Internet Financial Fraud Detection based on a Distributed Big Data Approach withNode2vec

[1] Gudura Raveendrababu, [2] Mohammad Abdul Hafeez, [3] Shaik Abeed Basha, [4] Indlapa Divya,
CSE Department,
[1,2,3,4] Assistant Professor, Dhruva Engineering Collage, Hyderabad.
Shree Engineering Collage, Hyderabad.

ABSTRACT. With the use of Decision Tree and Neural Network Algorithms, anti-money-laundering systems may have their false positive rate reduced. Methods and Instruments for Research: Classification accuracy for anti-money-laundering analysis is 87.75% for a recurrent neural network technique (N=10) and 69.60% for a decision tree (N=10). With a pre-test power of 0.8 and an alpha of 0.05, we calculate the sample size using GPower. Finally, when comparing mean accuracy (87.75%), Neural Networks perform better than Decision Trees. Both the accuracy and loss results are statistically significant ($p>0.05$) at the 0.557 level. Finally, the Mean Accuracy of Neural Networks Anti-Money Laundering Systems Is Higher Than That of Decision Trees.

Some of the terms used in this context are decision tree algorithms, anti-money laundering, machine learning, innovative detection, transparent transactions, and recurrent neural networks.

INTRODUCTION

A practice known as money laundering involves making large amounts of money that have been obtained unlawfully, via activities like drug trafficking, terrorism, or other major crimes, seem as if they came from a legitimate source. Reporting suspicious behavior is crucial to the battle against anti-money laundering (AML) and is mandated by laws and regulations. [1, 2]. The term "anti-money laundering" encompasses a wide range of regulations and policies put in place to prevent criminals from disguising their illicitly acquired funds as legal revenue. [3]. However, the economy and the people it services are its larger targets of danger. There has been a steady rise in financial crime in recent years [4,5]. Consequently, in order to prevent illegal and unethical behavior, AML must be detected and monitored [6]. Additionally, fewer terrorist strikes are possible as a result of AML detection since individuals are unable to finance terrorists [4, 5]. Additionally, AML regulations serve as safeguards against broader forms of fraud.

Data for this study comes from the banking industry's anti-money-laundering system database. Many scholars study on this, and it has resulted in many published studies. In the last five years, thirty publications have been published, with twelve appearing on Research gate and eighteen on IEEE Explore. the references [7], [8]. A secret key and security key are helpful since they are dynamic and change each time the user launches the application [9]. The account is secure since the key is produced only when the application is executed, making it impossible to access the account. Because One Time Passwords (OTPs) are so readily cracked and obtained by third parties, we utilize secret keys and security keys to protect our accounts and the users' financial data. The number ten. Due to the large number of banks in India, compiling the bank information is an enormous and complex task [7], [8]. Additionally, we have considered a great deal of data, including operational costs, sluggish procedures, high commissions, poor savings stimulation, online or virtual banking, ATM permanent network, etc. Improve search via technology, maintain regular cross-communication, use data analytics to spot trends, standardize systems, and provide training; these are the five pillars of anti-money laundering.

Banks must guarantee data safety and promptly identify fraud; nevertheless, the current approach has a significant false positive rate when evaluating red signal gaps. Existing anti-money-laundering detection mechanisms are becoming less effective as a result of the proliferation of internet transactions. This study's overarching goal is to provide a model for an anti-money-laundering system that reduces the false positive rate, which in turn improves the detection and prediction rates, simplifies the system's operation, and makes it more accurate.

MATERIALS AND METHODS

The study was carried out by the data analytics lab of the Saveetha School of Engineering. The Saveetha School of Engineering was recognized for their role in the planning, execution, and mistake repair processes. There is no need for human samples or data in this investigation. Our study does not need any kind of ethical clearance. Within the scope of this research, two categories have been established: decision trees and neural networks. Twenty samples were used in this investigation, with 10 data sets from each group. Compared to Decision Trees, Neural Networks achieve an accuracy rate of 87.75%. Using a pretest power of 0.8 and an alpha of 0.05, Gpower 3.1 is the program that is used. The dataset used in this research was made public on the GitHub platform.

We mined GitHub.com for the dataset that may aid in real-time prediction. Important variables such as counts, minimum and maximum amounts, bank IDs, and firm types are included in the dataset's 8 columns and 1575 rows. There is account data in the dataset. In order to fight money laundering, the data collected from this source is saved in a specific folder [10]. All requirements are satisfied by the acquired dataset for the purpose of testing the ML algorithm. Dividing the dataset in half is the next stage. Currently, we're considering allocating 50% to training and 50% to testing (or anti-money laundering). Following testing and training on the dataset and using the suggested Novel Neural Network method, the accuracy percentage will be calculated [3]. The pseudo code for the neural network is available in Table 1. Following dataset training and testing, the accuracy % may be ascertained by using the decision tree comparative technique [2]. The details of the pseudo code decision tree are shown in Table 2.

Program Code for a Neural Network-Based Anti-Money Laundering System (Table 1)

INPUTS: Anti-Money-Laundering Training Dataset
First, review the Anti Money Laundering test results. Next, we need to extract the properties related to anti-money laundering. Step 3: Retrieve the AML records. Fourth, evaluate the security key and secret key side by side. Next, implement the neural network algorithm. Algorithm for Neural Networks, Step 6: f(b+∑_(i=1)^n x_i w_i). Stage 7: Detection of Forecast Fraud. Step 8: Check the precision .
Classified as either authentic or deceitful

The second table shows the pseudo code for the decision tree-based anti-money-laundering system.

INPUTS: Anti-Money-Laundering Training Dataset

	First, review the Anti Money Laundering test results. Next, we need to extract the properties related to anti-money laundering. Step 3: Retrieve the AML records. Fourth, evaluate the security key and secret key side by side. Execute the Decision Tree Algorithm in Step 5. The Decision Tree Algorithm Formula (Step6) E(T,X)=∑_cϵX P(c) E(c) Step 7: Offers an approach to presenting algorithms. Retain precision in Step 8.
	Classified as either authentic or deceitful

To put the plan into action, the Python OpenCV package is used. Windows 10 was among the first platforms to try deep learning. The hardware configuration consisted of an Intel Core i5 CPU and a 4 GB RAM module. Sorting was done using the 64-bit system. The Java platform was used to implement this code. To guarantee an accurate outcome, the dataset is quietly handled while the program is executing.

STATISTICAL ANALYSIS

I utilized IBM SPSS version 21 as the statistical tool for the research. Ten samples are needed for dataset creation using SPSS's Decision Tree and Neural Network methods. The quantity and kind of enterprises are the dependent variables, while details such as maximum and minimum balances and bank IDs make up the independent variables.

RESULTS

For the purpose of statistical analysis, twenty individuals are used to explore novel decision trees and neural networks. Both decision tree and neural network methods undergo statistical examination. Various filtering algorithms are evaluated for their accuracy and grouping values. In addition to providing statistical information, these 20 data samples allow us to compare losses. Table 3 displays the group, accuracy, & loss statistics for the two methods—neural network and decision tree—that were previously discussed. Accuracy rate averages and standard deviations, together with information on the sample size, are all included in the group statistics table.

The anti-money-laundering system uses data that is organized into 31 columns with 8 widths, as shown in Figure 3.

SI.NO	Name	Type	Width	Decimal	Columns	Measure	Role
1	Group	Numeric	8	0	27	Nominal	Input
2	Accuracy	Numeric	8	4	27	Scale	Input
3	Loss	Numeric	8	2	27	Scale	Input

For each method, Table 4 displays the aggregate statistics, including the mean, standard deviation, and standard error mean. The given data set is subjected to the independent sample T test with a 95% confidence interval. Independent t-sample tests for the methods are shown in Table 5. We determine their mean loss by comparing the algorithms' accuracy. The mean loss and accuracy of a neural network and a decision tree are compared in Figure 1. Compared to decision trees, which achieved an accuracy of 69.50% in the Anti-Money Laundering system before, neural networks achieve an accuracy of 87.75%.

Figure 4: Statistical Analysis of Neural Networks and Decision-making Groups Averaging the measurement tree's variance, error bar, and measurement tree

	Group	Algorithm	N	Mean	Std Deviation	Std.Error Mean
Accuracy	1	Neural Network	10	87.7500	6.43657	2.03542
	2	Decision Tree	10	69.6000	5.50818	1.74184
Loss	1	Neural Network	10	12.2500	6.43657	2.03542
	2	Decision Tree	10	30.4000	5.50818	1.74184

TABLE 5. Two groups are subjected to an independent sample T-test in order to ascertain significance and standard error. There is no statistical significance with a 95% confidence interval since the p-value is more than 0.05 (0.557).

		Equality of variance test using Levene's formula		T-Test to See If Means Are Equal						
				t	df	Sig (2 - tailed)	Mean difference	Standard Error Distinction	95% certainty level of dissimilarity	
		F	Sig						Lower	Upper
Accuracy	**Equal variances assumed**	.358	.557	6.775	18	0.000	18.15000	2.67898	12.52167	23.77833
	Equal Variances not assumed			6.775	17.580	0.000	18.15000	2.67898	12.51202	23.78798
Loss	**Equal variances assumed**	.358	.557	-6.775	18	0.000	-18.15000	2.67898	-23.77833	-12.52167

	Equal Variances not assumed			-6.775	17.580	0.000	-18.15000	2.67898	-23.78798	-12.51202

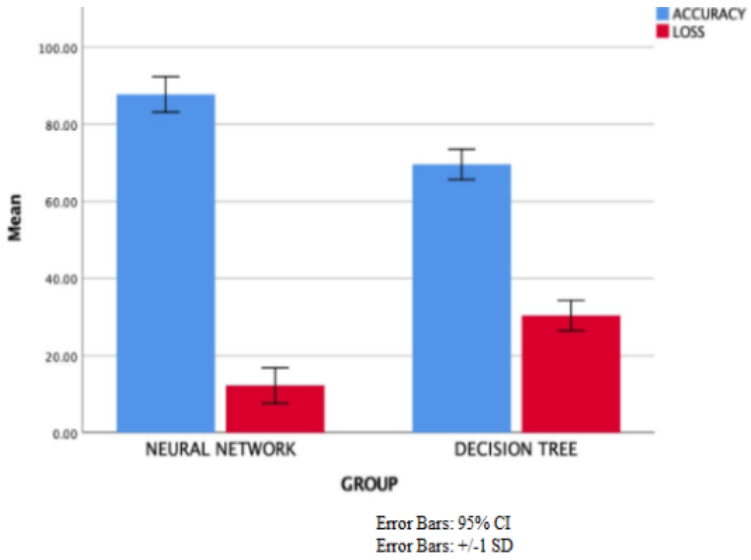

FIGURE 1. When comparing neural networks with decision trees, we look at their average accuracy and loss. On average, the Neural Network outperforms the Decision Tree. A Neural Network's standard deviation is somewhat less than a Decision Tree's X-axis standard deviation: The y-axis compares the mean detection accuracy (± 1 SD) of a back propagation neural network with that of a decision tree.

DISCUSSION

The study's significance level of p=0.557 indicates that the Neural Network approach seems to be more successful than the Decision Tree strategy. Unexpectedly low g-power ratings are the result of insufficient datasets. The Neural Network approach outperforms the Decision Tree method by an impressive margin of 87.75%. There was a median, interquartile range, and standard deviation of 2.03542 for neural network techniques in the 20-person sample. Compared to Decision Tree, Neural Network is obviously the superior method. According to the decision tree methods, the mean is 69.6000 and the standard deviation is 5.50818.

Research from research [5, 20]–[22] shows that a Neural Network technique achieves an accuracy of 87.75% whereas the Decision Tree algorithm only reaches 69.50%. Pictured here are figures [20] through [22]. Compared to earlier detection assessments using algorithms like Decision Tree Identification, the accuracy of anti-money laundering categorization using Neural Network techniques has reached 87.75%. The Decision Tree algorithm's 69.50% success rate pales in contrast to the Neural Network technique. Discoveries, including illicit funds, may be uncovered with the use of decision trees. Graphing the variation in Neural Net Accuracy and determining performance for fraud detection with various approaches might assist improve efficiency. references [23], [24][25][26] cite this work. We outperformed rival classifiers in our investigation.

The main problem with decision trees is that they often result in data outfitting. By just modifying the input data, decision tree algorithms may undergo a dramatic transformation in their overall structure. It is recommended that the previously mentioned study be enhanced in order to reduce the redflag gap false positive rate. Software used in the

banking and legal sectors may be integrated with the anti-money-laundering system to fulfill regulatory requirements that regulated firms must detect and disclose suspected instances of money laundering.

CONCLUSION

According to this anti-money-laundering research, Neural Networks achieve an accuracy of 87.75%, whereas Decision Tree Algorithms only manage 69.60%. Thus, it may be concluded that Neural networks seem to have far higher accuracy than Decision Trees.

REFERENCES

1. M. Ziemski, T. Wisanwanichthan, N. A. Bokulich, and B. D. Kaehler, "Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences," *Front. Microbiol.*, vol. 12, p. 644487, Jun. 2021.
2. S. DadashzadehFahim, R. Karimi, and Z. Faghihrasoul, "Bank Transaction Analyze for Recognize of Money Laundering Using Decision Tree Algorithm (DTA)." doi: 10.21203/rs.3.rs-685083/v1.
3. N. A. L. Khac, N. A. Le Khac, and M.-T. Kechadi, "Application of Data Mining for Anti-money Laundering Detection: A Case Study," *2010 IEEE International Conference on Data Mining Workshops*. 2010. doi: 10.1109/icdmw.2010.66.
4. M. Ziemski, T. Wisanwanichthan, N. A. Bokulich, and B. D. Kaehler, "Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences," *Front. Microbiol.*, vol. 12, p. 644487, Jun. 2021.
5. D. Chau and M. van Dijck Nemcsik, *Anti-Money Laundering Transaction Monitoring Systems Implementation: Finding Anomalies*. John Wiley & Sons, 2020.
6. C.-W. Zhang and Y.-B. Wang, "Research on application of distributed data mining in anti-money laundering monitoring system," *2010 2nd International Conference on Advanced Computer Control*. 2010. doi: 10.1109/icacc.2010.5487272.
7. D. Chau and M. van Dijck Nemcsik, *Anti-Money Laundering Transaction Monitoring Systems Implementation: Finding Anomalies*. John Wiley & Sons, 2020.
8. E. Waluyo, "UPAYA MEMERANGI TINDAKAN PENCUCIAN UANG (MONEY LAUNDRING) DI INDONESIA," *Jurnal Dinamika Hukum*, vol. 9, no. 3. 2009. doi: 10.20884/1.jdh.2009.9.3.235.
9. F. Thompson, "Debate: Financial panic, economic threat, and sovereign debt crises," *Public Money & Management*, vol. 32, no. 3. pp. 164–165, 2012. doi: 10.1080/09540962.2012.676271.
10. A. U. Bello, *Improving Anti-Money Laundering Compliance: Self-Protecting Theory and Money Laundering Reporting Officers*. Springer, 2017.