iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66490

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

A Survey on Online Payment Fraud Detection Techniques using Machine Learning Algorithms

Vulugundam Anitha[1], Ch. Siri[2], K. Sai Meghana[3], M. Joshna[4], G. Akanksha[5]
[1]*Assistant Professor, Dept. of Electronics and Telematics Engineering, G. Narayanamma Institute of Technology & Science (For Women), Hyderabad, Telangana, India*
[2,3,4,5]*B.Tech Students, Dept. of Electronics and Telematics Engineering, G. Narayanamma Institute of Technology & Science (For Women), Hyderabad, Telangana, India*

Abstract: ***Online transaction fraud detection has become a critical challenge with the rise of digital payment systems. This paper surveys various machine learning techniques employed in fraud detection, including Support Vector Machines (SVM-QUBO), Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, and Random Forest. The performance of each method is evaluated in terms of accuracy, precision, recall, and computational efficiency. The study also explores how different algorithms handle high-dimensional, imbalanced datasets and the impact of feature selection techniques. The results provide insights into the strengths and limitations of these algorithms, offering a comprehensive comparison for fraud detection in digital transactions.***

Keywords: Fraud Detection, Machine Learning, Random Forest, SVM-QUBO, Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Trees.

I. INTRODUCTION

In the era of e-trade development, the growth is unprecedented and there can be almost one billion transactions daily. While this evolution seems to assures convenience and adoption at any place in the world, it brings a large affirmation of virus-wannabe cyber practices such as identity theft, credit card fraud and even payment repudiation. Corrupt practices outlined above cut down direct financial outcomes for the business and a violation of trust quality for online shopping forums. Online fraud hence needs intelligent and appropriate measures for combating.

That is why the routine of the Machine learning (ML) became the primary statistical fighting shield to analyze and recognize the bizarre practices in real-time.

Thus, by prodding of sophisticated terminal-based ML algorithms [1], a prudent detection system has an obligation to safeguard web- transactions, ward off users from businesses, and maintain confidence in the e-market. However, developing such a model of fraud detection is not without some considerations as this article seeks to address. Its applicability in real life puts one in a position where one has to handle large number of transactions and so a good model should in principle be assumed capable of easily and effectively identifying potentially fraudulent transactions.

Machine learning is shown to be active in enhancing the detection algorithm through an investigator feedback loop. Building an application that is based on the World Wide Web to identify, predict, report any suspicious and fraudulent transaction, and block any transaction that the merchant may consider as fraud.

The solution would concurrently enhance online payment security and enable businesses to manage defaults that keep operations going with increasing threats. [2]

II. FRAUD DETECTION TECHNIQUES

A. Description of the Dataset

The information for 'Online Payments Fraud Detection' dataset is collected from the Kaggle repository for fraud detection. The dataset contains a predisposed number of observations for the fraud class therefore, we have resorted to much-use algorithms to obtain us better results with this unbalanced dataset. Also, the dataset had features with multiple datasets within the same column. Thus, to attend to this, we have opted for 'category encoders' to convert the strings to numerical values.

B. Steps in Machine Learning

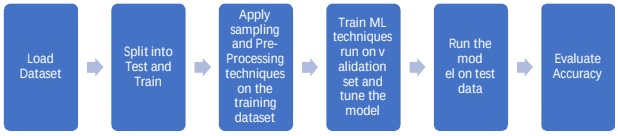


Fig. 1. Flowchart of steps involved in machine learning

Step 1: We feed the system with a lot of data and try to process it.
Step 2: Split the data into training data and testing data so that you should be able to analyze the effectiveness of the machine during the period that it has been trained.
Step 3: The raw data are then preprocessed with reference to the mentioned tables, and quantitatively encoded for categorical and scaled.
Step 4: Acquire complex behavior with respective training algorithms.
Step 5: On the generation of answers over the test data.
Step 6: Verifying results regarding various specifications of the performance criteria of correctness.

III. FRAUD DETECTION TECHNIQUES

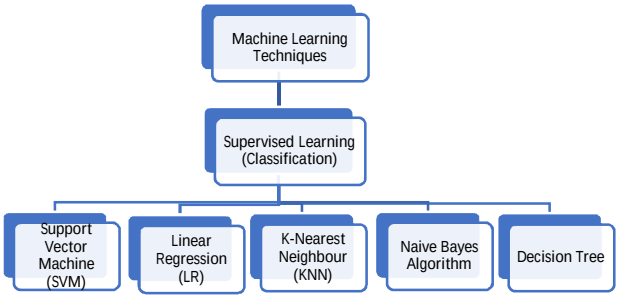


Fig. 2. Categorization of supervised learning algorithms

A. Support Vector Machine (SVM)

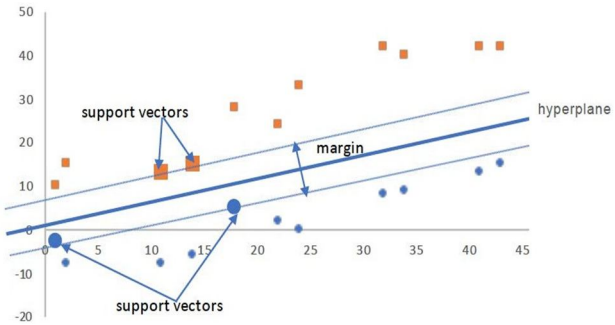


Fig. 3. Support Vector Machine (SVM) [5]

The Support Vector Machine (SVM) belongs to the most widely used algorithms in classification and regression problems. SVM is an algorithm that identifies the right hyperplane that will differentiate between data points belonging to different classes. By using the kernel function, Linearly Non- separable data are transformed to form linearly separable data in a higher dimension of the figure. Borrowing from the tradition of neural networks, a sigmoid kernel has been seen to have very impressive performance in capturing non-linear relationships. SVM balances sample space errors by using the function of the regularization parameter to allow control of the fitting accuracy of the model on the training data when compared to the generalization ability of the model. This is the very reason why SVM is not only flexible but also very effective; the ability to mix the two.[3]

B. Logistic Regression Algorithm

Logistic Regression is a model for the classification task, which conservatively identifies whether an instance belongs to one of two classes, for instance fraud (class 1) or not fraud (class 0) from a dataset it has previously learned from. Perversely, though, it is used for classification rather than regression because after performing transformations on the input variables using the logistic function also known as sigmoid, it produces the probability that the data at hand belongs to a particular class. The model identifies parameters (weights particularly) that form the best fit between the inputs and the target class in minimizing classification errors. After that, the data is classified as fraud (equal to 1) or non-fraud (equal to 0), based on the probability value that exceed the standard of 0.5.[4]

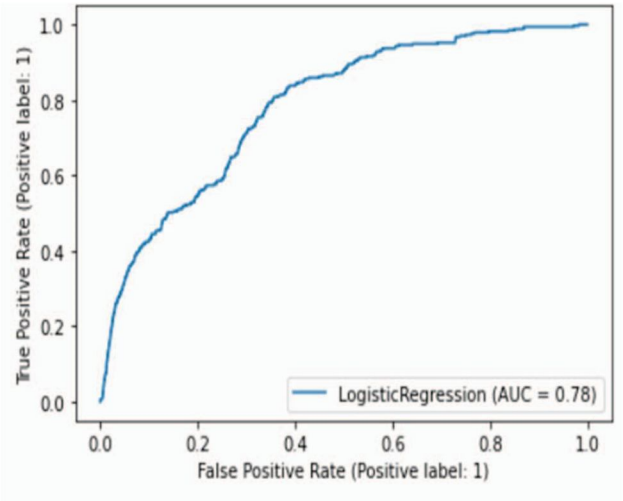


Fig. 4. ROC of Logistic Regression Algorithm [9]

C. KNN (K-Nearest Neighbors)

KNN (K-Nearest Neighbors): It is an ML approach used in addressing classification and regression problem in a particular data set. It was described as an instance-based learning technique though it indeed predicts based on instances in the training data. In KNN the new data point is categorized or in other words is allocated a class label or a number from the set of numbers that normally represents the class label or numbers for its “K” closest neighbor points within the training set. The metric used in this paper to determine K nearest neighbors is based on the distance between feature of the data points.

To apply the KNN algorithm, we need to define the following parameters:

1) K: The no of neighbours to take into consideration while deriving at any particular decision.
2) Distance Metric: Most often, the distance between two vastly different data, the distance measure adopted in the assessment of distance is the Euclidean distance measure.
3) To make predictions using KNN, we need to perform the following steps:
4) Calculate the difference of the distance between the new data point and all the data point in the training data.
5) Select k nearest neighbors on the basis of the distances which have been calculated.
6) The class label or the numerical value of one or the other is the value of the class label or the value of K nearest neighbors given to the new data point.

Here important note about KNN: Although it is simple and fairly easy to implement, it also has disadvantages. The estimation is highly sensitive to the value chosen for K and distance measure used in the algorithm. It also demands a big amount of memory space to accommodate the training data, although the computation of the distances between the data-points may take a lot if applied on big data sets. KNN has caught great attention in numerous applications such as image identification, textual categorization, and recommendation. It is also practiced with other computational methods like decision trees, neural networks in order to enhance the rates of forecasts.[5]

D. Naive Bayes Algorithm

Naive Bayes: An optimization technique applied when finding categories a problem. It is probability theory, referred to as Bayes' theorem that offers a means of calculating the probability of a particular hypothesis (in this case the class label) given the likelihood of the observed evidences (which in this case are the features or some attribute of the data) the observed evidence (in this case, the features or attributes of the data). Naive Bayes assume that the features are independent of each other given the class label or the number of feature vectors is large. This implies that the likelihood of a distinct profile of features is realized given one particular class label is found by the product of the likelihood of the different solitary features given that class label.

To apply the Naive Bayes algorithm, we need to define the following probabilities:

1) Prior Probability: The measure of the likelihood that each class label is associated with in the particular dataset.
2) Likelihood Probability: The particular likelihood that, in a given data set, each feature is going to be present in a certain class label.
3) Evidence Probability: The application probability for each feature to appear in the data set.

Naive Bayes is a simple and eager method that can handle enormous quantities of data and is resistant to obsolete characteristics. It does, however, presume that the traits are independent of one another, which may not always be the case in real-world circumstances. The Naive Bayes algorithmis extensively used in a variety of applications, including spam filtering, sentiment analysis, and document categorization.[6]

E. Decision Tree

Decision Trees (DTs) are actually trees that categorize instances according to one or many characteristics that is present in the data to be used for the model. In the decision tree each node represents a feature in the instance to classify, and each branch represents a possible value that the node can take.

Classification of instances proceeds through the root node and on through the branches until a class value is reached based upon feature values.

The learning process utilizes decision tree as the model of decision where mapping of observations of the item lead to conclusion on the target value of the item. The decision tree classifiers described in this paper use post- pruning in which the performance of the decision trees is tested, and unnecessary trees are eliminated using a validation set. Node elimination can be done on any given node, and the node can then be rewired to the most predominant of classes associated with instances connected to the node in the training phase.[7]

F. Random Forest

Random Forest is based on the principle of the boot strapping method that selects randomly a number of different trees that improving the pronosticative quality. Instead, it depends on the performance of decision trees, defining the relationship between them.

It intends to decrease the probability of error which arises from a single tree by making the trees independent of one another. Random Forest also tells where the input features are important so that they can claim a spot in predictions. The advantage of this over just one decision tree is that we can combine uncorrelated trees where several trees are constructed and the results are than averaged in order to provide a better and more accurate result. In short, using several trees together made possible because of uncorrelated trees works as an ensemble technique to reduce overfitting along with improved accuracy, making Random Forest a powerful tool in replacement of a standalone decision tree.[8]

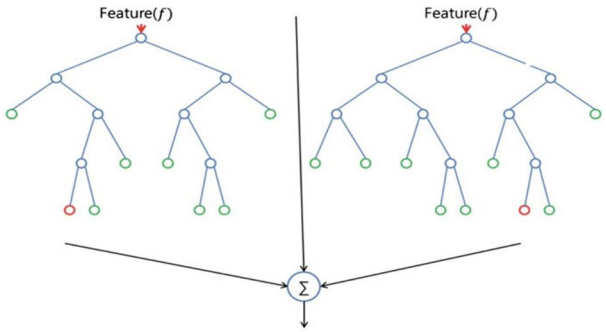


Fig. 5. Comparison of algorithms

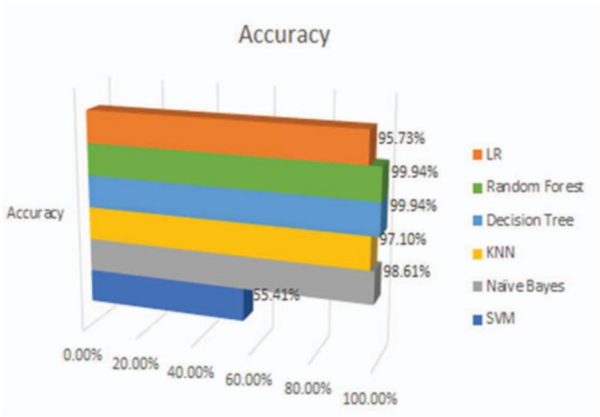


Fig. 6. Comparison of LR, Random Forest, Decision Tree, KNN, Naïve Bayes, SVM [9].

This paper comparatively assesses six applied methodologies under the machine learning category: Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest for identifying fraud in the online payment system. As much as pertains from the figure, these techniques were judged for performance based on their accuracy. The Decision Trees and Random Forests had very good performance, practically achieving 100 percent of accuracy, thus proving models suitability to work with complex relationships present in the data. It is also necessary to add that the Naive Bayes and the KNN algorithms behaved good giving high accurate values, which is suggestive to determine their efficiency in classifying transactions. Logistic Regression was able to perform as the primary basis for comparison between the methods, where SVM was slightly observed performing poorly compared to the other methods due to huge difficulty in handling the same dataset.

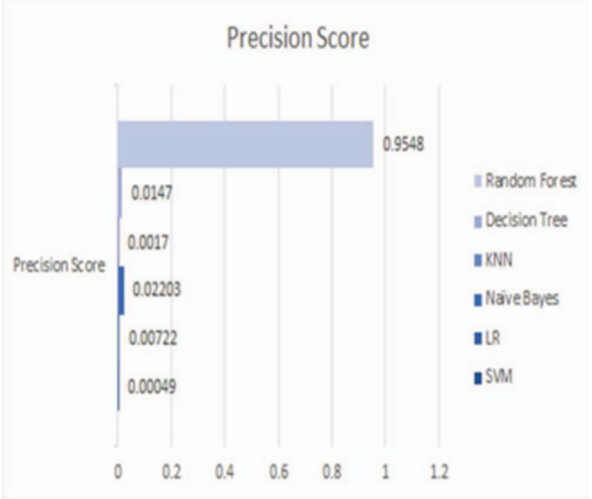


Fig. 7. Precision Score of Random Forest Decision Tree, KNN, Naïve Bayes, LR, SVM [9].

Precision values are given in figure 7 for SVM, LR, Naive Bayes, KNN, decision trees, and Random Forest. In the aspect of Precision, Random Forest had the best result which shows its efficiency in reduce False positives and classify the fraudulent transactions effectively. Similarly, Decision Trees and Naive Bayes had well 起 balanced precision scores. KNN took a moderate precision and both of them had reasonable efficiency in minimizing wrong prediction. SVM however was much slower in terms of precision, which shows the problem with the use of SVM when dealing with imbalanced data.[9]

In paper [6], the performance using different methods of machine learning for fraud detection is compared. The algorithms were run on a dataset, proving that the differences in their performance varied noticeably. SVM sor on average had low accuracy with as low as 0.7009 and committed as many as 8516 errors, which showed how hard it was for SVM sor to deal with the complexity of the dataset. On the other hand, for fraud detection, a better result was obtained using the Isolation Forest Algorithm which scored about 0.9974, with 73 errors only. The Local Outlier Algorithm also proved efficient as it had an accuracy of 0.9965 with 97 errors which poses its efficiency when handling intricate convictions as testified to by the myriad variety of patterns. This comparison thus sees it fit to point out that Isolation Forest is indeed better placed in the accuracy and management of errors compared to other methods available.[10]

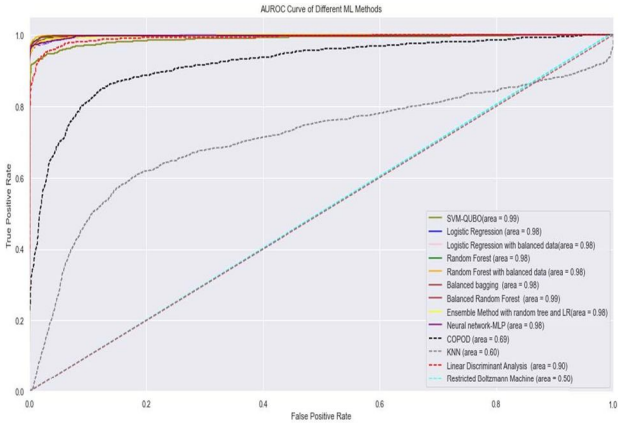


Fig. 8. AUROC curves of SVM-QUBO verses machine learning algorithms on ICCT dataset with no feature selection [11]

The study points out the difficulty of identifying fraud because of the unbalanced data distribution and the evolution of fraud schemes. It revisits the previous methods of Logistic Regression, Random Forest, and Neural Networks which are frequently used techniques but may drastically decline in efficiency if the data is not enhanced or the features are not selected properly. From the figure presented, the AUROC curves of various ML models in the ICCT dataset without feature selection are illustrated and notice that the proposed SVM-QUBO model achieve the highest AUROC value, which is equal to 0.99. Logistic Regression and Random Forest, which are the more classic algorithms, show certainly good results but slightly lower, which is evidence of efficient using quantum-inspired SVM- QUBO for this case. However, the survey also shows that the success of big data depends on the right combination of techniques and advanced algorithms in eliminating the frauds and reducing the level of false positives. [11]

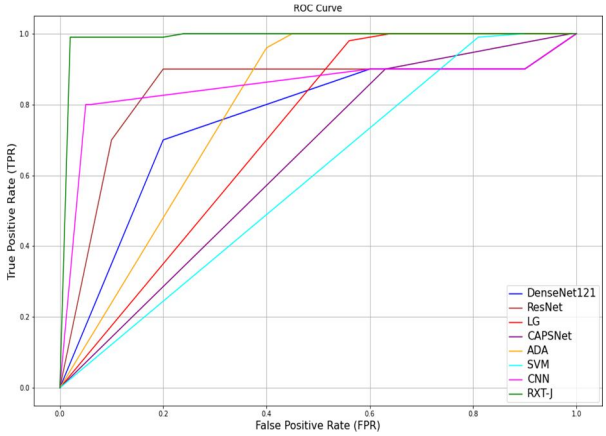


Fig. 9. ROC curve of the proposed method and existing methods on the IEEECIS dataset [12]

The study focuses on the difficulties of identifying fraud and the imbalanced and ever evolving nature of the data sets.It discusses a common set of ML approaches, including Logistic Regression, Random Forest, and Neural Networks, which were formerly used but had provides low or unsatisfactory performances without data pre-processing and feature selection. GRU and LSTM are both better suited at exploiting temporal dependencies, while concept drift is well handled by the more sophisticated SVM-QUBO coefficient optimized by quantum advisable heuristic. The result of the SVM-QUBO model is the highest AUROC of 0.99, which proves that the work's approach is more effective in dealing with datasets where the number of positives is significantly lower than that of negatives. Other methods of data balancing like SMOTE and feature engineering are also used to increase the accuracy and is a testament to the inherent capability of new generation models to provide a far superior results for accurate classification and eliminating false positives inherent in real- time fraud detection.[12]

A survey paper on the application of machine learning for online payment fraud detection reviews and discusses different ensemble and supervised methods for the domain. Graph has a high resistance to outliers and noise as it applies hyper parameters including the number of estimator, learning rate and max depth though it might lead to over-fitting if tuned wrongly. AdaBoost increases the accuracy of classification through the re-weighting of misclassified samples in the training data set with respect to weak learners such as the decision tree though it gets insensitive to noisy data and slows convergence rate. Bagging increases model resistance by repeating weak training of multiple independent models that are trained using bootstrapped training data, and could effectively handle high dimensionality. Random Forest goes beyond Bagging by growing decision trees with bootstrapped samples that employ a random selection of features; it performs well in relations between variables and with noisy data. SVMs are presented due to their capability of selecting appropriate hyperplanes to label classes, and their efficiency in managing non-linear and high-dimensional data through kernel functions at the expense of being sensitive to the selection of hyperparameters. These algorithms, used in fraud detection, image recognition and classification of text, respond to problems of imbalance in data sets, noise, and dynamic fraudulent behaviour; ensemble techniques usually show high-performance measures regarding increase in the efficiency and reliability of real schemes of classification.[13]

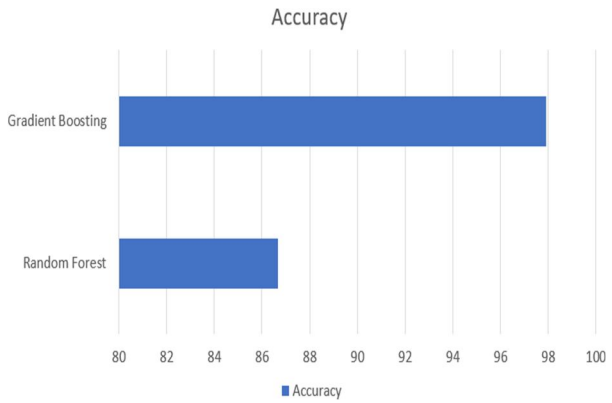


Fig. 10. Comparison of Result of Gradient Boosting and Random Forest [14]

The literature of the study on fraud detection using machine learning focuses on the various algorithms used in the detection of fraudulent transactions. Some of the skilled methods are Random Forest and Gradient Boosting which is quite famous for classification. Among them, Gradient Boosting has repeatedly shown higher results because of the possibility of sequential correction of the results and a good approach to unequal assortments, with such accuracy rates as 97,91% against 86,68% for Random Forest. These outcomes indicate that the ensemble method can improve the accuracy of fraud detection to a concerning level, despite problems of skewed datasets and dynamically changing patterns of fraud.[14]

Recent advancements in fraud detection highlight the efficacy of machine learning algorithms in enhancing classification performance. CSO-SVM has provided excellent classification, with an accuracy of 99.88% and sensitivity 99.42% compared with a specificity of 99.69% & AUC of up to 0.977. This model performs much better than all these models such as Logistic Regression 83%, Naïve Bayes 84%, Decision Trees 87%, Random Forests 88%, PSO 92.89%, and BOA 82.67% respectively. With feature selection through Competitive Swarm Optimization (CSO) and classification through SVM this paper surmounts issues of class imbalance and overheated computation setting the pace for efficient credit card fraud detection.

Method	Dataset	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
CSO-DCNN [7]	European credit card dataset	98.20	N/A	N/A	N/A
CNN [8]		97.27	N/A	N/A	0.966
Coarse KNN [9]		99.46	99.50	77.70	N/A
CSO-SVM		99.88	99.69	99.42	0.977

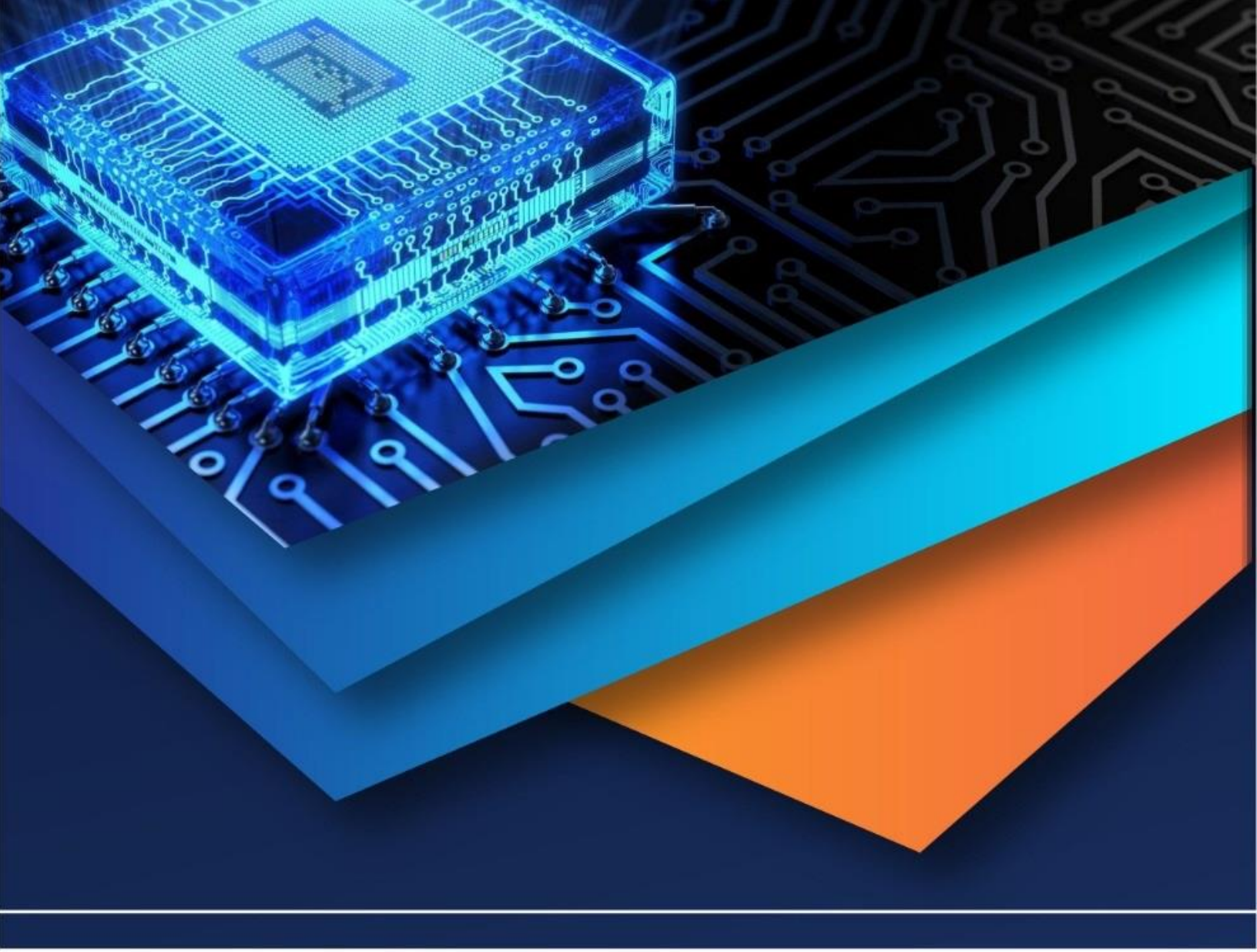
The table shows the current research on analysing fraud using machine learning has shown that optimization techniques improve the performance of classifiers. Based on the European credit card dataset, the authors used the CSO-SVM model which gave an accuracy of 99.88% and specificity of 99.69%, sensitivity of 99.42%, and the AUC of 0.977. Similarly, the KNN method obtained the accuracy of 99.46%, while the sensitivity of this method equalled 77.70% only. Other works like, CNN and CSO-DCNN achieved accuracy of 97.27% and 98.20% respectively, but they did not evaluate the performance of their models in other aspects. These results reaffirm the effectiveness of the proposed CSO-SVM model in terms of high accuracy and balanced performance measures in fraud detection frameworks. Thus, the proposed model presents a benchmark for the development of improved fraud detection models [15].

IV. CONCLUSION

In this paper, different methods of machine learning used in the online transaction fraud detection process have been discussed based on their efficiency and capability of dealing with the imbalanced and large dataset of the features involved. Here, SVM-QUBO, Logistic regression, KNN, Naïve Bayes, Decision Trees, and Random Forest each had their advantages and disadvantages. Although some algorithms achieved the highest speed and computation time, there were algorithms that offered a better distribution and higher accuracy and precision. The comparison of these algorithms is informative of the gain and loss functions of the real-world exposure of the scheme to the fraud detection system agenda and this study lays a good platform for the subsequent investigations.

REFERENCES

[1] Mr. Dipra Mitra, Dr. Shikha Gupta, Miss. Pavandeep Kaur, “An Algorithm Approach to Machine Learning Techniques for Fraud detection: A Comparative Analysis”, 2021, IEEE.
[2] Seyedeh Khadijeh Hashemi, Seyedeh Leili Mirtaheri, Sergio Greco, “Fraud Detection in Banking Data by Machine Learning Techniques”, 2023, IEEE Access.
[3] Sunita Singh Air, Mrs. Anubhooti Papola “Fraud Detection in Credit Cards Using Methods of Machine Learning”, 2024, Veer Madho Singh Bhandari Uttarakhand Technical University, Dehradun, Vol. 45 No. 3 (2024).
[4] Riham Muqattash, Faten Kharbat, “Detecting Mobile Payment Fraud: Leveraging Machine Learning for Rapid Analysis”, 2023, IEEE.
[5] Olawale Adepoju, Julius Wosowei, Shiwani lawte, Hemaint Jaiman, “Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques”, Global Conference for Advancement in Technology (GCAT), 2019.
[6] Vairam T, Sarathambekai S, Bhavadharani S, Kavi Dharshini A, Nithya Sri N, Tarika Sen, “Evaluation of Naïve Bayes and Voting Classifier Algorithm for Credit Card Fraud Detection”, 2022, IEEE.
[7] Maad M. Mijwil, Israa Ezzat Salem, “Credit Card Fraud Detection in Payment Using Machine Learning Classifiers”, (ISSN: 2321 - 5658) Volume 8- Issue 4, December 2020.
[8] Vandavasi Baba Mahesh, Koneru Venkata Sai Chandra, Lammata Shiva Prasad Babu, Velagal Aarthi Sowjanya, Dr Moulana Mohammed, “Clicking Fraud Detection for Online Advertising using Machine Learning”, KL University, 2024, IEEE.
[9] Darshan Aladakatti, Gagana P, Ashwini Kodipalli, Shoaib Kamal, “Fraud detection in Online Payment Transaction using Machine Learning Algorithms”, 2022, IEEE.
[10] Diksha Dhiman, Amita Bisht, Anita Kumari, Dr Harishchander Anandaram, Shaurydeep Saxena, Dr Kapil Joshi, “Online Fraud Detection using Machine Learning”, 2023, International Conference on Artificial Intelligence and Smart Communication (AISC).
[11] Haibo Wang, Wendy Wang, Yi Liu, Bahram Alidaee, “Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection”, 2022, IEEE Access.
[12] Abdulwahab Ali Almazroi, Nasir Ayub, “Online Payment Fraud Detection Model Using Machine Learning Techniques”, 2023, IEEE Access.
[13] Lochan S, Sumanth H V, Ashwini Kodipalli, Rohini B.R., Trupthi Rao, Pushpalatha V., “Online Payment Fraud Detection Using Machine Learning”, 2023, IEEE.
[14] U. Siddaiah, P. Anjaneyulu, Y. Haritha, M. Ramesh, “Fraud Detection in Online Payments using Machine Learning Techniques”, 2023, IEEE.
[15] Ashwini Gajakosh, R Archana Reddy, Myasar Mundher adnan, G Rajalaxmi, Pramodhini R, “Fraud Detection in Credit Card using Competitive Swarm Optimization with Support Vector Machine”, International Conference on Distributed Computing and Optimization Techniques (ICDCOT), 2024, IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY