Synthetic data generation in finance: requirements, challenges and applicability

Emilija Strelcenia*
Department of Creative Technology
Bournemouth University
Bournemouth, United Kingdom
strelceniae@bournemouth.ac.uk

Simant Prakoonwit
Department of Creative Technology
Bournemouth University
Bournemouth, United Kingdom
sprakoonwit@bournemouth.ac.uk

***Abstract*—Financial datasets possess susceptible, private and identifiable details about clients. The usage and distribution of such data for research outside a financial institution are strictly constrained due to privacy laws. One option to deal with this restriction is creating artificial data. The generation of fake data protects the confidentiality of customers' information. Data privacy is a prime concern in public opinion. This research study reviews various requirements and challenges for data generative techniques and handling synthetic data in finance.**

Keywords— class imbalance, financial data, synthetic data,

I. Introduction

Data scarcity is a significant issue since a large bulk of data is required in fraud detection for training machine learning frameworks. Deep learning, a machine learning subset, utilises the synthetic neural network to learn models from data. On the other hand, neural networks are a combination of multiple layers of neurons attached by weighted connections. Data augmentation is an efficient way to deal with this challenge [1]. In recent years, researchers have conducted various studies in this particular area.

Data collection is a time-consuming and way costly task to acquire data. One of the main issues being faced while training datasets is that there needs to be more training data available in many application domains [2]. In some areas, data collection is not possible. For instance, due to privacy concerns, original data training in credit card fraud detection is impossible. On the contrary, financial institutions need extensive data to detect monetary fraud cases. Data augmentation is an effective method used to generate data from existing data synthetically. Data augmentation saves both time and costs in gathering required data. Furthermore, it decreases sample inadequacy problems in machine learning models [3].

Large bulks of data streams are required to enhance the accuracy of deep learning frameworks. Besides limited training data, the need for more relevant data is a fundamental challenge to regress datasets. Data augmentation can provide solutions through different methods to improve the quality and size of training datasets to get a better outcome [4].

Furthermore, model over-fitting is also regarded as a big challenge. Over-fitting is a modelling error which occurs as the model closely fits the existing data set. Deep learning models require sufficient data to avoid the issue of over-fitting. In addition, while training a model on inadequate datasets, it becomes hard for the model to generalise perfectly for a new dataset. In addition, when tested for any new data, these models will only provide accurate predictions, making the model efficient. Therefore, the model needs more datasets to deal with the over-fitting challenge. However, data augmentation lessens the issue of over-fitting through training with a bulky set of appropriate data [5]. Furthermore, data augmentation regularises the model and enhances its ability to generalisation.

Besides the above challenges, imbalanced data is also a significant problem in dealing with real-life applications. This problem is prevalent in financial institutions, especially with credit card fraud detection, as fraud transactions are too few compared with legal transactions. In addition, deep learning models need data in considerable quantities to classify correctly. However, occasionally available data needs to be more balanced, which creates difficulty in training deep models and affects the overall accuracy [6]. According to [7], scarcity of data is a significant problem when building deep learning methods. Data availability is only possible in some fields to train a model. Many researchers have been introducing new ways to solve this issue in recent years. Data can be re-sampled to solve this challenge; however, data augmentation can help this issue by dealing with highly imbalanced datasets by creating data for training machine learning models. Various studies have been proposed to augment and run these techniques on datasets. The studies have also presented the association between deep learning methods and datasets.

Augmentation of data can be done through several approaches, for instance, the adversarial approach developed by [8]. This technique is a category of AI-based algorithms modelled to sort out the generative modelling challenge. In addition, this algorithm aims to explore a collection of training samples and learn the patterns of distributions which generate them. Other than that, the heuristic approach by [9] and the style transfer approach proposed by [10] are significant examples of approaches to augment data.

II. Requirements for Data Generative Techniques

Generally, researchers cannot use financial data for research purposes due to privacy law restrictions. One effective way to deal with this challenge is synthetic data generation.

Synthetic data can be defined as the data obtained through the generative method that represents real-world data traits [11]. Artificial data is produced synthetically. Synthetic data is used to enhance the efficiency of deep learning methods. This type of data can be used when data is limited. GANs can generate artificial instances that have data samples of real-world datasets. Thus, artificial financial data generation safeguards sensitive details about customers.

The primary aim of generative techniques is to synthesise new instances- highly linked to real data but cannot be mapped back to real-world data [12]. The data-generating process is dissimilar from most obfuscation methods, such as altering sensitive traits from datasets. This study highlighted three main requirements for generative frameworks to create synthetic financial data: a) The ability to generate multiple types of financial data, namely categorical, binary, complex as well as numeric; b) The generative process should have the ability to produce arbitrary numbers of data points and; c) the confidentiality of financial datasets has to be accurately tuned against how valuable, and near to real the data is.

Data streams vary based on the specific areas of interest, from time series data representations and tabular representations. It is noteworthy that credit card holders' primary consideration is privacy over security. On the other hand, the generative techniques only learn the characteristics of real datasets without affecting customers' privacy. Hence, if fraudsters gain full access to the generating algorithms and artificial data streams cannot abuse the original datasets.

III. Why do we need Synthetic Data?

Some benefits and applicability of generating artificial data in finance are mentioned below.

A. *Insufficient historical data*

Fraud detection in the credit card domain faces significant challenges due to the non-availability of real-world datasets for testing. One of the main reasons is regulatory barriers which restrict researchers from using and publishing the findings of their studies on these datasets. In addition, financial institutions need historical data to learn about major past events, such as recessions, shifting market trends, and market crashes. However, the institutions usually need historical data on certain circumstances. Therefore, synthetic data streams are helpful in such a setting as they act as counterfactual data to test inferences and strategies [13].

B. *Regulatory Limitations*

Deep learning methods based on Neural Networks (NN) are attaining considerable outcomes in multiple domains. Most of the time, the training phase of these methods needs representative and large datasets which contain sensitive details about customers. Privacy laws and regulations may avert companies from sharing private data with their customers, even within their departments. However, on the other hand, data is required for financial institutions for research and development. Hence, synthetic data is needed to fulfil the needs of financial institutions [14].

C. *Handling the imbalanced class issue*

One of the biggest challenges being faced in credit card fraud detection is the imbalance class problem. Conventional deep learning algorithms and other anomaly detection approach frequently fail due to this issue. Moreover, artificial data and data imputation approaches can deal with the class imbalance problem [15].

D. *Training through deep machine learning techniques*

Deep machine learning is usually carried through cloud services. The learning approach requires computing resources along with vast numbers of training data. Moreover, financial institutions need support to acquire accurate work data for training. However, artificial data can be incorporated for training models, which can then be reverted to implement real-world data. In addition, training on synthetic data protects against inference attacks, in which model parameters are utilised to extract training data [16,17].

E. *Data Sharing*

Data distribution offers better solutions as well as helps reduce technical issues within and between financial institutions. Furthermore, artificial data, with natural characteristics, let commercial banks share data without privacy laws and regulatory matters [18].

IV. Synthetic Data Generation Techniques

The subsequent section explores two approaches to generating synthetic financial data.

A. *Tabular data generation*

Scholars have introduced many approaches for generating tabular data. Similarly, many GAN-based generative models have also been introduced to create tabular data. For instance, variational auto-encoders (VAEs) and conditional GANs have been introduced, such as Table-GAN [19], CT-GAN [20], T-VAE [20] and PATE-GAN [21]. On the other hand, CT-GAN sorts out the restraints of former techniques by incorporating a cGAN framework and encoding for continuous and discontinuous variables to enhance the range of generated samples. Unfortunately, though, CT-GAN only covers some privacy concerns. In addition, PATE GAN alters the conventional GAN training process for generating tabular data that is differentially private.

B. *Artificial time series financial data*

Besides tabular data, much research has been done, and scholars have proposed several approaches to create artificial time series financial data. However, it is imperative to mention that more needs to be done to address privacy concerns. The modern techniques for generating artificial time series financial data are Quant-GAN [22] and the approach introduced by [23] in which CGAN in time series forecasting and modelling was introduced. These models are used for log returns of financial instruments and related time series models. However, it is pertinent to mention that these models provide no privacy guarantees.

V. Machine Learning Techniques for Synthetic Data Generation

The invention of this century, machine learning, replaces traditional methods and can operate on large datasets that are inaccessible to people immediately. Unsupervised learning versus supervised learning is two key divisions in machine learning strategies. Fraud detection can be done in any way, and the datasets should choose only how to use it. Anomalies are always to be recognised in supervised instruction. Over recent years, various frameworks have been used for detecting credit card-related fraud. The highly unbalanced databases appear to be the main hurdles to implementing ML for fraud detection. With such a small number of fraudulent payments, the majority are shown to be authentic in many sets of evidence. One of the investigators' greatest challenges is designing a framework for fraud prevention that is accurate, effective, and will produce few false positives while effectively identifying fraud activity [24,25].

Machine learning techniques play an essential role in many areas of processing data. One of these areas is the detection of credit card-based fraud. There are several ML techniques to detect this fraud using supervised methods, unsupervised methods and hybrid techniques [25].

In recent years, improved ML frameworks can apply various complex statistical calculations to big data settings, which computes the finding quickly. For example, machine learning techniques improve the classifier performance and detection rate. Further, it analyses the effectiveness of several classification techniques on even a highly skewed credit card fraud database, such as random forest, artificial neural networks, tree classifiers, Naive Bayes, supporting vector machines, gradient boosting classifiers, and logistic regression approaches [26].

A. Methodology

Begin gathering information about the credit card data set. Then, use machine learning (classification system) techniques. After that, utilise one-class classifiers and the Matthews correlation coefficient to further analyse the data before confirming multiple imbalances. The next step would be plotting the correlation matrix for the whole dataset. In machine learning models, the learning process involves two stages, the training stage and the testing stage. In the first stage, training data samples are denoted as inputs where the learning framework learns features. In the testing stage, the learned algorithm is applied in the execution engine to make test predictions. After that, divide data into subsets to train and test them. For instance, use 75% for training and 25% for testing. Finally, use formulae to calculate the total evaluation of different metrics, such as the confusion matrix, FPR, recall, accuracy, F1-Score and precision rate.

Further, the study [27] utilised the feedback mechanism to raise the detection rate and precision. The general overview of CC fraud datasets is shown in figure 1.

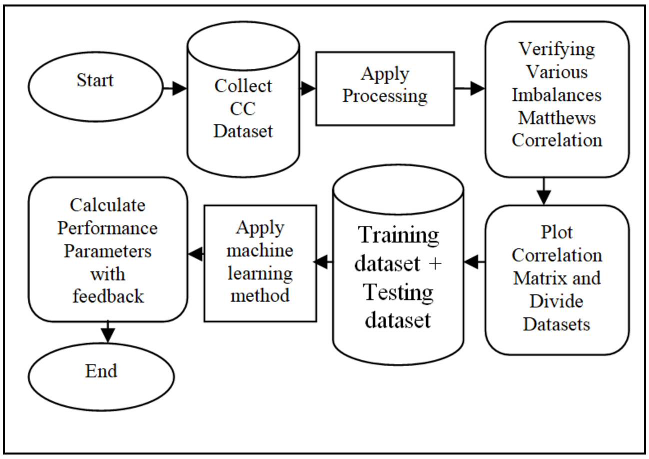
Start
Collect CC Dataset
Apply Processing
Verifying Various Imbalances Matthews Correlation
Plot Correlation Matrix and Divide Datasets
Training dataset + Testing dataset
Apply machine learning method
Calculate Performance Parameters with feedback
End

Fig. 1. Steps in CC Fraud Detection Using ML

Repeat steps 4 through 6 for different classifiers. It has been computed to measure performance using various parameters. In this data analysis, various classifier techniques of machine learning are implemented using Python programming language to confirm credit card fraud. The data set has been split into two groups during implementation: 30% for testing and 70% for training. For the CC dataset, many parameters have been estimated based on experiments.

B. Data Selection

Both the Kaggle website and another website provide a set of data from the ULB Machine Learning Community. The dataset comprised all 2013 credit card transactions made by Europeans. The total dataset has 284,807 transactions. Out of the total transaction, only 492 were fraudulent transactions. PCA and time, class, and quantity were additional non-PCA-based primary features that may apply primarily to V1 through V28 as the entire data sets appear to be highly unbalanced. Two subcategories, 0 and 1, have already been fully subdivided into the class. Whereas class 1 was exclusively a fraud, class 0 may be data gathered that was not a scam.

C. Findings

The findings of the study performed by [27] showed that RF algorithms have an accuracy of 95.988%. In contrast, SVM, Decision Trees, LR, GBM, NB have precision percentages of 93.228, 90.9%, 93.99%, and 91.2%, respectively, for ULB machine learning credit card fraud identification. Greater values were demonstrated to be accepted for any machine learning strategy as simply a superior performance method of precision, accuracy, recall, and F1-score.

Furthermore, [27] study argued that a manifestation of criminal dishonesty, credit card fraud has no doubt. Until machine learning techniques were introduced, fraud detection appeared to be a complex problem requiring much skill. However, it was an implementation of artificial intelligence and machine learning for improved results, ensuring that possibly the customer's assets are secure and not subject to manipulation. The classifier's feedback procedure aimed to increase its efficacy and detection rate. Except for tree classifiers, random forest, artificial neural networks, vector supporting machines, gradient boosting classifiers, Nave Bayes, and logistic regression techniques, as well as multiple accomplishments, observational analysis of the relevant machine learning strategies has been carried out. In order to determine the best-performing method, evaluation metrics, including recall, precision, F1-score, FPR, and accuracy, have been determined. Any approach with better results for these parameters can be considered the best-performing method. Random forest displays better outcomes in this case than other machine learning classifiers.

VI. CONCLUSION

To conclude, this literature review highlighted the problems faced while generating synthetic data in finance. In addition, this work has also discussed the importance of artificially generated data in finance. Furthermore, this review explored the applicability of financial data in the service sector, such as time series and tabular data. Finally, this review work also briefly explains different properties of desirable representation of artificial data. GANs are employed to augment data effectively. GAN is a class of generative models that can create new data based on actual training data. It consists of two neural networks, a G and a D, which work in opposite directions. The G generates newfangled data instances from actual data. Conversely, the D appraises the synthetic data for validity. The applicability of GANs is high, and they can be used in various fields, such as credit card fraud detection.

To further conclude, deep learning models have transformed our daily life since they are doing well in mitigating real-world challenges. However, the issue of data scarcity is a significant problem as a large quantity of data is required to test the authenticity of data. Augmentation of data is an effective way to deal with scarce data. Augmentation of data can be done via several techniques. Moreover, GAN has an excellent scope for future research as this data augmentation technique is beneficial in multiple fields.

VII. References

[1] S. Assefa, D. Dervovic, M. Mahfouz, R. Tillman, P. Reddy and M. Veloso, "Generating synthetic data in finance: opportunities, challenges and pitfalls," *the First ACM International Conference on AI in Finance,* no. 1, pp. 1-8, 2020.

[2] C. Bowles et al. "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv,* no. 2, p. 1810.10863, 2018.

[3] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data,* vol. 6(1), no. 3, pp. 1-48, 2019.

[4] A. Langevin, T. Cody, S. Adams and P. Beling, "Synthetic data augmentation of imbalanced datasets with generative adversarial networks under varying distributional assumptions: A case study in credit card fraud detection," *Journal of the Operational Research Society,* no. 4, pp. 1-28, 2021.

[5] G. Cawley and N. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research,* vol. 11, pp. 2079-2107, 2010.

[6] Z. Liu et al. "Self-paced ensemble for highly imbalanced massive data classification," *IEEE,* vol. IEEE 36th international conference on data engineering (ICDE), pp. 841-852, 2020.

[7] M. Bansal, D. Sharma and D. Kathuria, "A systematic review on data scarcity problem in deep learning: solution and applications," *ACM Computing Surveys (CSUR),* vol. 54(10s), no. 1, pp. 1-29, 2022.

[8] I. Goodfellow et al. "Generative adversarial networks," *Communications of the ACM,* vol. 63(11), pp. 139-144, 2020.

[9] A. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré, "Learning to compose domain-specific transformations for data augmentation," *Advances in neural information processing systems,* no. 30, pp. 1-11, 2017.

[10] L. Gatys, A. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 2414-2423, 2016.

[11] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and GANs," *Mathematics,* no. 10(15), p. 2733, 2022.

[12] V. Sampath, I. Maurtua, J. Aguilar Martín and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of big Data,* no. 8(1), pp. 1-59, 2021.

[13] E. Lopez-Rojas and S. Axelsson, "Using financial synthetic data sets for fraud detection research," *RAID,* no. 17, p. 485, 2015.

[14] M. Abadi et al. "Deep learning with differential privacy," *In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security,* pp. 308-318, 2016.

[15] F. Tanaka and C. Aranha, "Data augmentation using GANs," *arXiv,* pp. 1-16, 2019.

[16] E. Choi, S. Biswal, B. Malin, J. Duke, W. Stewart and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," *PMLR,* no. 68, pp. 286-305, 2017.

[17] D. Neu, J. Lahann and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," *Artificial Intelligence Review,* no. 55(2), pp. 801-827, 2022.

[18] C. Bowen and F. Liu, "Comparative study of differentially private data synthesis methods," *Statistical Science,* no. 35(2), pp. 280-307, 2020.

[19] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv,* pp. 1-16, 2018.

[20] L. Xu, M. Skoularidou, A. Cuesta-Infante and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in Neural Information Processing Systems,* vol. 32, pp. 1-11, 2019.

[21] J. Jordon, J. Yoon and M. Van Der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[22] M. Wiese, R. Knobloch , R. Korn and P. Kretschmer , "Quant GANs: deep generation of financial time series," *Quantitative Finance,* no. 20(9), pp. 1419-1440, 2020.

[23] R. Fu, J. Chen, S. Zeng, Y. Zhuang and A. Sudjianto, "Time series simulation by conditional generative adversarial net," *arXiv,* pp. 1-33, 2019.

[24] S. Khatri, A. Arora and P. Agrawal, "Supervised machine learning algorithms for credit card fraud detection: a comparison," *IEEE,* pp. 680-683, 2020.

[25] S. Khemakhem and Y. Boujelbene, "Predicting credit risk on the basis of financial and non-financial variables and data mining," *Review of Accounting and Finance,* vol. 17, pp. 316-340, 2018.

[26] V. Nasteski , "An overview of the supervised machine learning methods," *Horizons,* pp. 51-62, 2017.

[27] N. Trivedi, S. Simaiya, U. Lilhore and S. Sharma, "An efficient credit card fraud detection model based on machine learning methods," *International Journal of Advanced Science and Technology,* vol. 5, pp. 3414-3424, 2020.