Integrating Machine Learning Algorithms with Quantum Annealing Solvers for Online Fraud Detection

M. Sathyam Reddy[1], K. Soma Sekhar Goud[2], P. Sunil[3], D. Kruparao[4]

[1] Assistant Professor, [2, 3 & 4] Students

Department of Computer Science and Engineering,

Narasaraopeta Engineering College,

Sathyamreddym@gmail.com[1], somugoudkarpurapu@gmail.com[2], sunil.pulivarthi55@gmail.com[3], kruparao190@gmail.com[4]

ABSTRACT: The challenge of detecting fraudulent transactions in real time, proposing a novel framework utilizing quantum machine learning (QML) with Support Vector Machine (SVM) enhanced by quantum annealing solvers. Evaluating performance against traditional machine learning methods on bank loan dataset the results highlight the superiority of the quantum-enhanced SVM in both speed and accuracy for the bank loan data, while yielding comparable accuracy to other methods for credit card transactions. Feature selection significantly improves detection speed on dataset, albeit with marginal accuracy gains. This paper is proposed after careful study of algorithms like classification and regression, the algorithms that are useful for prediction model to get the best accurate value. With the quantum Annealing solvers, we are going to predict the fraudulent are non-fraudulent.

Key words: Cradit Card, Criminal Transactions, Quantum Annealing, Anomaly Detection, Quantum Algorithms

I. INTRODUCTION

Fraudulent transactions cost businesses a significant amount of money each year. In the US, businesses lose an average of $4 billion annually due to fraudulent transactions, while insurance companies in the UK face losses of around £1.6 billion [1] from fraudulent transaction claims. These losses not only include expenses for refunds, shipping, and other management costs but also result [2] in missed sales opportunities from trustworthy customers and damage to the company's reputation. The introduction of the paper Credit card fraud detection using machine learning [3] presented at the 4th International Conference on Intelligent.

Computing and Control Systems (ICICCS) in May Effective detection systems can help reduce these losses by identifying fraudulent transactions early. However, preventing and detecting fraud is challenging for several reasons. Firstly, the widespread use of mobile technologies has led to a substantial increase in online transactions, with a 110% rise in e-commerce transactions in the US alone in early 2020 compared to the previous year [4]. This surge in online activity has also led to an increase in web attacks targeting e-commerce retailers and associated fraudulent activities. Secondly, although there is a need for real-time or near real-time fraud detection [5] for online transactions, many existing systems are not fully effective as they only detect fraud after it has occurred.

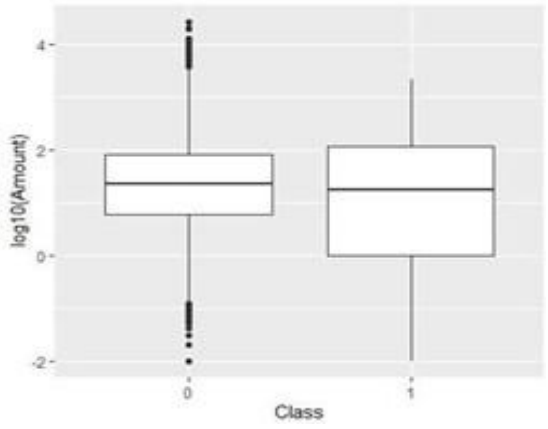

Fig 1: Distribution of amount on Normal (0) and Fraudulent (1)

The above figure (1) explains 0 and 1 of the box plot distribution of credit card transactions made by European cardholders in September 2013 [6]. It's not always straightforward to tell the normal transactions from the fraudulent ones because normal transactions can have extreme values. Fraudulent transactions often blend in with normal ones in terms of monetary amounts, making them

difficult to detect. To tackle these challenges, we're integrating quantum annealing solvers and machine learning algorithms for real-time fraud detection taken from the dataset. When dealing with online transactions, we encounter time series data, which can be either stationary or non-stationary. Stationary data remains consistent over time, while non-stationary data changes over time, showing trends and cycles [7]. Non-stationary data can be unpredictable, so we need to convert it into stationary data for modelling and forecasting.

II. LITERATURRE SURVEY

There have been several studies conducted on online fraud detection in various statistical and machine learning techniques. Here are some of the notable literature surveys:

In the research, numerous creators and Analysts have forced machine learning calculation to identify the sort of credit card extortion. We have detailed a few of the strategy within the writing. Maniraj [8] centres on information set examination and preprocessing, and the application of different inconsistency discovery calculations such as Nearby Exception Calculate and Segregation Woodland Calculation to PCA-transformed credit card exchange information.

Vaishnavi Nath Dornadula et.al [9] point to plan and create a novel extortion discovery strategy for Spilling Exchange Information, with an objective, to dissect the past exchange subtle elements of the clients and extricate the behavioural designs. At that point utilizing sliding window methodology, to total the exchange made by the cardholders from distinctive bunches so that the behavioural design of the bunches can be extricated individually.

The article "Upgraded credit card extortion location based on SVM-recursive include disposal and hyper-parameters optimization," created by N. Rtayli and N. Enneya[10] and distributed within the Diary of Data Security Applications in December 2020, presents an imaginative approach to reinforcing the discovery of credit card extortion through the application of machine learning strategies.

The paper titled "Nonstationary[11] time arrangement change strategies, An test survey," created by R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, and E. Ogasawara, and distributed in Knowledge-Based Frameworks in January 2019, gives a comprehensive audit of different strategies for changing nonstationary time arrangement information. Time arrangement information, which comprises perceptions recorded over successive interims of time, frequently display nonstationary behaviour, where measurable properties such as cruel and fluctuation alter over time.

The paper titled "Credit card fraud detection-machine learning methods," authored by Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, and Anderla A, presented at the 18th International Symposium INFOTEH-JAHORINA[12] in 2019, focuses on the application of machine learning techniques for the detection of credit card fraud.

The paper titled "Information awkwardness in classification. Test assessment," wrote by. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, and distributed within the Data Sciences diary in Walk 2020,[13] dives into the basic issue of information awkwardness in classification errands and gives an exploratory assessment of different procedures to address this challenge

PROPOSED SYSTEM

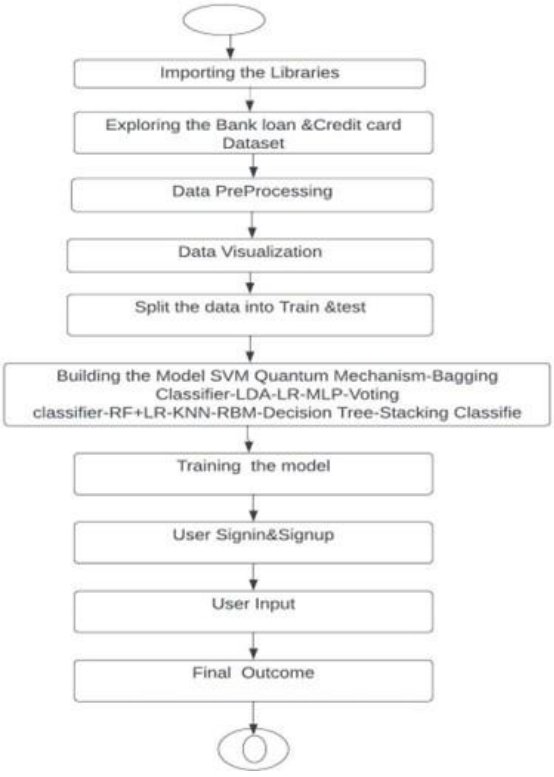

Fig 2: Activity Diagram

The above fig.2 portrays how framework's cycles stream. A movement graph has exercises, activities, changes, introductory plus last states, plus gatekeeper conditions, very much like a state chart Our Model is proposed based on certain criteria as follows:

A. Dataset Description:

The dataset holds information about credit transactions which has been made in a span of two days. The number of frauds has been calculated as 492 out of 284,807 transactions. The details have been given in form of positive and non-positive numerical values.

```
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

Fig 3: Data Description

The dataset contains 31 features which has been labelled as V1-V28 due to confidential reasons. The feature which has been revealed are Time and Amount of transaction.

B. Data Visualization

Data visualization is a way of showing data using graphics, such as charts, plots, infographics, and animations in below fig.(3)

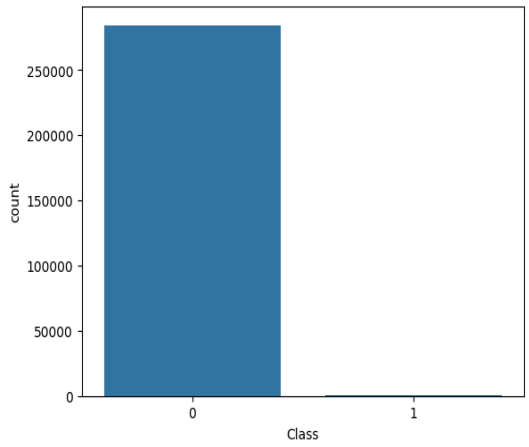

Fig 4: Approximate costing of online orders

C. Data Preprocessing:

Data preprocessing is a way of making raw data more suitable for analysis. It involves cleaning, transforming, and integrating data to make it more complete, consistent, and understandable. Data preprocessing helps to improve the accuracy and quality of data mining or machine learning results. In this project we have performed various preprocessing techniques which have helped clean the dataset into useful format.

III. EMPIRICAL FRAMEWORK AND METHODOLOGIES

We propose a extortion location system as appeared in Figure 3. The framework first verifies whether the input information is time series-based vs inactive, taken after by a stationary test to decide whether the time arrangement information are stationary or non-stationary. Since Expanded Anna Dickey Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) are two of the foremost commonly utilized factual test to analyse whether the arrangement of information is the stationary, this consider employments both tests [14], to assess whether the time arrangement information is stationary as appeared in Figure 5 through the unit root test. For nonstationary information. A few common detrending strategies such as control change, square root, and log change, will be connected to change over them into stationary. At that point the measurement decrease strategy is utilized to diminish the "noise" properties of the information.

Fig 5: Fraud Detection Frame Work

The machine learning approach of getting part capacities of SVM will be defined as QUBO, and after that the bit capacities recognized by quantum toughening solvers will be connected on prescient examination of extortion location. The execution of this QML extortion discovery framework will at that point be compared to the framework built with another conventional machine learning calculations. From the dataset it moved from the SVM

A. Random Forest

Having a place to the family of gathering strategies, Random Forest could be a broadly utilized machine learning algorithm. Outfit strategies combine different models to move forward the precision and Vigor of expectation [15]. By utilizing decision trees- straightforward models that fragment the feature space into littler locales based on input highlight

```
*Accuracy score for RF:  99.94382180275835

*Confusion Matrix for RF:
[[71083     8]
 [   32    79]]
```

Fig 6: Accuracy score of RF

The idea behind Random Forest is that by combining multiple decision trees, the model can capture a wider range of relationships between the input features, and the target variable.

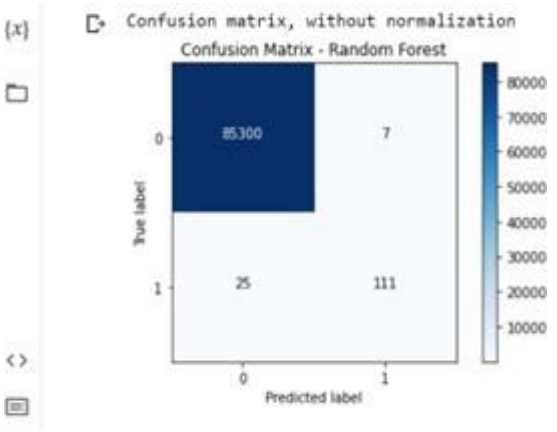

Fig 7: Confusion matrix for Random Forest

The thought behind Random Forest is that by combining numerous choice trees, the show can capture a more extensive extend of connections between the input highlights and the target variable.

B. Decision Tree

A choice tree could be a type of machine learning calculation that's used for classification and replace examination. It could be a tree-like shoe where each inner hub speaks to the result of the test, each department speaks to the result of the test, and each leaf hub speaks to a lesson name or numerical value decision trees are commonly [16] utilized in issues including classification, where the objective is to expect a categorical yield variable. By the by, choice trees have a inclination to overfit, especially in case the tree's profundity is over the top or the include number is substantial. Approaches like pruning.

```
*Accuracy score for DT:  99.91854161399961

*Confusion Matrix for DT:
[[71070    21]
 [   37    74]]
```

Fig 8: Accuracy score of DT

we predicted in the above figure assistance of our show one without any substitution which comes about in creation of dataset for each tree

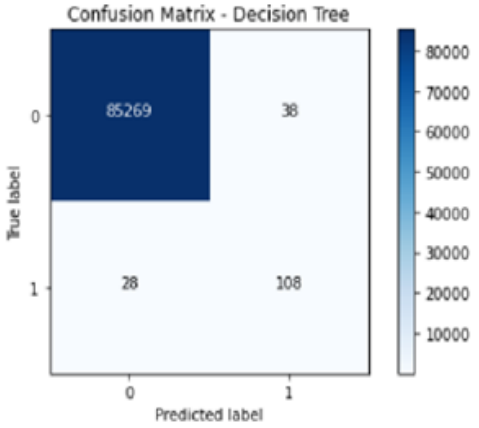

Fig 9: Confusion matrix for Decision Tree

By the above fig, choice trees have a inclination to overfit, especially in case the tree's profundity is over the top or the include number is substantial. Approaches like pruning and capping the tree's profundity can lighten overfitting. Besides [17], choice trees may not be the ideal determination for issues with interdependent features or circumstances with perplexing, nonlinear choice boundaries.

IV. RESEARCH DESIGN

The rise of extortion episodes makes it imperative to memorize more almost characteristics of datasets related with distinctive sorts of fakes. Such understanding will offer assistance to actualize and way better distinguish framework for extortion discovery. For this reason, this think about has chosen two datasets: Israeli cardholders' credit cards exchanges (ICCT) and bank credits application information [18]. The ICCT dataset is noontime-series based, contains 14,999 exchanges with 23.8D44fraudulent cases (3,571 in add up to) and 29 autonomous factors. For each 100 exchanges in this dataset, there are more than 23 extortion cases. The Advance dataset is time-series based, having 33,320 exchanges with 0.798% false exchanges (265 cases) and 122 free factors. Compared to ICCT dataset, in extra to being profoundly dimensional, the Advance information is 1000 cases [19]**.** Working with a period series-based, profoundly uneven, high-layered dataset, the results exhibit QML's

exceptional presentation as well as the convenience of our proposed extortion location methodology.

ACCURACY

Accuracy is a common metric used to evaluate the performance of a machine learning algorithms. According to It measures the proportion of correctly.

Model	Accuracy	Precisionn	F1 Score	Recall
LR	99.80%	39%	43%	41%
SVM	99.84%	83%	76%	79%
DT	95.91%	77%	66%	71%
RF	94.94%	90%	71%	79%

Table.1. Comparison of regression Algorithms

V. RESULTS

Both speed and precision comparison of SVM-QUBO and the twelve ml algorithms and conducted on each dataset. In terms of speed, the execution time for SVM-QUBO incorporates time for to make the preparing and testing records and envelopes, and (1) preparing and testing the modular. Untrue positive alludes to inaccurately distinguished the ordinary exchanges as false ones. In commerce, the fetched of a untrue positive frequently out weights a untrue negative. When a true-blue client is misidentified as a extortion, the negative involvement may lead to the misfortune of that client [20].

VI. CONCLUSION

QML has gotten a developing measure of consideration for its capability to resolve significant issues because of its computational capacities. To legitimize the speculation for potential execution gains, it is fundamental to recognize the main regions for QML application because of the challenges and cost of changing over issues into the QUBO design [21] expected for quantum processing. Fast and proficient misrepresentation identification is an incredible choice for QML arrangements because of the predominance of e-business, online exchanges, and the critical misfortunes brought about by fake exercises. We give a system identifying misrepresentation in such manner. Prior to deciding if the information is fixed or non-fixed, this structure will initially decide if it depends on time series.

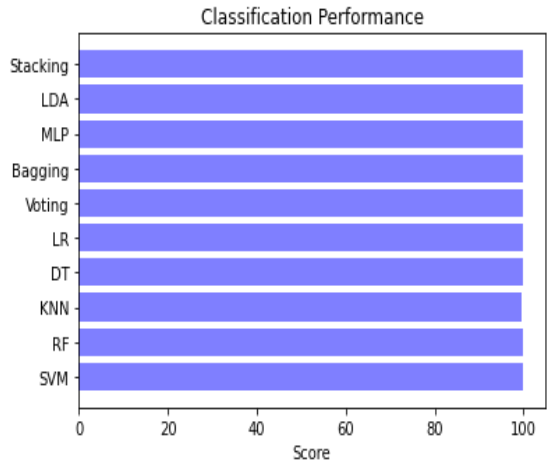

Fig 10. Classification Performance

The above figure utilized in this study is a critical stage toward constant misrepresentation discovery in light of its incredibly quick recognition speed. We are fostering a test system that will create datasets with fluctuating proportions of standard exchanges to fake exchanges as a feature of our continuous examination. The combined datasets utilize the boundaries from the benchmark examples that were utilized in this examination.

REFERENCES

[1] S P, Maniraj& Saini, Aditya & Ahmed, Shadab & Sarkar, Swarna, “Credit Card Fraud Detection using Machine Learning and Data Science”, International Journal of Engineering Research 2021 and. 08. 10.17577/IJERTV8IS090031.

[2] D.Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2019, pp. 1-5

[3] L. Columbus. (2020). How E-Commerce’s Explosive Growth is Attracting Fraud. [Online].

[4] LexisNexis. (2020). Lexis/Nexis Solutions for 2020 True Cost of Fraud Study: ECommerce/Retail Edition.[Online]

[5] R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, and E. Ogasawara, ‘‘Nonstationary time series transformation methods: An experimental review,’’ Knowl.-Based Syst., vol. 164, pp. 274–291, Jan. 2019.

[6] F. Itoo, Meenakshi, and S. Singh, ‘‘Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection,’’ Int. J. Inf. Technol., vol. 13, pp. 1503–1511, Feb. 2021.

[7] N. Rtayli and N. Enneya, ‘‘Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization,’’ J. Inf. Secur. Appl., vol. 55, Dec. 2020, Art. no. 102596.

[8] P. Hájek and R. Henriques, ‘‘Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods,’’ Knowl.-Based Syst., vol. 128, pp. 139–152, Jul. 2017.

[9]. S.Dhankhad, E. Mohammed and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 122-125

[10] Kaithekuzhical Leena Kurien and Dr. Ajeet Chikkamannur, " Detection and Prediction for Credit card Fraud Transactions using Machine Learning," International Journal of Engineering Sciences & Research Technology, V(8), n(3), march 2019, pp.199-208

[11] F. Itoo, Meenakshi, and S. Singh, ‘‘Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection,’’ Int. J. Inf. Technol., vol. 13, pp. 1503–1511, Feb. 2021.

[12] R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, and E. Ogasawara, ‘‘Nonstationary time series transformation methods: An experimental review,’’ Knowl.-Based Syst., vol. 164, pp. 274–291, Jan. 2019.

[13] A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, ‘‘A customized classification algorithm for credit card fraud detection,’’ Eng. Appl. Artif. Intell., vol. 72, pp. 21–29, Jun. 2018.

[14] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, ‘‘Data imbalance in classification: Experimental evaluation,’’ Inf. Sci., vol. 513, pp. 429–441, Mar. 2020.

[15] D. Willsch, M. Willsch, H. De Raedt, and K. Michielsen, ‘‘Support vector machines on the D-wave quantum annealer,’’ Comput. Phys. Commun., vol. 248, Mar. 2020, Art. no. 107006.

[16] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, ‘‘Credit card fraud detection using machine learning,’’ in Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS), May 2020, pp. 1264–1270.

[17] Asha RB, Suresh Kumar KR, “Credit card fraud detection using artificial neural network”, Global TransitionsProceedings,Volume2,Issue1,2021,Pages3541,ISSN2666285X.

[18] Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection-machine learning methods. In: 18th international symposium INFOTEH-JAHORINA (INFOTEH); 2019. p. 1-5.

[19] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Proc Comput Sci. 2019;165:631–41

[20] Abhishek L. Optical character recognition using ensemble of SVM, MLP and extra trees classifier. In: International conference for emerging technology (INCET) IEEE; 2020. p. 1–4

[21] Mohamed Jaward Bah, Mohamed Hammad “Progress in Outlier Detection Techniques: A Survey” Hongzhi Wang, of the 2019 IEE. Aibus, S. et al. 2007. Application of Classification