https://doi.org/ 10.33472/AFJBS.6.7.2024.232-239

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

PREVENTION OF CREDIT CARD FRAUD TRANSACTION USING GA FEATURE SELECTION WITH MACHINE LEARNING

[1]Sindhuri Suseela Mantena, [2]T. Venkateswara Rao, [3]Battula Ravindra Chanti Babu, [4]T. Prasad, [5]D. Naresh

[1]Assistant Professor, Department of Computer Science and Engineering, SRKR Engineering College (A), SRKR Marg, China Amiram, Bhimavaram, Andhra Pradesh, India

[2, 4]Assistant Professor, Department of Computer Science & Engineering, Sree Vahini Institute of Science and Technology (A),Tiruvuru, N.T.R District, Andhra Pradesh, India

[3]Assistant professor, Department of information technology, Lakireddy Bali Reddy college of engineering ,L.B Reddy Nagar, Mylavaram, N.T.R District, Andhra Pradesh, India

[5] Assistant professor, Department of Electronics and communication engineering, Lakireddy Bali Reddy college of engineering ,L.B Reddy Nagar, Mylavaram,N.T.R District, Andhra Pradesh, India

Article History
Volume 6,Issue 7, 2024
Received: 24 Feb 2024
Accepted : 27 Mar 2024
doi: 10.33472/AF5BS.6.7.2024.232-239

ABSTRACT: Financial fraud in credit card transactions is the use of a card without authorization across a number of platforms in an effort to gain an unfair advantage. Since this kind of fraud results in financial loss and security flaws, establishing an efficient detection method is crucial for solving the problem. Credit Card Fraud (CCF) has grown to be a serious issue for both cardholders and the companies that accept credit cards over time. Application level frauds and transaction level frauds are the two levels on which CCF operates. This analysis uses the Genetic Algorithm (GA) as a feature selection technique, concentrating on the application level of CCF identification. There are 2 phases in this.In the first phase, eight (8) traits were selected as the best match features. The selection and consideration of an additional set of eight (8) traits, also known as the second priority characteristics was required in the second stage. On the credit card dataset which is an imbalance dataset, the supervised machine learning methods DT,K-NN, and SVM were used to identify CCF. The first priority features are the most crucial elements.According to the experimental results of the framework, the results also demonstrated that the KNN algorithm surpassed DT and SVM in terms of accuracy, sensitivity, and precision.

KEYWORDS: Credit card, Imbalance dataset, Genetic algorithm, Fraud Detection, Decision Tree (DT), K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM)

I.INTRODUCTION

Financial transaction fraud is discovered by labelling a transaction connected to a user as an outlier because it deviates from the user's typical behaviour. To increase the business value of online transactions, the detection process is frequently done in real-time [13]. Transaction data is typically saved on an operational data store, and frauds are identified after a claim has been decided upon by examining the transaction data. Post-adjudication fraud discovery frequently has a significant negative impact on the value of the firm due to the expense of collection, significance threshold of the amount involved, and likelihood of successful collection. Financial institutions can track fraudulent

actions and take the necessary corrective action in real time by analysing historical fraud tendencies.

Credit Card Fraud (CCF) is a problem that has many implications for businesses, banks, and regular customers. Financial fraud has increased over the past decade as a result of the development of contemporary technology like mobile computing and the internet [14] For instance, an account holder's credit card information could be stolen by a criminal who would then use it to carry out fraudulent transactions. The activities have the ability to play a role in how illegal organizations or groupings, such terrorist groups and drug cartels, manage their finances.

It is impossible to go a day without encountering a bank, whether an online platform or a physical transaction, as banking and finance are essential sectors in our day-to-day operations. The banking information system has greatly improved both the private and public sectors' efficiency and viability. Due to the widespread usage of e-commerce, internet technology, online banking, and improvements in mobile intelligent devices, particularly online transaction operations carried out through web payment gateways like Alipay, PayPal, and others. credit cards are widely accepted as a form of payment. As credit card transactions become the preferred mode of payment for both online and offline purchases, Credit Card Fraud (CCF) rates are rising alarmingly quickly [1].

Financial fraud has grown to be a serious concern with wide-ranging implications for individuals, organizations, the government, and the finance industry. An illegal or criminal deception intended to produce financial or personal gain is referred to as fraud. CCF is concerned about the unauthorized use of credit card information for transactions. Both offline and online credit card transactions are possible [8]. Credit cards are used physically or by scanning them with a device during physical transactions; Cardholders often provide their card details, expiry dates, and card security number over the phone or online when making digital purchases. CCF has not been able to stall effectively despite the numerous permission strategies in place. Two methods are frequently used to avoid loss from fraudsters : fraud identification and avoidance. Monitoring cardholder transaction behaviour with the intention of identifying whether an incoming transaction is coming from the cardholder or from criminals is known as fraud detection. While fraud prevention is a defensive strategy that aims to prevent fraud transactions occurrence in the first instance. The two main types of fraud detection are anomaly detection and Misuse detection. Misuse detection employs categorization techniques to determine whether or not an incoming transaction is fraudulent. This kind of technique typically uses a model to learn about the numerous fraud tendencies already in existence. When an incoming transaction deviates from the typical transaction pattern, anomaly detection determines whether it is a potential fraud by building a historical transaction model for the card holder's typical transaction behaviour. An anomaly detection algorithm, which requires enough successful training data to model a card holder's typical transaction pattern [6].

With the development of big data, manual techniques have become more impractical because they require a lot of time and are ineffective for detecting fraudulent activities. However, financial institutions have increasingly been driven to use computational tools for control and prevention of CCF challenges. Additionally, the number of users and online transactions has increased, placing substantial workloads on these systems.

One of the main methods used to prevent and identify CCF problems is the data mining strategy. According to CCF,The method of dividing transactions into two categories—genuine and fraudulent transactions . SVM, DT, and K-NN performance was analysed using CFF data in while testing for capabilities of SVM, DT, and K-NN.

II. LITERATURE SURVEY

Awoyemi J. O., Adetunmbi A. O., Oluwadare S. A., et al [9] analyzed how NB, KNN, and logistic regression performed in terms of detecting fraud in highly skewed credit card data. The trials employed a dataset of 284,807 transactions from European cardholders. To pre-process the skewed data, they used a hybrid technique that included under-sampling and over-sampling. The findings demonstrated that KNN outperforms other techniques as evidenced by the accuracy levels attained by KNN, NB, and logistic regression classifiers.

For Fraud Miner, Hegazy M., Madian A., Ragaie M., et al., [12] suggested the Enhanced LINGO clustering technique. This advancement replaces the Fraud Miner's Apriori method with the construction of Frequently Patterns using the LINGO clustering algorithm, which summarizes the customer's profile based on whether his transactions were genuine or fraudulent. Their simulated test transactions' findings demonstrated that LINGO created significant summary patterns more effectively than the Apriori Algorithm's output.

Loo C. H. U. K., Randhawa K., Member S., et al. [5] used hybrid AdaBoost and most voting procedures to assess the effectiveness of model. It was contrasted with an actual credit-card dataset collected from a commercial bank with a publicly accessible credit card dataset. The majority vote method produced the best MCC score.

The performance of two kinds of random forest models on a real-world B2C credit card dataset transactions was studied by Liu G., Xuan, Zheng L., Li Z., Wang S., and Jiang C., et al. [7]. Random-tree-based random forest (I) and CART-based random forest(II) are the models. According to findings, Random Forest I produced accuracy whereas Random Forest II produced excellent accuracy.

Condensed Nearest Neighbor (CNN) algorithm was proposed by Priyadarshini Y. I., Vardhani P. R, Narasimhulu Y., et al. [2] as a potential framework for an unique data mining technique. A nonparametric classification technique called CNN seeks to build a condensed set while maintaining the samples that are crucial for making decisions. With fewer comparison characteristics, the training set will be more compact, use fewer queries and memory resources.

Twelve standard models and hybrid methods that incorporate AdaBoost and majority voting techniques were used by Randhawa K., Loo C. K., Seera M., Lim C. P., and Nandi A. K. et al. [3] to increase the accuracy rates of credit card fraud detection. They were judged based on benchmark data as well as actual world data. The methods' benefits and drawbacks were briefly explored. The performance metric was the Matthews Correlation Coefficient (MCC). Data was submitted to noise to assess the algorithms' robustness. Additionally, they have shown that the extra noise has no impact on the majority voting process.

To resolve the target leakage problem that existed in GB algorithms, Gusev G., Vorobev A., Dorogush A. V., Prokhorenkova L., and Gulin A. et al. [4] developed the Catboost algorithm. Even if LightGBM and Catboost demonstrated their effectiveness in addressing the GBM's current limitations, XGBoost outperforms with important features. To

address class imbalance, this used imbalance XGBoost with integrated weighted and targeted loss.

To improve their ability to recognize CCF, Jha S., Tharakunnel K.,Bhattacharyya S., and Westland J. C., et al. [15] compared various ML classifiers. There are a number of issues with credit card detection, some of them include: the dynamic nature of fraudulent behaviour patterns, whereby fraudulent transactions can occasionally look legitimate; the accessibility and extreme unbalance of credit card transaction datasets; the optimum feature extraction for such models; and the use of the proper performance evaluation metric on skewed CCF data.

III. PREVENTION OF CREDIT CARD FRAUD TRANSACTION USING GA

The block diagram of elimination of credit card fraud transactions utilizing GA is presented in Fig.1.

There are three sections to this study presentation. The dataset was initially gathered from the UCI repository. Second, redundancy in the dataset was eliminated through pre-processing the data. Three ML algorithms are being used in the third stage of CCF detection. This research endeavour aims to enhance the ML model that selects features for CCF using GA as an attribute selection approach. On a credit dataset, the performance of the GA is assessed using the KNN, DT, and SVM.

In this learning, the credit dataset was used to categorize the transactions as legitimate or fraudulent. The UCI repository's dataset is only partially available. The dataset consists of 1000 instances (credit candidates) and 21 attributes (14 of which are definite/insignificant and seven of which are numerical). In the dataset, access signifies a person who acquires credit from a bank, with each indication describing a particular member's credit standing, whether good or bad. Based on a specific set of characteristics, every person is categorized by having excellent credit or bad credit.

By using cleaning and transformation procedures, data preprocessing converts inconsistent and incomplete real-world data into processed data that is understandable. The quality of the data is crucial for every machine learning algorithm since poor data quality can harm the classifier's effectiveness. To improve the model's effectiveness and reduce its training period, the redundant and irrelevant features are removed. Two real-world datasets from Kaggle were used in this research project. One of the datasets, an online transactional data set, was donated by an e-commerce payment service provider to help machine learning programmers combat the biggest loss caused by frauds that occur globally.
By reducing duplicated features in the dataset, the Feature Selection (FS) strategy enhances learning performance.

The GA methodology is a highly well-liked method in the study of evolutionary computation. It resembles a natural selection. It is extensively used in business, engineering, and a variety of other fields. Getting the best possible answer to an issue is the preferred course of action. Three fundamental operators make up GA: crossover, mutation, and selection. Based on fitness function, Selection separates the best-fit individuals from the available population group. Crossover occurs when the first half of the second record is merged with the second half of the first record. Bits are randomly switched from 0 to 1 and vice versa with mutation.

Support Vector Machine (SVM): The SVM, a supervised machine learning method, is used for both classification as well as regression applications. But classification issues frequently make use of SVM. The SVM implementation models each data point as a point in an n-dimensional space, with each component's

estimate at a particular location (where n is the number of features). The elements are then categorised using the hyper-plane that separates the two classes generally.

Decision Tree: For data D training samples, trees are built using high entropy inputs. These trees are built rapidly and easily using the top-down recursive Divide-And-Conquer (DAC) approach. On D, tree trimming was done to get rid of unnecessary samples.

K-Nearest Neighbor (KNN): K-Nearest Neighbors (KNN), one of the most basic Machine Learning algorithms, is used for both regression as well as classification. New data points are characterized by KNN calculations using the data and resemblance measures (such as the distance function). In its most basic form they say that the KNN formula accepts that things in comparison are close to one another. Classification in KNN takes into account the majority vote for its neighbours. The class with the closest neighbours receives the data point. The accuracy of k estimation may increase as the number of nearest neighbours increases. The predictor variables are the dependent or target variable. The chances of developing cardiovascular sickness is the situation's target variable, and the KNN algorithm will forecast the outcome.

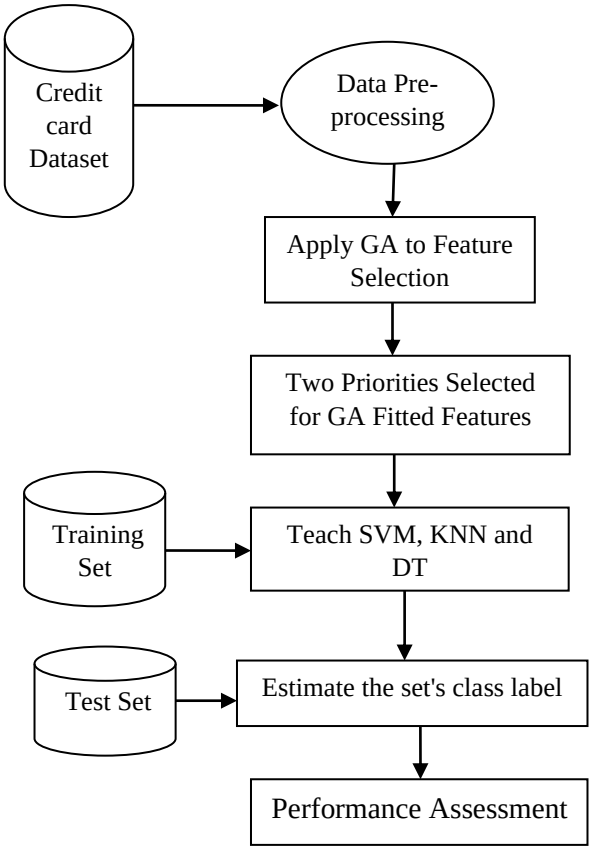

Fig.1: Block Diagram of Prevention Of Credit Card Fraud Transaction Using GA Feature Selection With Machine Learning

With the idea that predictors are conditionally independent, the training dataset is fed to the model in the first phase, which is referred to as the learning stage to determine the limitations of a probability distribution. In order to forecast and assess the probability of type posterior for each class, ML classifier be fed the test dataset and novel data during the prediction phase of the second stage. Using the test dataset's greatest posterior probability, a classification is made in following step.

Accuracy, Precision, and Sensitivity are utilized to calculate the performance analysis for the proposed system.

IV. RESULT ANALYSIS

This section discusses the proposed hybrid Machine Learning for analysing the results

and predicting credit card fraud transaction. In compared to other methods, the presented categorization method has the highest accuracy. This result reveals clearly that the characteristics and ML methods are successful in accurately predicting fraud transaction when compared to well-established models. This system makes use of DT, KNN and SVM classifiers. The performance is evaluated based on the classified instances namely TP, TN, FP, and FN which are defined as follows:

True Positive (TP): If a sample is predicted correctly as positive and actually it is positive.

True Negative (TN): If a sample is correctly predicted as negative and actually it is negative.

False positive (FP): If a sample is incorrectly predicted as negative but actually it is positive.

False Negative (FN): If a sample is predicted incorrectly as positive but actually it is positive.

Table 1: Performance Comparison Table

ML classifiers	Accuracy	Sensitivity	Precision
SVM	75.8	79.8	83.5
KNN	97.2	96.3	95.2
DT	74	80.3	81.9

Based on these values the performance metrics like sensitivity, precision, accuracy, specificity and classification error are measured for performance evaluation of this system.

Accuracy: It is a performance parameter that measures the system's capacity for accurate prediction, and it is represented as

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}\text{x } 100 \quad \text{-(1)}$$

Precision: The ability of a system to create only useful outcomes is measured by precision.

$$\text{Precision}=\frac{TP}{(TP+FP)}\text{ x } 100 \text{ –(2)}$$

Sensitivity: Recall or True Positive Rate (TPR) are other names for it. It is a performance parameter that measures the system positively.

$$\text{Sensitivity}=\frac{TP}{TP+FN}\text{ x } 100 \text{ –(3)}$$

Table.1 displays the results of the performance assessment of the GA algorithm with ML for preventing credit card fraud transactions.

Compared to Decision Tree, KNN and SVM classifiers, the KNN classifier has greater performance in terms of accuracy, precision and sensitivity. Fig. 2 shows accuracy, Sensitivity, Precesion Comparison of Decision Tree, KNN and SVM classifiers with various thresholds. X-axis shows classification and Y-axis represents percentage (%).

Fig. 2 shows comparative graph of accuracy for DT, KNN and SVM for prevention of fraud credit card transaction.

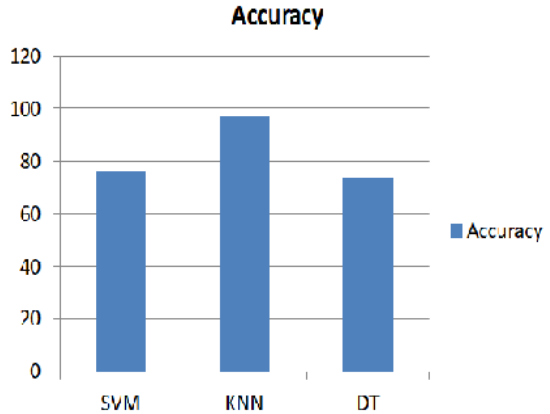

Fig.2 Comparative Graph Of The Accuracy

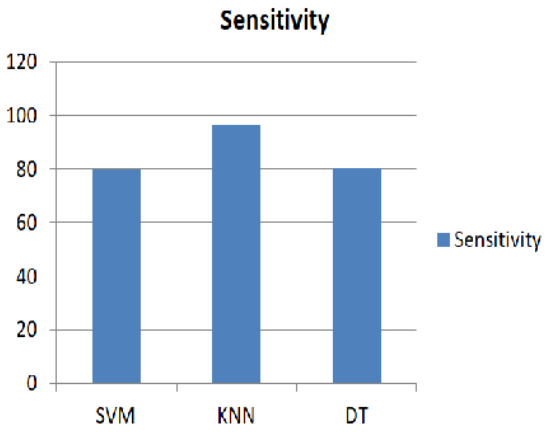

Fig.3: Comparative Graph Of The Sensitivity

Fig 3 shows comparative graph of sensitivity for DT, KNN and SVM for fraud credit card transaction.

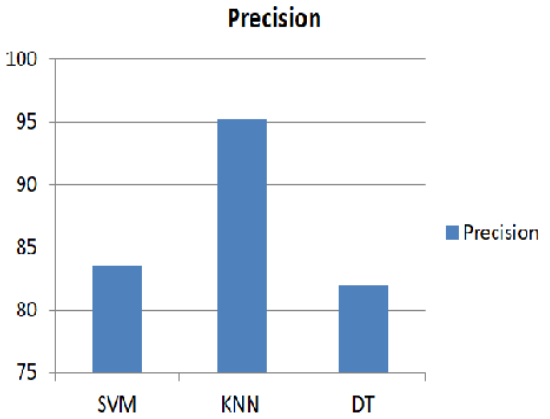

Fig. 4: Comparative Graph Of The Precision

Comparison graphs of DT, KNN, and the SVM model for preventing the fraud credit card transaction is shown in Fig. 4.

As a result, the Prevention of Credit Card Fraud Transactions Using GA Feature Selection with Machine Learning outperformed other ML classifiers in terms with to precision, sensitivity, and accuracy for predicting Credit Card Fraud Transactions.

V. CONCLUSION

Credit cards are becoming a common method of payment because of recent technological advancements. The operation's security flaws have made fraud a developing trend, losing millions of dollars each year. To lower the rate of credit card payment fraud, a fraud detection and prevention plan is needed. In the paper, machine learning-based CCF detection was provided. The proposed method improved the classification algorithm by prioritizing feature selection in a genetic algorithm. According to the results, RF outperformed other classifiers in terms of accuracy while choosing the first priority features. The optimum strategy advised for credit card detection systems is first priority feature selection. On the credit card dataset, an unbalance dataset, CCF is detected using the supervised machine learning techniques of DT, K-NN, and SVM. The first key focus attributes are the most crucial components, according to the experimental results of the framework. The results also showed that the KNN technique performed better than DT and SVM in terms of reliability, proximity, and precision.

VI. REFERENCES

[1] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," in Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE, 2020, no. Iciccs, pp. 1264–1270.

[2] P. R. Vardhani, Y. I. Priyadarshini, Y. Narasimhulu, Á. Nonparametric, *CNN* Data Mining Algorithm for Detecting Credit Card Fraud. Singapore.: Springer Singapore, 2019.

[3] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi," credit card fraud detection using Adaboost and majority voting". IEEE Access, vol. XX, pp,2018.

[4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, " Cat Boost : unbiased boost ing wit h cat egorical feat ures," in Advances in neural informat ion processing systems, pp. 6638–6648, 2018.

[5] K. Randhawa, C. H. U. K. Loo, and S. Member, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," IEEE Access, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[6] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, pp. 1–6.

[7] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, pp. 1–6.

[8] A. O. Adewumi and A. A. Akinyelu, "A survey of machine-learning and nature-inspired based credit card fraud detection techniques," Int. J. Syst. Assur. Eng. Manag., vol. 8, no. 2, pp. 937–953, 2017.

[9] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using Machine Learning Techniques : A Comparative Analysis," in International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1–9.

[10] Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. "Credit card fraud detection: a realistic modeling and a novel learning strategy." IEEE transactions on neural networks and learning systems, vol. 29, no. 8, pp. 3784-3797, 2017.

[11] Babu, S. Kishore, S. Vasavi, and K. Nagarjuna. "Framework for Predictive Analytics as a Service using ensemble model." In 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 121-128, IEEE, 2017.

[12] M. Hegazy, A. Madian, and M. Ragaie, "Enhanced Fraud Miner : Credit Card Fraud Detection using Clustering Data Mining Techniques," Egypt. Comput. Sci. J., vol. 40, no. 03, pp. 72–81, 2016.

[13] Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective." Expert systems with applications, vol. 41, no. 10, pp. 4915-4928, 2014.

[14] W. Zhou, G. Kapoor, "Detecting evolutionary financial statement fraud", Decision Support Systems, pp. 570–575, 2011.

[15] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, no. 3, pp. 602–613, 2011.