Secure Machine Learning for Fraud Detection Using Encrypted Online Payment Transactions

Ratna Pillai, Nawaz Sheikh, Naumaan Kazi, Digvijay Rai
x18134297, x18134637, x18130208, x18134645
Advanced Data Mining - Group A
MSc in Data Analytics

***Abstract*— In today's digital world, most of the transactions are online and one of the prominent issues which customers and companies are facing is fraudulent transactions. Fraudulent transactions have been a troublesome for both the customers and companies. With the advancement in technology in the recent years, many industries have been using data mining and machine learning techniques to predict or forecast the future unforeseen events using historic data. Although, this is a boon for many industries, sectors like healthcare, finance and areas where confidential information of customer is handled, it becomes difficult to outsource the data for analysis due to data regulation issues. In order to overcome this scenario, data mining field has also introduced many methods to analyze the sensitive data without having to worry about the confidentiality issues. This paper focuses on applying machine learning techniques on encrypted data for effective classification of fraudulent transactions.The novel approach followed in this paper is application of hyper parameter tuned algorithms for Gaussian Naive Bayes, Light GBM and XGBoost apart from conventional classification alogrithms such as SVM and Logistic regression to detect frauds on already encrypted data. The models will be evaluated using f1 score, AUC and kappa scores to derive the best performing model.**

Keywords: Credit Card Frauds, Data Mining, Homomorphic Encryption, Paillier

I. INTRODUCTION

In a fast-paced life, customers prefer online banking over offline banking to a great extent. This is only possible due to the advancement of technology in the form of application development using mobile platforms. This enables the users to perform real-time tasks with a single touch which saves time and energy rather than following conventional methods. The transaction volume of payments has increased and credit card companies have been rapidly growing over the years. Paypal Inc has processed the total payment volume of 143 billion USD in the third quarter of the year 2018. Frauds in online payment have become more common due to the rapid expansion of the online payment marketplace. [1] There are many ways in which online payment fraud may occur, but the most common one is unauthorized use of online payment with a debit card or credit card.

A fraud detection system should be sufficient enough to detect transactions which are fraudulent with high efficiency and accuracy. Although it is essential to avert corrupt players from implementing fraudulent transactions, and also make sure that genuine users are not getting interrupted from gaining access to the payment system. Nowadays, companies, industries, and other private institutions encrypt the data before storing in order to maintain data confidentiality and to get access of it using analytical and computational services. A Fully Homomorphic Encryption (FHE) technique permits the analytical and computational facilities to be applied on encrypted data.

One approach to protecting data confidentiality is to encrypt it before outsourcing it to the third party. This way, users of the data may be limited, but recent advancement in technology helped us to perform operations on encrypted data without decrypting them. The FHE is an encryption technique that allows random processes on the ciphertext. Gentry constructed the first FHE technique, and following procedures have quickly turn out to be to more practical, with improved parameters and performance. Gentry's technique and many of the following FHE techniques are known as Somewhat Homomorphic Encryption (SHE) technique as a fundamental building block, and use of that method is called bootstrapping to expand SHE to an FHE technique. SHE techniques implements multiplication and addition on encrypted data, but it is only limited to a small amount of computation which it can perform, as encryption includes the accumulation into ciphertexts in terms of noise. Homomorphic operations on ciphertexts allows the innate noise terms to expand, and these noise terms should not surpass a specific bound. However, at this moment, the effectiveness and pace of these computations are extremely low, causing performance issue while performing FHE through its best capacity.
Unlike performing Principal component analysis for dimensional reduction, it has also been used in encrypting the source data by maintaining the same variance as that of the original data. Since FHE is still not practical, substitution techniques such as usage of PCA acts as a better option. This project aims at applying multiple machine learning algorithms on a data that is already masked using PCA. Class imbalance is handled using undersampling tehcniques and k-fold validations have been further applied for sampling of training and testing data. Classification algorithms such as XGBoost, Logistic Regression, Support Vector Machine, Random Forest, Guassian Naive Bayes and Light GBM have been implemented for deriving the best performing models.

The models are evaluated using F1 score, (Area Under Curve) AUC and Kappa scores. In addition, these models have been optimized using hyper parameter tuning.

A. *Research Question:*

"*To what extent can we accurately classify the credit card payment transactions as fraud using encrypted data with all masked fields?"*

B. *Research Objectives:*

Following objectives are aimed to achieve based on the research question:

- To observe the importance of synthetic data in predicting the fraudulent transactions and applying the same methods on real-world back data.
- To perform classification using conventional models such as Logistic regression and Support Vector Classifier.
- To perform and observe the latency of classification using advanced data mining models such as Light GBM, XGBoost and Random Forest classifier.
- To optimize the best performing model using Hyper parameter optimization.

With this research question, the banking industries will be benefited to a great extent as they will be able to share the encrypted versions of data on which a data analyst can apply techniques like PCA to transform the masked information into numerical values. Also, a data analyst will be able to perform the mining of knowledge from real transaction data. This paper is further structured as following: II) Related work, III) Research Methodology, IV) Evaluation & Results V) Conclusion and Future Work

II. RELATED WORK

A. *Conventional fraud detection studies*

In the past, many researchers have conducted studies for detecting frauds in financial transactions using machine learning. These studies have been implemented on different areas in finance such as credit card or debit card payments, bank card enrollments, mobile payment transactions and online payments [2]. The chief executive member of Future Analytics in Belgium has evaluated a case study on fraudulent transaction detection using machine learning and artificial intelligence [3]. The author has stated that the performance of the underlying model can be significantly improved with new features added to the original data. The ideal steps to make the model more useful in the case study evaluated was to first explore the original features in the dataset, identify the most relevant features using techniques such as PCA and then study the true positive rate classifications which is also called as sensitivity or recall as this metric clearly measures the number of fraudulent transactions rightly identified as fraud. The author indicated that the recall should be maximized and the true negatives rate should be minimized to achieve a best performing model. Further the performance of the classifier was evaluated using characteristic such as ROC (Receiver Operating Characteristic) which was 95% with the model having new features added to it. Combining both unsupervised and supervised ML techniques, the author [1] has implemented feature selection to examine the impact of the predictor variables on fraud occurrence and then application of classification models using multiple algorithms. Using conventional algorithms resulted in a poor F1 score of 0.49 and hence the author had to switch to regression models for improved results. On the other hand, an interesting study was conducted by [1] to detect fraudulent transactions when the debit card/credit card details are entered by the customer for using mobile wallet. This process is called bankcard enrolment and for this process, the authors have implemented a classification model on the China Union pay dataset which comprises of multiple device platforms for mobile payments such as Huawei, Samsung, Apple, Xiaomi and Meizu to detect frauds using machine learning algorithms like gradient boosting decision trees, random forest and logistic regression. Based on the experiments carried out, XGBoost algorithm performed the best with an average recall of 75%, which improved the performance by 22. Paysim is a simulator developed by the author where using legitimate data is a big problem [4]. The dataset of mobile money transaction used for fraud detection is not available easily due to data privacy constraint. This simulator was developed to remove the burden of research which used to harness these data and to perform the research. By this it simulates real type data it is very similar to the original data-sets. The techniques used here are agent-based simulation and other statistical mathematics. The attributes it simulated where CASH-IN, CASH-OUT, TRANSFER, PAYMENT and DEBIT. It concludes by generating synthetic data that are used for detection of fraud. The data can aid many organizations like government, school and financial sectors. Using this the privacy of the customer will always be maintained. In finding fraud for financial sector machine learning is a very popular approach. This similar approach is presented by [5] where the author uses Logistic Regression and the Support Vector Machine for more accurately finding fraud in payment transactions. As the data was taken from Kaggle the same data generated from study [4] for the analysis. The data here was highly imbalanced. The data was split in three 70% for testing, 15% for testing and remaining 15% for CV. The factors chosen were Payment, debit, cash-in, cash-out and transfer where to check the performance we have used precision, F1-Score and Recall. Looking at the results logistic Regression performs the best and whereas Linear SVM gives less score. In the following research paper fraud is detected by behaviour analysis [6]. For this tool is used in prediction of security at dynamic. The use of predictive security analyzer is used, and the goal of this study is to find any patterns in money laundering in synthetic way where the behaviour is on the same properties on real life transaction. The PSA would take the real time information from the execution environment, security space and the models. Here normal ROC is not used but higher performing EPC configuration is used In order to reduce the fraud for prediction it is necessary to use feature to select false positive. From the study [7] we observe that

Author	Model Used
[7]	Deep Feature Synthesis algorithm
[9]	Logistic Regression and Gradient Descent
[10]	Hyperplane Decision, Decision Tree and Navie Bayes
[11]	Homomorphic Algorithm
[12]	Full Homomorphic Classify using Navie Bayes
[13]	Support Vector Machine
[6]	Predictive Security Analyzer (PSA)
[4]	Not Applicable

one in five cases are actually fraud and also one in six customers had valid transaction declined. For addressing this the research has used Deep feature Synthesis which will automatically give the behavioural features which are on transaction as they are historical data. The classifiers used are random forest and the research has got 237 features on every transaction. The model was tested on the data which was for the multinational bank and did comparison of the solution. The study has successfully got the false positive by 54% and also saved thousands of euros from the point of stakeholders. Using supervised machine learning in detecting the fraudulent transaction in credit card with using real life data [8]. As it is known that fraud in credit card data I highly imbalance. The author had employed algorithms to implement using ensemble learning method a super classifier. The novelty brought to this study were inclusion of variables which might improve the accuracy in finding fraud. The approach used were both supervised and unsupervised learning methods. There were numerous machine learning algorithms used by the authors where Logistic Regression is the most promising followed by random Forest and XGBoost classifiers. As the under sampling was done to balance the unbalance data. Based on these past studies, it is evident that much of the past research have been done on synthetic data due to unavailability of the real time data because of data regulation and privacy issues. Hence the need for encryption arises in this field to analyze the data for deducing patterns belonging to the domain are unable to share the sensitive data.

B. Homomorphic Encryption Studies

Unlike the authors [3] and [1] who have just applied algorithms to classify fraud transactions, [14] takes in to consideration that there are privacy issues associated as the banks are compromising in sharing the customer data to data analysts and it can lead to more vulnerability in the form of data leakage or stealing of the potential data. These security challenges have been highlighted from the data mining perspective by the authors and to prevent this security compromise, data mining uses the privacy preserving data mining (PPDM) technique. Also, there are some important security methods like encryption and decryption which are suggested by the authors to conduct on the datasets to prevent any unauthorized usage. In addition, they have also implemented Naıve Bayes algorithm to predict the class level to produce effective and quick results. Similary, [15] also highlights that the data privacy concerns have to be taken care as the predictions performed on the data might contain some confidential information. They have implemented a Homomorphic encryption technique on the data and built a logistic regression and gradient descent model on a set of real numbers to accurately classify the data. With the descent model implementation, the performance of the model was significantly improved which takes 3.6 minutes to process the 0.02 GB data. Alternatively, [16] have studied the application of l2 logistic regression on the encrypted data using Newton Raphson methods and Shamir's secret key encryption techniques. They have proposed an architecture where the respective business users will aggregate the data and encrypt them for the data mining algorithms to be further performed. For their experiments, they have created synthetic data sets for Insurance, parkinson's disease and genomic data. Additionally, they have also proposed to encrypt the intermediate results such as the summary statistics obtained so that the privacy breach does not occur. Accurate results have been generated and evaluated using the R^2 statistic by comparing the correlations of encrypted data with the real world framework.

On the other hand, author [17] has projected two non-synergistic privacy conserving classification models and executed them using a Fully Homomorphic Encryption based privacy retaining Naıve Bayes classifier. This proposed technique reduces the communication between the user and the server, which makes it more secure during the protocol execution process. Also, this proposal is robust to cyber-attacks using quantum computers. The outcomes of this result show that the two projected models require 70-90 seconds approximately to categorize test datasets. Even though this classification speed might not seem to be practical, but it can be achieved with the advancement in Fully Homomorphic Encryption technology. In this paper, the author [18] has proposed a homomorphic encryption technique which will be applied on a publicly available data to maintain its integrity. The homomorphic encryption method's computational complexity was primarily based on the number of multiplications to be performed on encrypted data. Confidential Machine Learning technique was used for binary classification which was dependent on least-squares solutions of polynomial approximations attained from a few steps of gradient descent. The researcher aimed to show that machine learning algorithms can be applied on encrypted data, and confidentiality of data can be retained during training and testing of the data. In this paper, the author [19] has projected a privacy-conserving cloud-based and biometric verification system for multiple-party, which depends on pre-trained DNN (deep neural network) to carry out feature extraction using transfer learning. It does feature extraction from fingerprint and iris images for the biometric verification and fake/actual detection. CNNOptLayer, a novel algorithm, was used for optimizing the process. The Paillier Chunckwise was used to encrypt and bit-mask the biometric features. The outcome shows that the system achieved a verification F-measure score of 95.47% with zero false positives when combined features of fingerprint and iris inputs were verified. Similarly, [11] have also stated there are ample

amounts of risks in computing when a work is deployed to a third party which leads to risks of authentication, security, confidentiality and privacy factor. Homomorphic Encryption is a state of the art which erases the problems of confidentiality and privacy of the data which is stored on the cloud. Users gets the ability of performing computational operations on cipher text originating an encrypted result which can be decrypted later on to fetch the results. RSA Cryptosystem and Paillier Cryptosystem also known as Multiplicative Homomorphic Encryption and Additive Homomorphic Encryption respectively are the cryptographic techniques and are widely used public key cryptosystems. If any one of the cryptographic technique is used, then it leads to the encryption being partially homomorphic encryption. Whereas, it is a fully homomorphic encryption when both the cryptographic techniques are used together. [9] The costs of computation with HE-based solutions is relatively high as there are many number of iterations in the algorithms used for optimization naming gradient descent (GD). Binary classification can be done using ensemble GD for logistic regression. MNIST and Credit dataset were taken from UCI machine learning repository in order to implement ensemble GD for logistic regression. The iteration count reduces to 60% when the above stated method is used. Along with this, running time is reduced to 60-70% on the total procedure on the encrypted data and the storage size is also reduced to 30-40%. Random initials are allotted to the partial datasets iterating them in a series and the average is taken to get the final result of the experiment. [13] Machine learning task has been addressed which uses encrypted training data. Three parties naming Data Owners, an Application and the Authorization Server were included in the model. Private data is owned by the data owners. Application trains the model to apply the machine learning algorithm on it. Partial homomorphic cryptosystem is handled by the Application so that encrypted data gets access and computations are handled. Support Vector Machine (SVM) is selected because it gives excellent performance in supervised classification. Two SVM training algorithms naming SMO and PEGASOS are implemented for benchmarking purposes. An unsupervised clustering stage has been proposed which is named as semiparametric SVM scheme. Classification of tumors as malignant or benign by keeping the patients data private has been proposed. [12] Private Naive Bayes classification is designed using fully homomorphic encryption (FHE) to compute encrypted medical data in order to keep the patients data private. Accuracy observed for unencrypted and encrypted data is 96.03% with a precision of 95.10%. Naive Bayes, Decision Tree, Linear SVM, k-Nearest Neighbors and Logistic regression are the algorithms used to classify the breast cancer as malignant or benign. The highest accuracy has been achieved by Linear SVM and has the value 97.06%. [10] Classification is widely used in genomics prediction, financial prediction, face recognition and spam detection. It is very important to keep the data private. Hence, hyperplane decision, decision trees and Naive Bayes protocols are used to classify the data. AdaBoost is combined with all the three protocols. The models were trained using sci-kit learn and the five datasets are medical datasets. The classifiers are implemented in C++ using Boost, Google's Protocol Buffers, HELib and GMP for the implementation of Fully Homomorphic Encryption. Evaluation is done using building blocks library which used multiplexer classifier and Viola and Jones face detection. Efficiency of the classifiers and library were demonstrated on real datasets. As we know client data is sensitive in nature and it raises many concerns over its privacy. Where client uses cloud services to store and analyse its data is called as MLaaS (Machine Learning as a Service). The author [20] conducts this research to solve this issue using deep neural network. For the classification of MLaaS and also it has it focus on using CNN technique. It has a task of achieving this by twerking the current CNN method. The first step taken was using the activation function like Tanh, ReLU and Sigmod. Low degree of polynomials was also used for efficiency and practical usage. The CNN was also train in such a way it is approximate with the LDP that original activation function. There was 99.25% accuracy obtained using MNIST optical technique and this score is a start-of-art. And also achieved 164000 prediction per hour.

Author	**Models Used**
[2]	Naïve Bayes, Lositics Regression, OneR, Decision Tree, C4.5 Algorithm, Random Forest
[1]	Gradient Boosting Decision Tree (GBDT) algorithm , Logistic Regression, Random Forest
[3]	Artifical Intelligence
[19]	CNN
[14]	Bayesian classification algorithm
[15]	Logistic Regression
[18]	Linear Means Classifier, Fisher's Linear Discriminant Classifier
[15]	Navie Bayes, Hyper-Plane and Decision Tree
[16]	Logistic Regression and Newton-Raphson Method

III. RESEARCH METHODOLOGY

Methodology plays an important role in solving a research problem systematically and logically. Since the research problem is based on the financial industry, as a de-facto CRISP DM (Cross Industry Standard Process) has been selected.

A. Business Understanding:

Before starting any implementation, it is important to understand the business requirements thoroughly. For this research project, we have already covered the business requirements with respect to fraud detection in financial transactions in the related work section II

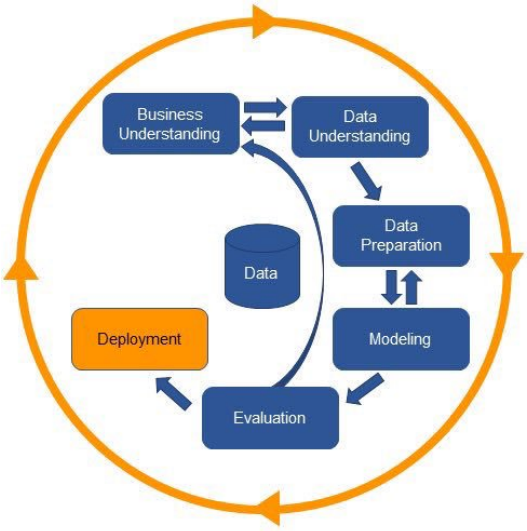

Fig. 1: CRISP DM for Credit card Fraud Detection

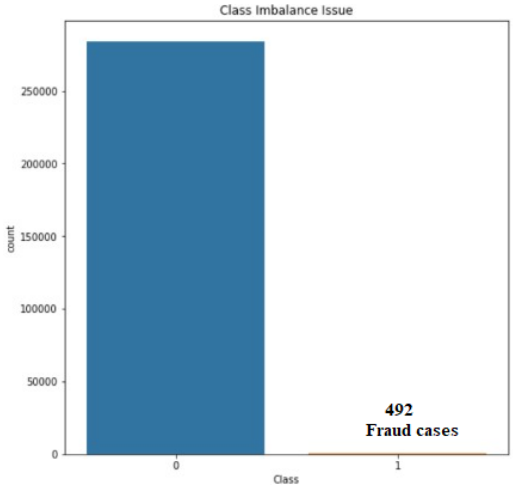

Fig. 2: Class Imbalance

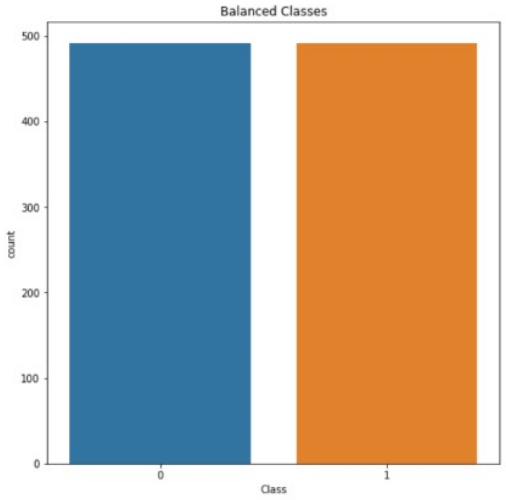

Fig. 3: Class Balance

B. Data Understanding:

The credit card fraud data comprises of 2 days transaction details of card holders in Europe with 2,84,807 rows. The data considered for this implementation addresses two requirements of any financial industry. Firstly, the dataset in consideration has masked information about the transaction details with only Transaction Amount, Class determining whether the transaction is fraud or not and the elapsed time between multiple transactions. Masking of the remaining 28 columns have been done as a part of maintaining the confidentiality of transaction data. Techniques like Principal Component Analysis (PCA) have been applied on the data to transform it and make them available as simple numeric values for any researchers to perform any analysis. Secondly, with the class field in dataset, machine learning algorithms can be applied to achieve the classification results. Thus, the dataset already is fully homomorphic encrypted and hence is ideal for this data mining project.

C. Data Preparation:

As all the predictor variables are already transformed and encoded with numeric values using PCA, they cannot be transformed any further. However, class imbalance and k-fold cross validations have been performed as a part of exploratory data analysis.

1) Handling Class Imbalance & 10 fold validation: Out of the two hundred thousand rows, only 492 rows belong to fraudulent transactions which indicates the the class is highly imbalanced. Below Figure 4 shows the under sampled classes.

Hence, random under sampling has been done on the data to balance the data for building classification models. Below Figure 4 shows the classes after balancing.

2) Training Model:: For fraud classification models, we have applied the 10 fold cross validation and split the training and testing data in 80:20 ratio.

D. Modelling:

1) Logistic Regression: Logistic regression is a basic classification algorithm which can be applied when the dependent variable is categorical. It works on the basis of hypothesis that how accurately the dependent variable is classified for the given independent variables as an input.

2) Random Forest: Random forest is evolved from decision trees which is also a classifier.[1] Classification is provided for the input data by each decision tree where the random forests chooses the most voted predictions as the result. A large number of weak classifiers enables a random forest to form a strong classifier.

[1]https://www.sciencedirect.com/topics/engineering/random-forest

TABLE I: Classification Summary - Without parameter tuning

Without Parameter Tuning							
Algorithms	F1 Score	Precision	Recall	AUC	Kappa Score	Accuracy	MPE
Gaussian Naive Bayes	88.48%	96.05%	82%	89.62%	80.27%	90.46%	9.60%
LightGBM	94.38%	94.38%	94.38%	94.87%	89.75%	94.92%	5%
XGBoost	94.97%	94.44%	95.50%	95.43%	90.78%	93.01%	4.50%
Support Vector Classifier	94.25%	96.47%	92.13%	94.67%	89.71%	90.98%	5.07%
Random Forest Classifier	94.25%	96.47%	92.13%	94.67%	89.71%	94.92%	5.07%
Logistics Regression	92.57%	94.18%	91.01%	93.19%	86.63%	92.75%	6.50%

3) Support Vector Classifier: A supervised machine learning algorithm used for both regression or classification constitutes a Support Vector Machine.[2] Each data is plotted as a point in the n-dimensional space (where number of features is n) as each feature is a value of a particular coordinate. By finding the hyperplane that differentiate the two classes, classification is performed.

4) Gaussian Naive Bayes: Classification can be done using Naive Bayes algorithm.[3] Real attributes can be extended using Gaussian Naive Bayes.To summarize the distribution with real-valued inputs, mean and standard deviation is calculated of input values.Gaussian Probability Density Function assists in finding the probabilties of the input variables.

5) XGBoost: XGBoost is a decision tree based machine learning algorithm which uses a framework of gradient boosting.[4] Decision tree-based algorithms are considered to be best when working with structured or tabular data. System Optimization and Algorithmic Enhancements are the key features in the improvement of XGBoost. Bias variance tradeoff is a parameter in XGBoost.[5] Execution speed and model performance are the two major implementation reasons in XGBoost.

6) Light GBM: A tree based learning algorithm is used by LightGBM which uses gradient boosting framework.[6] LightGBM enables faster training with higher efficiency, providing better accuracy with support of parallel and GPU learning and capability to handle large-scale data.

IV. EVALUATION OF RESULTS

For this research, we have evaluated classification algorithms using F1 score, Kappa score, Precision, Recall, Mean Percent Error and 10 fold cross validation accuracy score. Summary of results are presented in I.

[2]https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[3]https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[4]https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[5]https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[6]https://lightgbm.readthedocs.io/en/latest/

1) Logistic Regression: All the variables in the data set except dependent variable have been used for classifying the fraud transactions.

The performance of the basic logistic regression model evaluated using confusion matrix was significant score with (Accuracy: 92.7%, Kappa: 88.6%, Recall: 91.01%, Precision: 94.1% and Error Rate: 6%)

2) Random Forest: We have applied random forest on the training model using all the variables.

Performance of Random Forest obtained are (Accuracy: 92.1%, Kappa: 89.7%, Recall: 91.01%, Precision: 96.4% and Error Rate: 5%)

3) Support Vector Machines: We have used Linear SVC to predict the performance of the SVC.

SVC performance is also good with scores (Accuracy: 90.9%, Kappa: 89.7%, Recall: 92.1%, Precision: 96.4% and Error Rate: 5%)

4) Gaussian Naive Bayes: Performance of the Gaussian Naive Bayes model is evaluated with scores (Accuracy: 90.4%, Kappa: 80.2%, Recall: 82.02%, Precision: 96.05% and Error Rate: 9%)

5) XGBoost:: XGBoost implementation on the data provided scores (Accuracy: 95.4%, Kappa: 90.7%, Recall: 95.5%, Precision: 94.4% and Error Rate: 4%)

6) LightGBM:: Performance obtained with Light GBM with scores (Accuracy: 94.9%, Kappa: 89.7%, Recall: 94.3%, Precision: 94.3% and Error Rate: 5%)

A. Hyper Parameter Tuning:

Hyperparameter is a parameter that cannot be learned from regular training process. GridSearchCV and RandomizedSearchCV are the two best strategies for hyperparameter tuning.[7] Out of the six algorithms used, XGBoost and LightGBM with best results. However, after tuning, Light GBM generated better performance results.Table II shows the summary of classification results with tuning. Below are the best hyperparameters which are selected and implemented on LightGBM model.

[7]https://www.geeksforgeeks.org/ml-hyperparameter-tuning/

TABLE II: Classification Summary - With parameter tuning

With Parameter Tuning							
Algorithms	F1 Score	Precision	Recall	AUC	Kappa Score	Accuracy	MPE
LightGBM	94.91%	95.45%	94.38%	95.33%	90.76%	95.43%	4.5%
XGBoost	91.62%	91.11%	92.13%	92.36%	84.64%	93.52%	7.6%
Support Vector Classifier	92.65%	93.18%	92.13%	93.28%	86.66%	92.925%	6.59%
Random Forest Classifier	93.33%	92.30%	94.38%	93.95%	87.72%	93.52%	6.09%
Logistics Regression	92.13%	92.13%	92.13%	92.82%	85.65%	93.26%	7.10%

- subsample : 0.5
- random_state : 501
- objective : 'binary'
- num_leaves : 65
- min_split_gain : 0.7000
- min_data_in_leaf : 3
- metric : 'accuracy'
- max_depth : 8
- learning_rate : 0.03
- colsample_bytree : 0.9
- boosting_type : 'gbdt'

B. Visualization of Light GBM

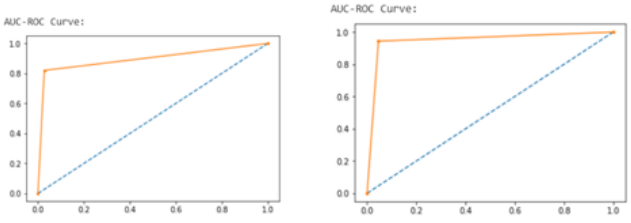

Fig. 4: Light GBM AUC-ROC: Basic vs. Tuned curves

AUC-ROC Curve:: Classification results are evaluated using standard AUC-ROC curve. ROC curve determines the performance of the model and AUC depicts the measure of class separation. For light GBM, it is clearly evident that after tuning the AUC has increased from 94.2% to 95.4% which makes Light GBM the best performing model.

Confusion Matrix:: Figure 5 confusion matrix of Light GBM for basic and tuned models:

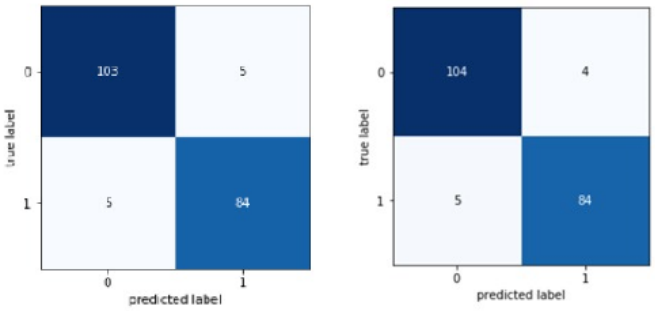

Fig. 5: Light GBM Confusion Matrix: Basic vs. Tuned

V. CONCLUSION AND FUTURE WORK

With this project, we have tried to perform data mining algorithms on the masked credit card transactions data to predict fraud which is ideal in scenarios where the financial industry is unable to share their sensitive data for analysis. Instead, they can simply encrypt the data and apply homomorphic encryption techniques such as PCA transformation to maintain the confidentiality of the data and share for research purposes. We have trained the models using conventional classification algorithms such as Logistic regression and SVC which has resulted in significant results. Further, we have also applied advanced algorithms such as Light GBM, XGBoost and Gaussian Naive Bayes which performed exceptionally well and is ideal for banking domains as these algorithms use the system GPU and provide results quickly with minimal latency. Based on the results evaluation, we can conclude that Light GBM is the best performing algorithm with 95.43% accuracy and 95.3% AUC score which is also more than the state-of-the art performance of 93% [8] achieved so far using Random Forest. Future work of this project would be to perform data mining on a complex financial dataset such as loan applications to predict defaulters.

REFERENCES

[1] H. Zhou, H.-f. Chai, and M.-l. Qiu, "Fraud detection within bankcard enrollment on mobile device based payment using machine learning," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 12, pp. 1537–1545, 2018.

[2] D. Choi and K. Lee, "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System," *IT convergence practice (INPRA)*, vol. 5, no. 4, pp. 12–24, 2017.

[3] C. Soviany, "The benefits of using artificial intelligence in payment fraud detection: A case study.," *Journal of Payments Strategy & Systems*, vol. 12, no. 2, pp. 102–110, 2018.

[4] E. Lopez-Rojas, A. Elmir, and S. Axelsson, "3 Paysim : A financial mobile money simulator for fraud detection," *28th European Modeling and Simulation Symposium, EMSS 2016 :*, no. January, pp. 249–255, 2016.

[5] J. Besenbruch, "Fraud Detection Using Machine Learning," 2018.

[6] R. Rieke, M. Zhdanova, J. Repp, R. Giot, and C. Gaber, "Fraud detection in mobile payments utilizing process behavior analysis," *Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013*, no. September 2015, pp. 662–669, 2013.

[7] R. Wedge, J. M. Kanter, K. Veeramachaneni, S. M. Rubio, and S. I. Perez, "Solving the false positives problem in fraud prediction using automated feature engineering," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11053 LNAI, pp. 372–388, 2019.

[8] https://www.kaggle.com/gpreda/credit-card-fraud-detection-with-rf-auc-0-93

[8] S. Dhankhad, E. A. Mohammed, and B. Far, "Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study," *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, pp. 122–125, 2018.

[9] J. H. Cheon, D. Kim, Y. Kim, and Y. Song, "Ensemble method for privacy-preserving logistic regression based on homomorphic encryption," *IEEE Access*, vol. 6, no. c, pp. 46938–46948, 2018.

[10] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine Learning Classification over Encrypted Data," pp. 1–34, 2015.

[11] V. Biksham and D. Vasumathi, "Homomorphic Encryption Techniques for securing Data in Cloud Computing: A Survey," *International Journal of Computer Applications*, vol. 160, no. 6, pp. 1–5, 2017.

[12] A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei, "Private naive bayes classification of personal biomedical data: Application in cancer data analysis," *Computers in Biology and Medicine*, vol. 105, no. December, pp. 144–150, 2019.

[13] F. J. González-Serrano, Á. Navia-Vázquez, and A. Amor-Martín, "Training Support Vector Machines with privacy-protected data," *Pattern Recognition*, vol. 72, pp. 93–107, 2017.

[14] B. B. Jayasingh, M. R. Patra, and D. B. Mahesh, "Security issues and challenges of big data analytics and visualization," *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, pp. 204–208, 2016.

[15] A. Kim, Y. Song, M. Kim, K. Lee, and J. H. Cheon, "Logistic regression model training based on the approximate homomorphic encryption," *BMC Medical Genomics*, vol. 11, no. Suppl 4, 2018.

[16] W. Li, H. Liu, P. Yang, and W. Xie, "Supporting regularized logistic regression privately and efficiently," *PLoS ONE*, vol. 11, no. 6, pp. 1–20, 2016.

[17] H. Park, P. Kim, H. Kim, K. W. Park, and Y. Lee, "Efficient machine learning over encrypted data with non-interactive communication," *Computer Standards and Interfaces*, vol. 58, no. July 2017, pp. 87–108, 2018.

[18] T. Graepel, K. Lauter, and M. Naehrig, "ML confidential: Machine learning on encrypted data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7839 LNCS, pp. 1–21, 2013.

[19] M. Salem, S. Taheri, and J.-S. Yuan, "Utilizing Transfer Learning and Homomorphic Encryption in a Privacy Preserving and Secure Biometric Recognition System," *Computers*, vol. 8, no. 1, p. 3, 2018.

[20] E. Hesamifard, H. Takabi, and M. Ghasemi, "Deep Neural Networks Classification over Encrypted Data," pp. 97–108, 2019.

APPENDIX

Team Contribution:

Name	Contribution
Digvijay Rai	25%
Naumaan Kazi	25%
Nawaz Sheikh	25%
Ratna Pillai	25%