**Random Forest for Credit Card Fraud Detection**

# I. INTRODUCTION

Credit cards are frequently used as a result of the growth of mobile clever devices and e-commerce. Card-not-present transactions (i.e., online transactions without a real card) [1] are more common, especially since web payment gateways like PayPal and Alipay handle all credit card procedures. An online transaction is now simpler and more practical thanks to credit cards. However, transaction frauds are on the rise and cost businesses a lot of money every year [18] [19]. By 2020, losses are anticipated to climb at double-digit rates each year [2]. Fraud detection is the technique of keeping track of a cardholder's transaction patterns to determine if an incoming transaction was made by the cardholder or someone else [10].In general, abuse detection and anomaly detection are the two main types of fraud detection techniques [15]. In order to assess whether an incoming transaction is fraudulent or not, misuse detection uses categorization techniques. Such an approach typically requires knowledge of the current fraud kinds in order to create models by studying the various fraud patterns. Building a profile of a cardholder's typical transaction behavior using historical transaction data is the goal of anomaly detection, and if a new transaction deviates from the expected transaction behavior, consider it to be possible fraud. However, an anomaly detection system requires enough subsequent sample data to define a cardholder's typical transactional behavior.

# II. RELATED WORK

To effectively address the issue of credit card theft, it is beneficial for us to have a thorough understanding of fraud detection technologies. The study in [16] offers a thorough analysis of the difficulties and issues in fraud detection research. There are essentially two types of credit card theft: behaviour fraud and application fraud 3]. Application fraud is when thieves use other valid cardholders' information or fabricate bogus information to obtain new credit cards from issuing firms. When someone uses a card fraudulently, they take the account information and password from the legitimate cardholder and use them to make purchases. Nearly all of the current research on the detection of credit card fraud focuses on recording cardholder behaviour patterns and identifying fraudulent transactions using these patterns. In order to detect fraud, Srivastava et al. [5] use a hidden markov model (HMM) to represent the sequence of transaction attributes in credit card transaction processing. An HMM is initially trained using the cardholder's typical behaviour. The current transaction is regarded as fraudulent if the trained HMM rejects it with a high likelihood. However, they merely take into account the transaction amount as a feature in the procedure.Amlan et.al [8] propose a method using two-stage sequence alignment which combines both misuse detection and anomaly detection [15]. In their method, a profile analyzer is used to determine the similarity of an incoming sequence of transaction on a given credit card with the legitimate cardholder’s past spending sequence. . Then, the unusual transactions traced by the profile analyzer are passed to a deviation analyzer for possible alignment with the past fraudulent behavior. Sahin and Duman [12] make a comparison between decision tree and support vector machine (SVM) in detecting credit card fraud. They divide a dataset into three groups which are different in the ratio between fraudulent transactions and legitimate one and they develop seven decision tree and SVM based models and test them in these datasets. the conclusion is that random forest has the best performance among these techniques with the process of aggregation. However, the aggregated method in this work fails to detect a fraud in real time. Below is the figure of our labels column in the dataset.

Table

Description automatically generated

# III. RANDOM FOREST

Our classifier is random forest [20]. Decision tree models' [23] prevalence in data mining can be attributed to their algorithm's simplicity and adaptability to various data attribute kinds. Single-tree models, however, may be susceptible to certain training data and are simple to overfit [11]. Ensemble approaches, which integrate a number of individual judgements in some way to tackle these issues, are more accurate than single classifiers [21]. One of the ensemble approaches is random forest, which combines a number of tree predictors so that each tree is reliant on a different random independent dataset and all the trees in the forest have the same distribution [20]. The capability of a random forest is influenced by the association between various trees as well as the strength of each individual tree. The effectiveness of a random forest increases with the strength of each individual tree and decreases with the association between distinct trees. The randomization of trees, which uses bootstrapped samples and randomly chooses a subset of data attributes, is what gives them their diversity.Despite the possibility of some cases in our dataset being mislabeled, random forest is still resistant to noise and outliers. The algorithm that has been used are two in this detection which are as follows: Algorithm I describes the process of producing a type-I random forest:

## Algorithm I:

Input: Dataset D and the number of trees NT.

Output: A random forest.

For i = 1 to NT:

A following algorithm II describes the process of producing a type-II random forest:

Algorithm II:

Input: Dataset D, the number of trees NT and the threshold T of Gini impurity

Output: A random forest

For i = 1 to NT:

The method of node splitting is the primary distinction between the two algorithms. In type-I random forest, data are distributed according to the attribute with the least amount of gini impurity; in type-II random forest, data are dispersed according to the distance between records and two centres. Each of them has advantages and disadvantages of their own. Because Algorithm I must determine the distances between computing centres and all records, it may be faster in computing centres but slower in scattered records. Algorithm II would distribute data more quickly, even though it might compute gini impurity for attributes more slowly. The confusion metrices obtained from the random forest classifier has been shown in below figure.

A picture containing text, monitor, screen, screenshot

Description automatically generated

Moreover, the parameters obtained on the dataset is also described below:

The model used is Random Forest classifier

The accuracy is 0.999

The precision is 0.974

The recall is 0.785

The F1-Score is 0.870

The Matthews correlation coefficient is0.874

# IV. CONCLUSIONS

Our study makes use of a real-world B2C dataset of credit card transactions. There are still certain issues, like as skewed data, even if random forest produces good results on tiny sets of data. Solving these issues will be the main focus of our upcoming efforts. It is necessary to enhance the random forest method itself. The voting process, for instance, counts all base classifiers equally, even though some of them might be more significant than others. As a result, we also attempt to improve this method.

# VI. REFERENCES

[1] Gupta, Shalini, and R. Johari. ”A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant.” International Conference on Communication Systems and Network Technologies IEEE, 2011:22-26.

[2] Y. Gmbh and K. G. Co, “Global online payment methods: Full year 2016,” Tech. Rep., 3 2016.

[3] Bolton, Richard J., and J. H. David. ”Unsupervised Profiling Methods for Fraud Detection.” Proc Credit Scoring and Credit Control VII (2001):5– 7.

[4] Seyedhossein, Leila, and M. R. Hashemi. ”Mining information from credit card time series for timelier fraud detection.” International Symposium on Telecommunications IEEE, 2011:619-624.

[5] Srivastava, A., Kundu, A., Sural, S., and Majumdar, A. (2008). Credit card fraud detection using hidden markov model. IEEE Transactions on Dependable and Secure Computing, 5(1), 37-48.

[6] Drummond, C., and Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. Proc of the Icml Workshop on Learning from Imbalanced Datasets II, 1–8. [7] Quah, J. T. S., and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732.

[8] Kundu, A., Panigrahi, S., Sural, S., and Majumdar, A. K. (2009). Blastssaha hybridization for credit card fraud detection. IEEE Transactions on Dependable and Secure Computing, 6(4), 309-315.

[9] Shi, E., Niu, Y., Jakobsson, M., and Chow, R. (2010). Implicit Authentication through Learning User Behavior. International Conference on Information Security (Vol.6531, pp.99-113). Springer-Verlag.

[10] Duman, E., and Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications, 38(10), 13057-13063.

[11] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). Data mining for credit card fraud: a comparative study. Decision Support Systems, 50(3), 602-613.

[12] Sahin, Y., and Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines. Lecture Notes in Engineering and Computer Science, 2188(1).

[13] Mota, G., Fernandes, J., and Belo, O. (2014). Usage signatures analysis an alternative method for preventing fraud in E-Commerce applications. International Conference on Data Science and Advanced Analytics (pp.203-208). IEEE.

[14] Behdad, M., Barone, L., Bennamoun, M., and French, T. (2012). Natureinspired techniques in the context of fraud detection. IEEE Transactions on Systems Man and Cybernetics Part C, 42(6), 1273-1290. [15] Ju, W. H., and Vardi, Y. (2001). A hybrid high-order markov chain model for computer intrusion detection. Journal of Computational and Graphical Statistics, 10(2), 277-295.

[16] Bolton, R. J., and Hand, D. J. (2002). Statistical fraud detection: a review. Statistical Science, 17(3), 235-249.

[17] Vlasselaer, V. V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., and Snoeck, M., et al. (2015). Apate : a novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38-48.

[18] Chan, P. K., Fan, W., Prodromidis, A. L., and Stolfo, S. J. (2002). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and Their Applications, 14(6), 67-74.

[19] RONG-CHANG CHEN, TUNG-SHOU CHEN, and CHIH-CHIANG LIN. (2006). A new binary support vector system for increasing detection rate of credit card fraud. International Journal of Pattern Recognition and Artificial Intelligence, 20(02), 227-239.

[20] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[21] Dietterich, T. G. (2000). Ensemble methods in machine learning. , 1857(1), 1-15.

[22] Abeel, T., de Peer, Y. V. and Saeys, Y. Java-ML: A Machine Learning Library, Journal of Machine Learning Research, 2009, 10, 931-934

[23] Quinlan, J. R. (1986). Induction on decision tree. Machine Learning, 1(1), 81-106.

[24] Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). Classification and regression trees. Biometrics, 40(3), 358