1)$$

The equation 1 also describes relationship in the population making inferences about the relationship for the larger population. The study used the equation 2 to model the error that may occur as a result of over and under-predictions.

$$e_i = y_i - \hat{y}_i \quad (2)$$

## IV. RESULTS AND DISCUSSION

MATLAB was used for modelling the classifier proposed in this study The Fitch tree function in MATLAB takes features variable matrix X and target vector Y and other name-value argument to return a tree structure. Figure 3 shows log of activities in MATLAB that generated rules and tree graph.

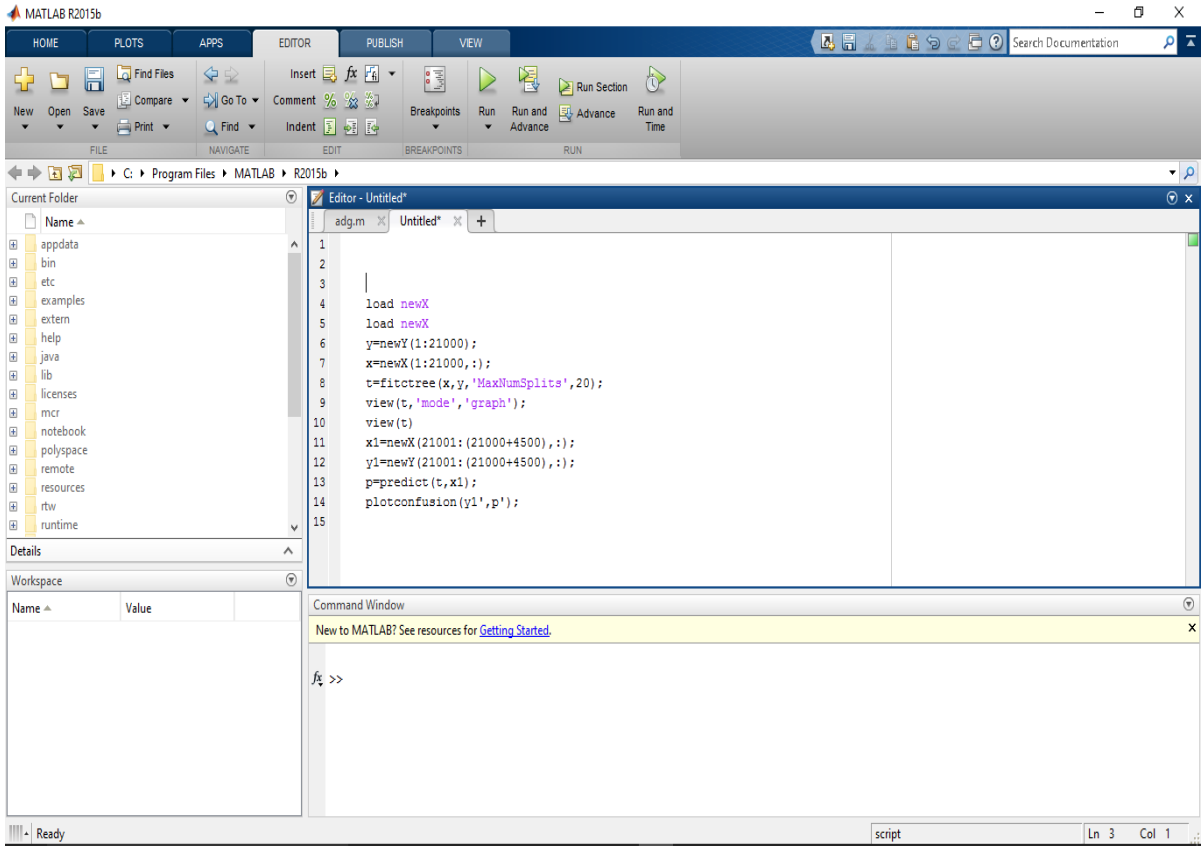



**Fig 3:** Generation of rules and tree graph in MATLAB



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

After the tree has been trained the rule of the tree was extracted using the MATLAB function view (tree). With the rule of the tree a graphical user interface was generated which was designed using JAVA as it is shown in figure 4. The generated decision tree rules show the operations of the decision tree.

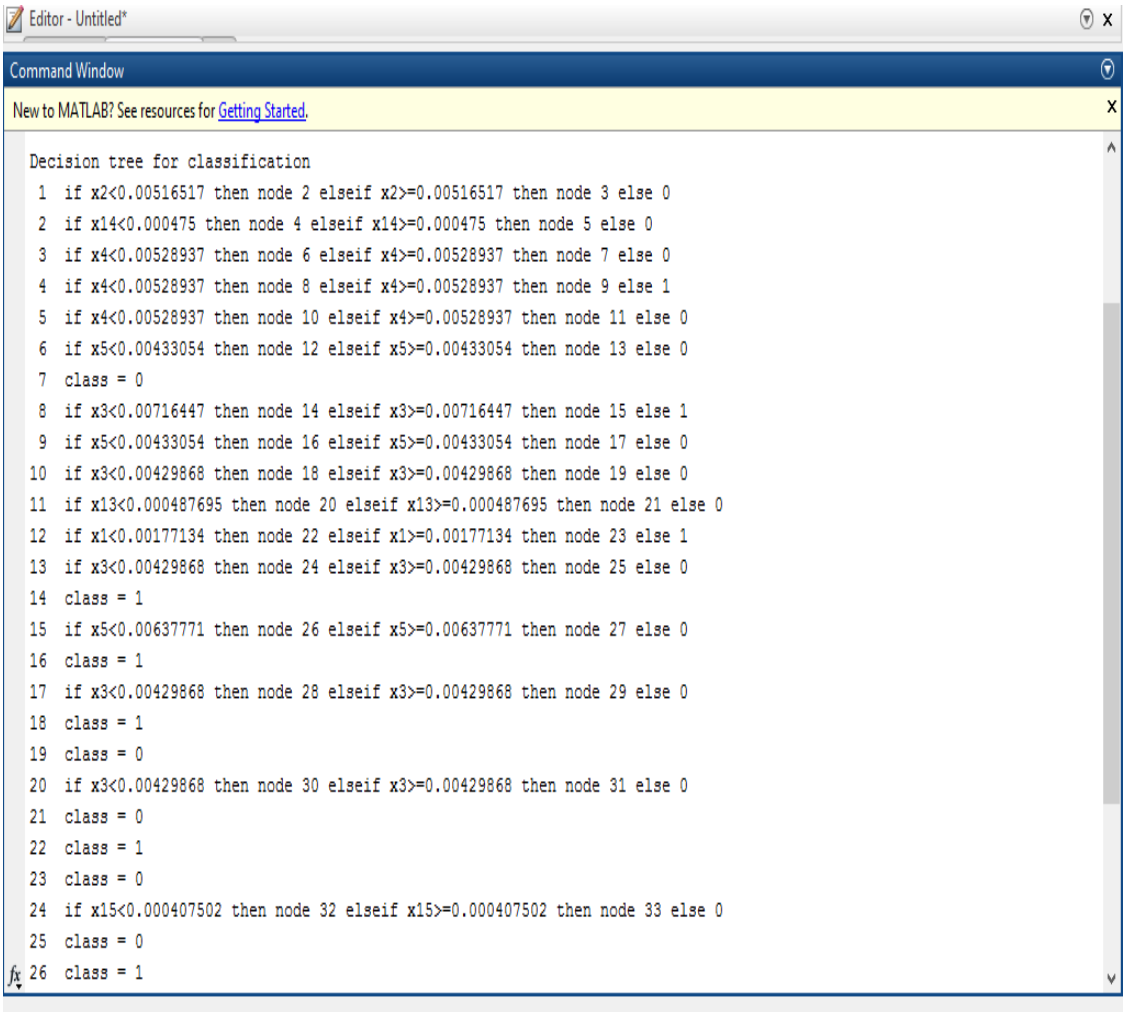



**Fig 4:** rules for detecting fraud

The attributes of credit card dataset transactions were collected, and the dataset were divided into subsets using the attributes. Decision tree nodes containing decision attributes were used to create decision tree using MATLAB shown in figure 5. The diagram in figure 6 however, shows the decision tree for detecting fraudulent activities in credit card.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

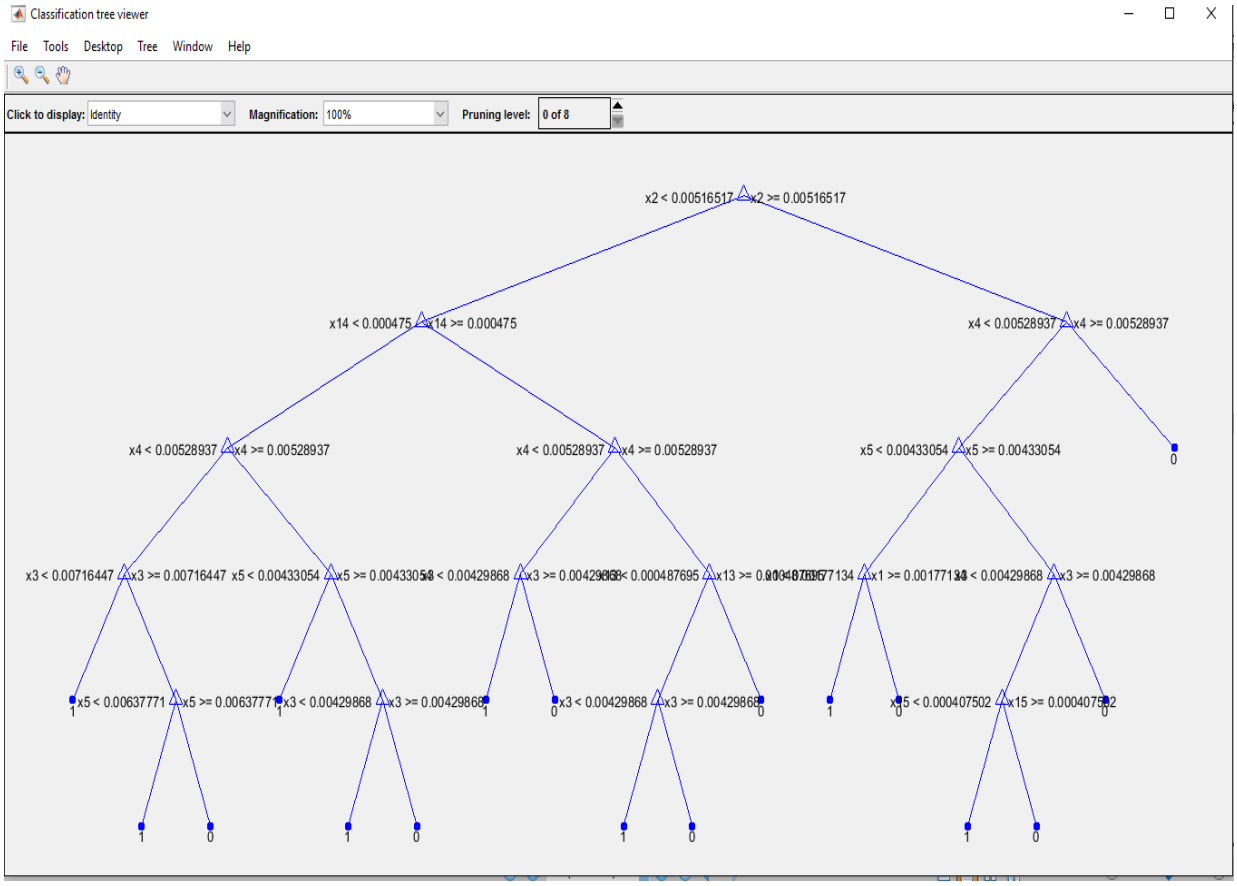



**Fig 5:** Decision Tree for detecting fraudulent activities in credit card

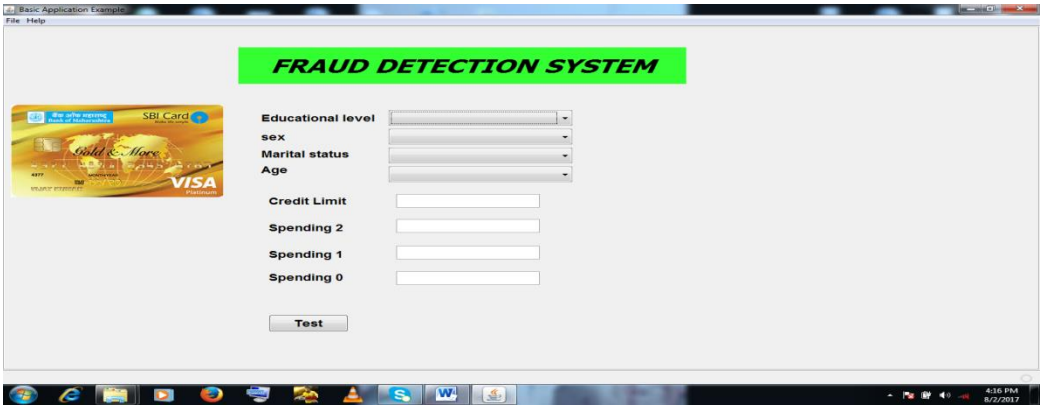



**Fig 6:** Diagram depicts system detection of Fraud.

A secondary dataset of credit card transactions was downloaded from UCI machine learning repository which contains cases of fraudulent and non-fraudulent transactions with their corresponding transaction details. A dataset of 17 features was collected, trained, and normalized to get the best 8 features of dataset that were used to model the detection system, which is shown in table 1and 2 respectively.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

**Table 1:** shows the collected and trained data:

| S/N | INPUT VARIABLE | DOMAIN | NODE SYMBOL |
|---|---|---|---|
| 1 | Credit limit | Numeric | X1 |
| 2 | Sex | Nominal data<br>1= male<br>2= female | X2 |
| 3 | Education | Nominal data<br>1= graduate<br>2= postgraduate<br>3= secondary school<br>4= others | X3 |
| 4 | Marital status | Nominal data<br>1= married<br>2= single<br>3= other | X4 |
| 5 | Age | Numeric | X5 |
| 6 | Spending 1 | Numeric | X13 |
| 7 | Spending 2 | Numeric | X14 |
| 8 | Spending 3 | Numeric | X15 |
| 9 | Spending's | Numeric | X6 |
| 10 | Spending's | Numeric | X7 |
| 11 | Spending's | Numeric | X8 |
| 12 | Spending's | Numeric | X9 |
| 13 | Spending's | Numeric | X10 |
| 14 | Spending's | Numeric | X11 |
| 15 | Spending's | Numeric | X12 |
| 16 | Spending's | Numeric | X16 |
| 17 | Spending's | Numeric | X17 |

The dataset was normalized in such a way that each feature (column) has values between zero (0) and one (1). This was achieved using equation 3.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

**Table 2:** The best 8 dataset used in modelling the detection system

| S/N | INPUT VARIABLE | DOMAIN | NODE SYMBOL |
|---|---|---|---|
| 1 | Credit limit | Numeric | X1 |
| 2 | Sex | Nominal data<br>1= male<br>2= female | X2 |
| 3 | Education | Nominal data<br>1= graduate<br>2= post graduate<br>3= secondary school<br>4= others | X3 |
| 4 | Marital Status | Nominal data<br>1= married<br>2= single<br>3= other | X4 |
| 5 | Age | Numeric | X5 |
| 6 | Spending 1 | Numeric | X13 |
| 7 | Spending 2 | Numeric | X14 |
| 8 | Spending 3 | Numeric | X15 |

The table 3 shows the output data which are classified based on being fraudulent or non-fraudulent transaction. A zero (0) value as output means non-fraudulent transactions while a one (1) as output means a fraudulent transaction.

**Table 3:** The Table summarizes the input and output transformation.

| OUTPUT | CLASS |
|---|---|
| 0 | Non-fraudulent |
| 1 | Fraudulent |

A set of 17 features with 21,000 observations were used during the training process of the decision tree. During the training, a maximum split of 20 was used to avoid overfitting of the dataset. Although 17 features were employed but from the tree graph of the trained model, only 8 features were automatically selected and the most important used in growing the tree. After the training, a total of 4,500 observations reserved were used to test the model. This testing was achieved by calling the MATLAB function predicts (Tree, X), where "Tree" is the trained model and X is the matrix of the testing input data (4,500 reserved observations without their target). The predict function returns a column matrix (P) of 4,500 rows with the entry of each row being 0 or 1. If the value or any row of P is 0 then the corresponding row in the test date X is predicted to be non-fraudulent otherwise it is fraudulent. In other to gain more insight of prediction accuracy of the model, the predicted matrix (P) and actual target Y of the test data was used to construct a confusion matrix as show in figure 4.2. This confusion matrix shows that the model has 81.6% accuracy while making misclassification error of 18.4%. This result shows that a high accuracy was reached using the proposed model. The diagram in figure 7 shows the confusion matrix to test for system accuracy.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

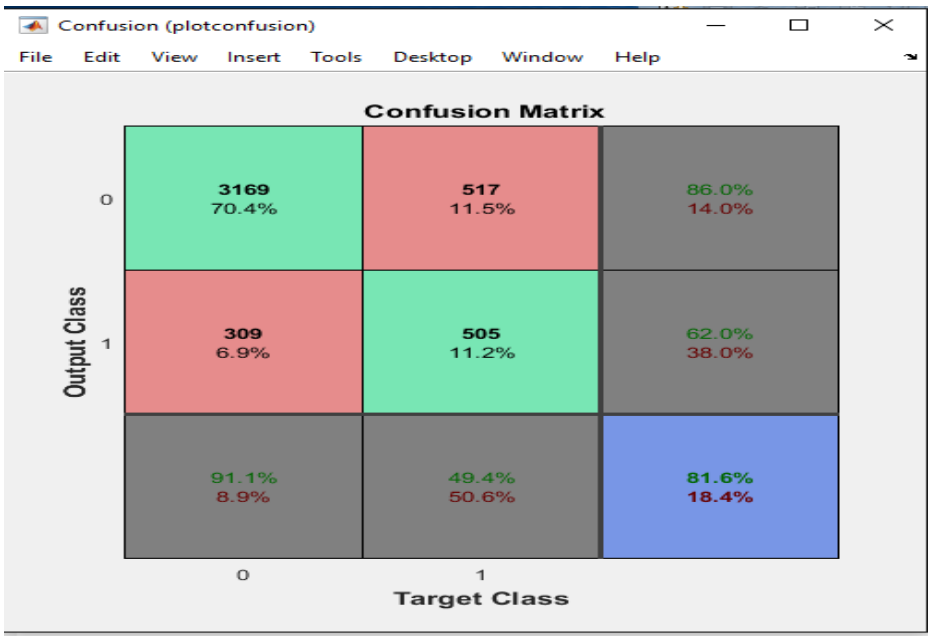



**Fig 7:** Confusion Matrix to test system accuracy.

## V. CONCLUSION

This study was able to achieve its aim by building a machine learning model that is capable of detecting fraud in credit card transactions. The decision tree algorithm was implemented with regression analysis for proper classification of fraudulent transaction. The result of this study shows improved performance.

## Conflict of Interest

The authors declare that there is no conflict of interest

## REFERENCES

[1] R. Anderson, R. "The Credit Scoring Toolkit: theory and practice for retail credit risk management and decision automation", New York: Oxford University Press, 2007.

[2] S. Renu and S. Suman, "Analysis on Credit Card Fraud Detection Methods". International Journal of Computer Trends and Technology (IJCTT), vol.8, no.1, pp. 45 – 51, 2014.

[3] V. Dheepa, R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods", International Journal of Recent Trends in Engineering, Vol. 2, no. 3, pp 126 – 128, 2019.

[4] Y. Sahin, E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines", Proceeding of International Multi-Conference of Engineering and Computer Statistics, Vol. 1, 2011.

[5] N. Sivakumar and R. Balasubramanian, "Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques". International Journal of Computer Science and Information Technologies, Vol. 6, no. 2, pp. 1379-1386, 2015.

[6] T. I. Monika, and M. Mrigya, "Credit Card Fraud Detection". International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, no. 1, pp 39 – 42, 2016.

.

[7] V. Bhusari and S. Patil, "Study of Hidden Markov Model in Credit Card Fraudulent Detection". International Journal of Computer Applications, Vol. 20, no. 5, pp 33 – 36, 2011.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[8] J. Zhing and A.A. Ghorbani, "Improved competitive learning neural network for network intrusion and fraud detection", Neurocomputing, vol.75, no. 1, pp. 135-145, 2012.

[9] R.C. Chen, M.L. Chiu, Y.L. Huang and L.T. Chen, "Detecting Credit Card Fraud by Using Questionnaire Responded Transaction Model Based On Support Vector Machines". *Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning, vol. 3177, October 2004, Pp. 800– 806, 2014.*

[10] R.C. Chen, S.T Luo, X. Liang, and V.C. Lee, Personalized Approach Based on SVM And ANN For Detecting Credit Card Fraud. Proceedings of the IEEE International Conference on Neural Networks and Brain, October 2005, Pp.810–8152015.

[11] B. Siddhartha, K.T. SanjeevJha, and W.J. Christopher, Data mining for credit card fraud: A comparative study. *Elsevier, Decision Support Systems, Vol. 50 Pp. 602–613, 2011).*