Synthetic Data Generation and Impact Analysis of Machine Learning Models for Enhanced Credit Card Fraud Detection

Ahmed Abdullah Khaled [1], Md Mahmudul Hasan [2, 3], Shareeful Islam[3,4], Spyridon Papastergiou[4,5], Haralambos Mouratidis[6]

[1] Quantonova, Cambridge, UK
[2] R&D Team, MediprospectsAI Limited
[3]School of Computing and Information Science
Anglia Ruskin University, Cambridge, UK
[4]Research and Innovation, MAGGIOLI S.P.A., Italy
[5]Department of Informatics, University of Piraeus, Greece
[6]Institute for Analytics and Data Science, University of Essex, UK
csewebmail@gmail.com, {md.hasan,shareeful.islam}@aru.ac.uk
spyros.papastergiou@maggioli.gr, h.mouratidis@essex.ac.uk

Abstract. The financial industry is currently experiencing a substantial shift in its operating landscape because of the swift integration of technology. This transformation brings with it potential risks and challenges. Heightened occurrence of online fraud is one the key concerns for this sector, which has been exacerbated by the growing prevalence of online payment methods on e-commerce platforms and other websites. The identification of credit card fraud is a challenging task due to nature of imbalanced transactional data to detect and predict any fraudulent activities. In this context, this paper provides a unique approach to create synthetic dataset to tackle imbalanced issue for credit card fraud detection. The approach adopts Synthetic Minority Over-sampling Technique (SMOTE) technique for balancing dataset. An experiment is performed using several ML models including SVM (Support Vector Machines), KNN (K-Nearest Neighbours), and Random Forest to demonstrate the feasibility of using synthetic data. In this study, we have combined resampling techniques like SMOTE for oversampling the minority class with ensemble methods and appropriate evaluation metrics like the F1-score to improve the imbalanced data. The result from the experiment compared with widely used public datasets to evaluate the model performance. The analysis reveals an imbalance in the real ULB (Université Libre de Bruxelles) dataset, with the positive class (frauds) comprising a mere 0.172% of all transactions. The findings clearly show that the Random Forest model performs better than other modes with outstanding precision, recall, accuracy, and F1 score values to detect fraudulent transactions and reduce false positives.

Keywords: Credit card fraud, Online Transactions, Synthetic Dataset, Imbalanced Dataset, Feature Transformation, Random Forest

1. Introduction

Credit card fraud is identifying which relates with unauthorized purchases, account takeover, and other forms of financial fraud. In recent years, credit card fraudulent activities have become momentous concern for the online transaction. A recent Experian report indicated that fraud rate rose to 18% in the last three months of 2022 in UK [1]. The number of fraud victims in USA has risen to more than 150 million in 2021[2], There are several factors contribute to this frightening concern notably increased number of digital transactions and online shopping, financial implication for a credit card fraud incident from customer and business perspective , and others [3] .As financial institutions strive to ensure the integrity of payment systems, detecting and preventing fraudulent activities has become a top priority. Additionally, fraudsters are becoming more sophisticated, so it is more important than ever to have effective ways to detect and prevent online financial fraud. Traditional fraud detection systems are not always able to keep up with the latest fraud techniques. Machine learning can be used to develop more sophisticated models that can detect fraud more accurately.

Within this context, this synthetic dataset can be used to detect fraudulent transactions in real time using ML models. Hence, the objective is to create models with few false positives and high accuracy. This effort helped to create a more secure and safe financial environment for everyone by tackling these issues. To this end, this paper makes three main contributions.

Firstly, creation of a new synthetic dataset for credit card fraud detection, which is different from widely used public datasets like the well-known ULB (Université Libre de Bruxelles) dataset. The benefit of using synthetic dataset is that it reduces imbalanced nature of the data and ensures data augmentation. Secondly, we used the Synthetic Minority Over-sampling Technique (SMOTE) technique to effectively balance the synthetic dataset manually to address any imbalances. Thirdly, three ML Models are used to demonstrate the performance of different ML algorithms and the getting result was compared with widely recognised public datasets to assess the performance of the model. The findings reveal that the real ULB dataset exhibits an imbalance, with the positive class (frauds) representing a mere 0.172% of all transactions.

In this paper, section 2 describes the related work. Proposed synthetic data generation procedures for credit card fraud detection has been explained in section 3. Experimental setup and analysis of this study have described in the following section 4. Finally, concluding remarks and future direction have been mentioned in section 5.

2. Related Work

There are several works that focus on different techniques for synthetic data generation and sampling credit card fraud detection which are relevant to our work. A hybrid approach is used to develop and evaluate a predictive model for detecting credit card fraud which combines machine learning and deep learning techniques [5]. Specifically, the model incorporates BiLSTM, MaxPooling, and BiGRU operations and is compared against six traditional machine learning classifiers for efficiency measurement. The paper also addresses the challenges posed by imbalanced and complex datasets in the field of fraud detection, to provide a more secure and reliable system for both consumers and companies. The main objective of the Cluster-based Over-sampling with noise filtering (KMFOS) approach as mentioned in [6] is to address the class imbalance problem in Software Defect Prediction (SDP) models. Traditional over-sampling methods like SMOTE often generate non-diverse synthetic instances and introduce unnecessary noise, making it difficult for classifiers to accurately identify defective instances. KMFOS aims to improve upon this by first dividing the existing defective instances into K clusters and then generating new, more diverse defective instances through interpolation between instances from different clusters. An analytical method, based on Bayes Minimum Risk theory, to correct the bias induced by under-sampling [7]. This correction allows for well-calibrated probability estimates and guides adjusting the classification threshold to account for the change in class priors. A novel under-sampling method called Particle Stacking sampling is used for addressing class imbalance which offers computational efficiency and minimizes information loss, thereby mitigating the bias towards the majority class commonly seen in classifiers [8]. Another work that considers both random under-sampling and SMOTE, to balance the dataset before training [9, 10]. Most class samples are randomly discarded in random under-sampling, whereas SMOTE method is used to generate synthetic examples for the minority class. This involves identifying the five nearest neighbours for each sample in the fraud class and creating new synthetic samples along lines connecting the sample and its neighbours. Unlike the random over-sampling method, SMOTE generates new samples to mitigate overfitting. By utilising these preprocessing methods, authors successfully achieve a balanced dataset for training, overcoming the potential negative effects of dataset imbalance. Credit card fraud detection is investigated by comparing the performance of naïve Bayes, k-nearest neighbour, and logistic regression classifiers on the (Université Libre de Bruxelles) ULB dataset comprising 284,807 European transactions [11]. A hybrid sampling technique balances the skewed data by under-sampling legitimate transactions and oversampling fraudulent ones. Various metrics assess performance, revealing that the k-nearest neighbour algorithm is most effective. Various ML algorithms is used to detect fraudulent transactions in a real-world dataset generated from European cardholdersULB dataset by [12]. The work considers SMOTE to alleviate the issue of class imbalance and results show that SMOTE technique can solve the issue of class imbalance in credit card fraud datasets. Moreover, they have demonstrated that pairing the AdaBoost method with several ML models can increase the performance of the proposed framework. Several ML algorithms, i.e., CatBoost, LightGBM, and XGBoost and class weight-tuning hyperparameters are used to improve credit card fraud detection by [4]. Additionally, deep learning is used to fine-tune the hyperparameters, including the proposed weight-tuning technique.

Based on our comprehensive analysis of recent literature, it is evident that there is a notable lack of studies addressing the current situation, particularly in the domain of imbalanced datasets and online financial fraud detection. Therefore, we have addressed this gap. The core contribution of our work is twofold. Firstly, it demonstrates the efficacy of machine learning models when trained on synthetic datasets, a method that presents a novel approach in the context of financial fraud detection. Secondly, we provide a comparative analysis of our results against previous research that employed public datasets, such as the IEEE-CIS Fraud Detection and Université Libre de Bruxelles (ULB) datasets. This comparison not only validates the reliability of synthetic datasets in fraud detection but also offers insights into the evolution and improvement of machine learning techniques in this domain. By bridging the gap between simulated and real-world data, our research contributes to the field of financial security, offering potential pathways for financial institutions to enhance their fraud detection mechanisms and safeguard customer transactions in an increasingly digital world.

3. Proposed Synthetic data development for fraud detection

The proposed approach considers Synthetic Minority Over-sampling Technique (SMOTE) technique for synthetic data generation. A real-world-like synthetic dataset is generated, balanced with SMOTE, fed into diverse ML models, and evaluated for accurate fraud identification. The aspect of this work is the creation and utilization of a synthetic dataset, balanced using SMOTE, for training fraud detection models, which is then compared with models trained on public datasets to assess its effectiveness.

A further step involved building machine learning models with three different algorithms: Random Forest (RF), k-nearest Neighbours (KNN), and Support Vector Machines (SVM). Every model was trained and evaluated twice: once on the unbalanced dataset and once on the balanced dataset produced by SMOTE. Regarding credit card fraud detection, the purpose of this dual strategy was to find out how balance dataset affected each algorithm's performance. A detailed comparison study has been conducted to assess the performance of our models. We compared our results with those from earlier research using publicly available datasets. We evaluated our models' performance concerning industry standards, including the Université Libre de Bruxelles (ULB) dataset, which served as a standard for evaluating the effectiveness of fraud detection algorithms. Standard measures including precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were used to evaluate each model's performance. These criteria were selected to provide a thorough assessment of the models' capacity to detect fraudulent transactions, considering both false positives and false negatives. The proposed framework has seven major components as illustrated in Figure 1.

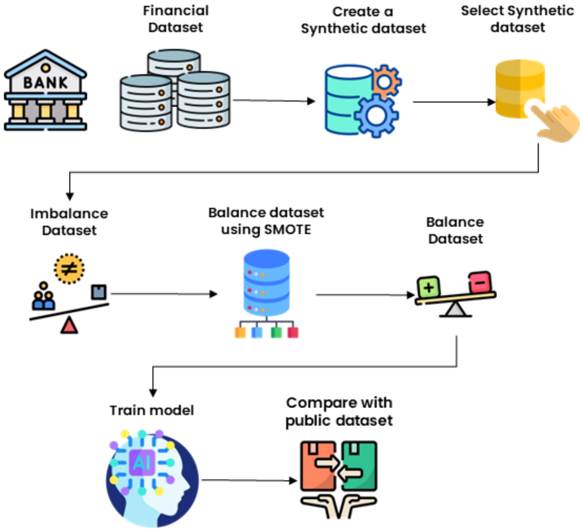

Fig. 1. High Level Architecture

Financial Dataset Acquisition: The process begins with data acquisition from financial institutions, symbolised by a bank building and data storage units. This raw financial dataset likely contains transaction records with various features.

Synthetic Dataset Creation: A synthetic dataset is created from the financial dataset. This involves using algorithms to generate data that simulate real transaction patterns. The synthetic dataset is designed to reflect the characteristics of genuine transactions while protecting privacy or addressing data scarcity.

Selection of Synthetic Dataset: A subset of the synthetic data is then selected for use. This step involves choosing a specific portion of the generated data that best serves the purpose of the study or has the desired properties, like a particular balance of classes or a specific range of transaction values. Generating transactions based on the customer and terminal profiles created earlier is a crucial step in this data generation process. To achieve this, a function is defined to find the list of terminals for each customer within a given radius. The spatial proximity of the customer's location to the available terminals governs this process. Transactions are then generated for each customer over a set number of days. The average number of transactions per day derived from the corresponding customer profile is used to create a Poisson distribution from which the number of transactions for each day is drawn. The transaction timestamp (TX_DATETIME), the customer's ID (CUSTOMER_ID), the terminal's ID (TERMINAL_ID), the transaction amount (TX_AMOUNT), the transaction's duration in seconds (TX_TIME_SECONDS), and the duration in days (TX_TIME_DAYS) are all included in every generated transaction.

Balancing the Dataset Using SMOTE: Initially, the synthetic dataset is imbalanced, which is a typical scenario in fraud detection where fraudulent transactions are far less frequent than legitimate ones. To correct this imbalance, SMOTE is applied, which synthesizes new examples from the minority class, in this case, the fraudulent transactions, to achieve a more balanced class distribution.

Final Balanced Dataset: The result of the SMOTE application is a balanced dataset, where the class distribution is now even, allowing for more effective training of machine learning models. This balanced dataset is ready for use in model training.

Machine Learning Model Training: The balanced dataset is used to train a machine-learning model. This is depicted as an AI brain, indicating the development of a predictive model capable of distinguishing between different types of transactions based on the balanced data provided.

Comparison with Public Dataset: After the model is trained, its performance is compared against models trained on public datasets. This step is crucial to validate the effectiveness of the synthetic data approach and to benchmark the model's performance against established datasets known in the field, such as the IEEE-CIS Fraud Detection and ULB datasets.

4. Experimental setup

The purpose of the experiment is to evaluate the effectiveness of syntactic data for fraud detection using several ML models. To accomplish this, firstly, we have prioritised the authentic transaction characteristics of the dataset, including transaction frequency and patterns, as well as the accurate representation of fraudulent activities. These elements are crucial in the development of realistic synthetic datasets for financial applications. Secondly, the development of a systematic method to generate a synthetic dataset, which methodically emulates real-world financial transactions, employing identified key factors to craft scenarios that closely resemble actual transaction environments. Finally, the production of a financial transaction-oriented synthetic dataset for validation purposes, utilising the outlined methodology with variations tailored to address the imbalance between legitimate and fraudulent transaction instances and enhancing the representativeness of fraudulent behavior within the data.

Furthermore, the credit card fraud detection process has been enhanced by implementing three machine learning algorithms: K-Nearest Neighbours (KNN), Random Forest (RF), and Support Vector Machines (SVM). To ensure comprehensive results, the synthesised dataset has been divided into subsets for testing and training, considering both balanced and unbalanced datasets. Each model has undergone thorough training and refinement using the training set, with the utilisation of cross-validation techniques to optimize hyperparameters and improve generalisation. The performance of each model has been assessed on both balanced and unbalanced datasets, using various evaluation measures such as accuracy, precision, recall, F1-score, and ROC AUC [16]. Finally, the models' ability to minimise false positives and accurately identify fraudulent transactions has been evaluated. These models are compared with the results of earlier studies to assess their effectiveness and how they measure up against established benchmarks in detecting credit card fraud by using the following parameters.

- **Fraudulent Ratio:** The number of valid (non-fraudulent) transactions is determined by subtracting the count of fraudulent transactions from the total number of transactions in the dataset.
- **Total Transaction Count**: The total number of transactions counted includes both valid and fraudulent transactions.

The fraudulent ratio is calculated using the following formula:

$$Fradulent\ ratio = \frac{number\ of\ fraud\ transaction}{number\ of\ total\ transaction}\ x\ 100\%$$

The extracted features from the raw dataset were insufficient in providing adequate information, as illustrated in Figure 2. As a result, the machine learning models struggle to identify patterns within the data. Consequently, this leads to a decrease in accuracy and prolongs the training process, as the models encounter difficulty in comprehending the underlying data patterns.

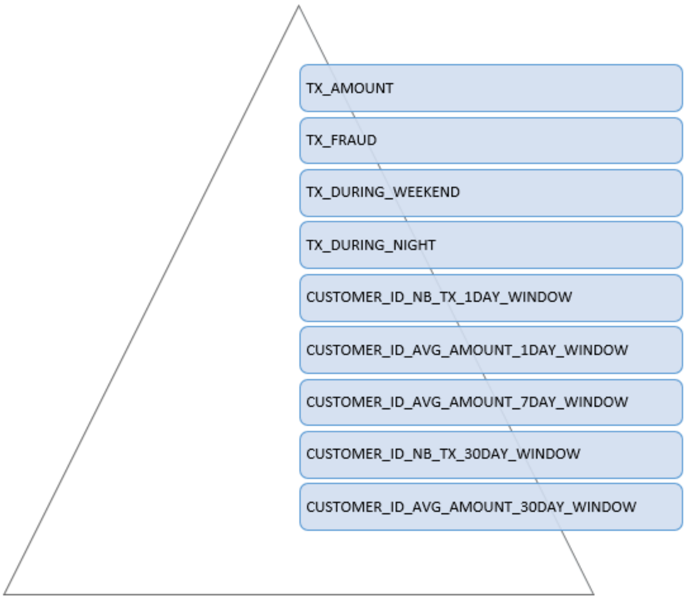
TX_AMOUNT
TX_FRAUD
TX_DURING_WEEKEND
TX_DURING_NIGHT
CUSTOMER_ID_NB_TX_1DAY_WINDOW
CUSTOMER_ID_AVG_AMOUNT_1DAY_WINDOW
CUSTOMER_ID_AVG_AMOUNT_7DAY_WINDOW
CUSTOMER_ID_NB_TX_30DAY_WINDOW
CUSTOMER_ID_AVG_AMOUNT_30DAY_WINDOW

Fig. 2. Features after Transformation.

4.2 Results and Discussions

The following features as depicted in Table-1 represents the average expenditure of a customer over a period of 30 days by the synthetic dataset, providing a comprehensive understanding of the customer's overall spending patterns.

Table 1. Sample Customer Profile for Different Transactions

CUSTOMER_ID	x_customer_id	y_customer_id	mean_amount	std_amount	mean_nb_tx_per_day
0	49.460165	22.808310	29.270023	14.635011	1.585320
1	37.731510	99.657423	43.778734	21.889367	3.087576
2	76.053669	31.000935	37.921414	18.960707	1.407059
3	14.546686	97.266468	91.371952	45.685976	2.239828
4	31.359075	88.820004	69.084441	34.542221	1.564350
5	50.718412	52.410350	93.160088	46.580044	2.285492
6	66.833757	5.225869	36.070444	18.035222	0.225607

Moreover, developing a strong customer-terminal relationship is a critical aspect of our study to develop a real-world artificial data. This is achieved by utilising a radius-based approach. Both customers and terminals are assigned x and y coordinates, allowing us to calculate the distance between them within a Cartesian coordinate system. If this distance falls below a predetermined radius, we consider the terminal accessible to the customer as illustrated in Figure 3.

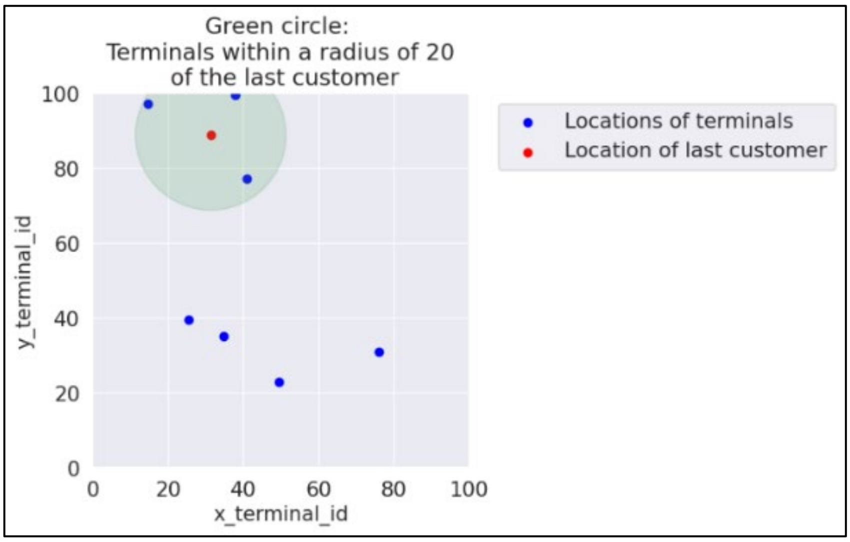

Fig. 3. Customer Interaction Terminal

Most transaction amounts are concentrated in smaller values within the distribution. As for transaction times, they exhibit a Gaussian distribution centred around midday when observed daily. These distributions align with the simulation parameters utilized in the earlier sections. After that, the fraud categories are added to the dataset. At first, all the TX_FRAUD column values were made zero which means not fraud. Then five attack scenarios were designed, and the fraud categories were entered as value one. The transaction table is created without a fraud scenario as shown in this Table 2.

Table 2. Synthetic Transaction Generation for Single Customer

TX_DATETIME	CUSTOMER_ID	TERMINAL_ID	TX_AMOUNT	TX_TIME_SECONDS	TX_TIME_DAYS
2018-04-01 22:22:31	0	5	14.97	80551	0
2018-04-01 17:16:41	0	5	27.05	62201	0
2018-04-02 14:41:26	0	4	37.75	139286	1
2018-04-03 16:48:08	0	5	18.41	233288	2
2018-04-02 09:55:10	0	5	33.30	122110	1
2018-04-05 03:55:12	0	4	16.28	359712	4
2018-04-05 00:59:44	0	1	24.18	349184	4

The distribution of valid and fraudulent transactions is visualized in Figure 4 and Figure 5, both before and after the application of SMOTE methods. The dataset contains a total of 142,486 transactions, out of which 13,778 were identified as fraudulent. The imbalance shown here is typical in fraud detection scenarios, where fraudulent transactions are much rarer than legitimate ones, posing a challenge for machine learning models to learn the patterns of fraud effectively.

It is important to note that the initial dataset was imbalanced, thus requiring the use of SMOTE to create a balanced dataset for improved results. Once the dataset was balanced, our algorithm was deployed for training and testing purposes.

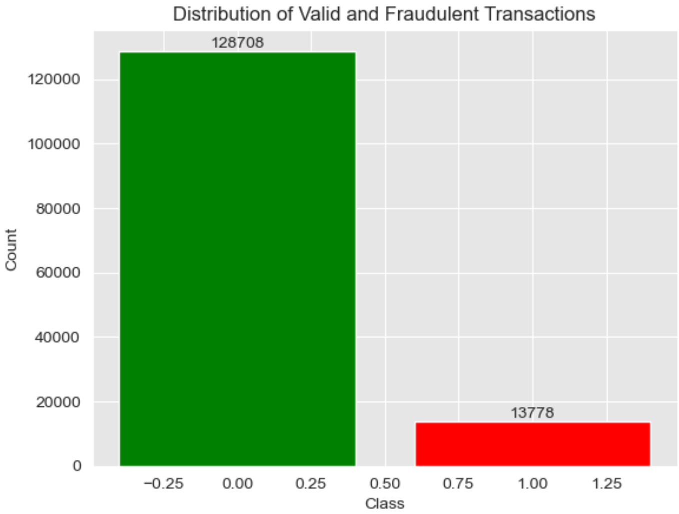

Fig. 4. Fraudulent and Valid Transaction Ratio for Synthetic Dataset (Imbalanced)

To detect instances of credit card fraud, we adopted a rigorous approach. By utilising a well-calibrated synthetic dataset, we meticulously divided it into subsets for both testing and training purposes. It is imperative that each model undergoes comprehensive training and refinement using the designated training set.

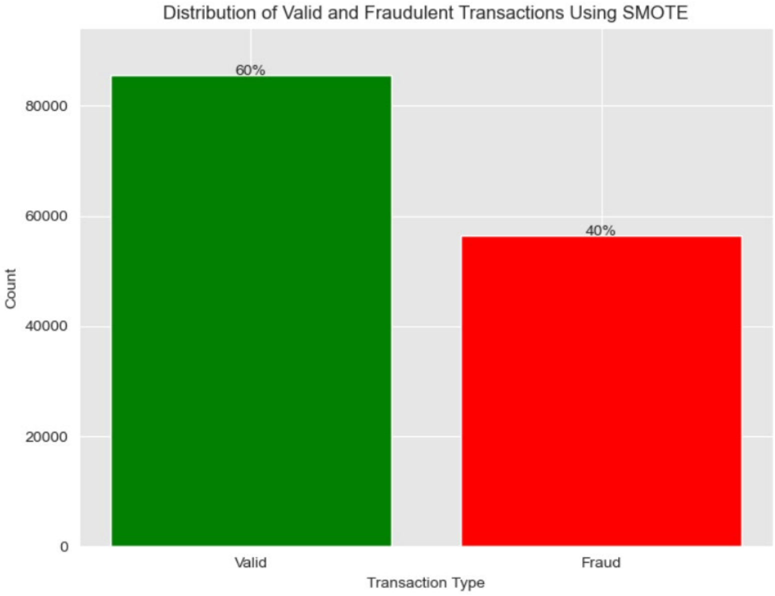

Fig. 5. Fraudulent Ratio for Synthetic Dataset (Balanced)

Table 3 presents an overview of the performance metrics achieved by a Support Vector Machine (SVM) model trained on an imbalanced dataset. Despite achieving a relatively high accuracy of 0.9138, the model exhibits a notable challenge in recall, with a value of 0.1065. The perfect precision score of 1.0 indicates the SVM's confidence in predicting the positive class. However, the F1 Score of 0.19249 reflects a trade-off between precision and recall, suggesting that the model's predictive capability is hindered by its low recall.

The performance metrics of machine learning models trained on both balanced and imbalanced datasets. The models that have been examined include Support Vector Machine (SVM), k-Nearest Neighbours (KNN), and Random Forest. The assessed metrics encompass F1 Score, Accuracy, Precision, and Recall, which collectively provide insights into various aspects of the model's performance. Upon closer examination of the findings in the table, it becomes evident that the Random Forest model consistently outperforms other models across multiple measures and datasets. Specifically, when trained on the unbalanced dataset, the Random Forest model achieved an Accuracy of 0.9434, Precision of 0.9851, Recall of 0.4195, and an F1 Score of 0.5885. Similarly, on the balanced dataset, it exhibited an Accuracy of 0.9073, Precision of 0.9727, Recall of 0.7430, and an F1 Score of 0.8425.

Table 3. Performance Metrics for Imbalanced Dataset

	Imbalanced Dataset
Accuracy	SVM: 0.9138, KNN: 0.9298, RF: 0.9434
Precision	SVM: 1.0000, KNN: 0.8793, RF: 0.9851
Recall	SVM: 0.1065, KNN: 0.3164, RF: 0.4195
F1 score	SVM: 0.19249, KNN: 0.4653, RF: 0.5885

The Random Forest model, developed using a balanced dataset, exhibits consistent and reliable performance across multiple metrics as shown in Table 4. It demonstrates commendable accuracy, precision, recall, and F1 score values, indicating its effectiveness in handling both positive and negative instances.

Random Forest outperforms the other models, demonstrating superior accuracy (98.10%), precision (99.32%), and F1 score (90.21%), with a notably high recall (98.65%) as well. These results suggest that the Random Forest model, with its ensemble approach and ability to capture complex data relationships, is highly effective for fraud detection tasks, offering robust generalization capabilities and excelling in the accurate identification of fraudulent transactions.

Table 4. Performance Metrics for Balanced Dataset

	Imbalanced Dataset
Accuracy	SVM: 0.9038, KNN: 0.899, RF: 0.9810
Precision	SVM: 1.0000, KNN: 0.157, RF: 0.9932
Recall	SVM: 0.1265, KNN: 0.011, RF: 0.9865
F1 score	SVM: 0.17249, KNN: 0.021, RF: 0.9021

The Random Forest model has demonstrated exceptional performance due to its ensemble approach and ability to capture complex relationships within the data effectively. This ensemble technique effectively mitigates overfitting, resulting in robust generalization on both imbalanced and balanced datasets. The model's accurate identification of positive instances is evident through its impressive Precision values on both datasets. Additionally, the strong Recall demonstrated on the balanced dataset indicates its capability to capture a portion of positive class instances accurately. In contrast, the SVM and KNN models show varying outcomes. Despite the SVM model's perfect Precision on the unbalanced dataset, its low Recall has limited its usefulness. Similarly, the KNN model consistently achieves high Precision values but has relatively low Recall, mainly when applied to unbalanced datasets.

In general, the Random Forest model demonstrates its superiority among the evaluated models by effectively balancing Accuracy, Precision, Recall, and F1 Score on both imbalanced and balanced datasets. Its utilization of ensemble-based methodology enables it to overcome the challenges associated with class imbalances, further enhancing its predictive capabilities. It excelled in maintaining high Precision while successfully identifying a number of true positive cases. This characteristic augment the model's reliability and suitability for practical, real-world applications.

5. Conclusion

Considering the increase in online transactions, it is crucial to address the issue of credit card fraud. The objective of this study was to generate reliable synthetic datasets and develop accurate machine learning models capable of detecting fraudulent transactions. The goal is to protect individuals, businesses, and financial entities from potential financial losses and security breaches. The use of synthetic datasets offers several benefits as it allows for the creation of a comprehensive dataset that closely resembles real credit card transactions. This dataset encompasses a wide range of scenarios and variables, which serves as a solid foundation for subsequent model development and testing. The creation of the dataset involved various meticulous processes, including the building of client and terminal profiles, transaction simulation, introduction of fraud scenarios, and computation of fraud ratios. Thanks to the abundance of data, our machine learning models were well-prepared for training and testing.

The findings of this project is encapsulated in its novel approach to generating a synthetic dataset designed explicitly for the challenge of financial fraud detection. By leveraging a simulator to produce data that closely mimics real transactional behaviour and addressing class imbalance through the application of SMOTE, this research stands out in its ability to create a realistic and balanced training ground for machine learning models. This is further enhanced by integrating robust classification algorithms and strategic comparisons with traditional public datasets, which collectively demonstrate the synthetic data's efficacy and contribute to advancing fraud detection methodologies. The project's distinctive contribution is its methodical blend of synthetic data generation, sophisticated balancing techniques, and pragmatic model evaluation. It establishes a new benchmark for machine learning applications in the financial sector. In the future, it is recommended to explore the integration of advanced deep learning algorithms, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to analyse the impact and conduct a comparative analysis. Leveraging deep learning models and real-time monitoring can enable the automatic acquisition of intricate patterns and relationships within the data, thereby potentially enhancing the accuracy of fraud detection.

Acknowledgement

This work was supported by the European Union's Horizon Europe Project, SEPTON- Security Protection Tools For Networked Medical Devices, funded under grant agreement No. 101094901 and CyberSecDome — An innovative Virtual Reality based intrusion detection, incident investigation and response approach for enhancing the resilience, security, privacy and accountability of complex and heterogeneous digital systems and infrastructures, funded under grant agreement No. 101120779.

References

1. What's credit card fraud detection and why is it important? https://www.capitalone.com/learn-grow/privacy-security/credit-card-fraud-detection/.
2. 2023 Credit Card Fraud Report /, https://www.security.org/digital-safety/credit-card-fraud-report/
3. S. Varga, J. Brynildsen, and U. Franke, "Cyber-threat perception and risk management in the Swedish financial sector," Compute Security, vol. 105, p. 102239, Jun. 2021, Doi: 10.1016/j.cose.2021.102239.
4. S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud Detection in Banking Data by Machine Learning Techniques," IEEE Access, vol. 11, pp. 3034–3043, 2023, Doi: 10.1109/ACCESS.2022.3232287
5. H. Najadat, O. Altiti, A. A. Aqouleh, and M. Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning," in 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, Apr. 2020, pp. 204–208. doi: 10.1109/ICICS49469.2020.239524.

6. L. Gong, S. Jiang, and L. Jiang, “Tackling Class Imbalance Problem in Software Defect Prediction Through Cluster-Based Over-Sampling with Filtering,” IEEE Access, vol. 7, pp. 145725–145737, 2019, doi: 10.1109/ACCESS.2019.2945858.
7. A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating Probability with Undersampling for Unbalanced Classification,” in 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa: IEEE, Dec. 2015, pp. 159–166. doi: 10.1109/SSCI.2015.33.
8. Y.-S. Jeon and D.-J. Lim, “PSU: Particle Stacking Undersampling Method for Highly Imbalanced Big Data,” IEEE Access, vol. 8, pp. 131920–131927, 2020, doi: 10.1109/ACCESS.2020.3009753.
9. Y.-R. Chen, J.-S. Leu, S.-A. Huang, J.-T. Wang, and J.-I. Takada, “Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets,” IEEE Access, vol. 9, pp. 73103–73109, 2021, doi: 10.1109/ACCESS.2021.3079701.
10. Z. Zhang and S. Huang, “Credit Card Fraud Detection via Deep Learning Method Using Data Balance Tools,” in 2020 International Conference on Computer Science and Management Technology (ICCSMT), IEEE, Nov. 2020, pp. 133–137. doi: 10.1109/ICCSMT51754.2020.00033.
11. J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” in 2017 International Conference on Computing Networking and Informatics (ICCNI), IEEE, Oct. 2017, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.
12. E. Ileberi, Y. Sun, and Z. Wang, “Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost,” IEEE Access, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.
13. S. K. Hashemi, S. L. Mirtaheri, and S. Greco, “Fraud Detection in Banking Data by Machine Learning Techniques,” IEEE Access, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.
14. S. K. Hashemi, S. L. Mirtaheri, and S. Greco, “Fraud Detection in Banking Data by Machine Learning Techniques,” IEEE Access, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.
15. F. Itoo, Meenakshi, and S. Singh, “Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection,” International Journal of Information Technology (Singapore), vol. 13, no. 4, pp. 1503–1511, Aug. 2021, doi: 10.1007/S41870-020-00430-Y/METRICS.
16. R. Kumar, A. Aljuhani, D. Javeed, P. Kumar, S. Islam, A.K.M. N. Islam. Digital Twins-enabled Zero Touch Network: A smart contract and explainable AI integrated cybersecurity framework, Future Generation Computer Systems,Volume 156.,2024