Financial Fraud Prevention with Synthetic Data Generation using GAN

Article *in* Arya Bhatta Journal of Mathematics and Informatics · September 2022

DOI: 10.5958/2394-9309.2022.00068.3

2 authors:

Rashi Jaiswal

14 PUBLICATIONS **67** CITATIONS

SEE PROFILE

Brijendra Singh

University of Lucknow

35 PUBLICATIONS **302** CITATIONS

SEE PROFILE

Aryabhatta Journal of Mathematics & Informatics Vol. 14, No. 2, July-Dec., 2022
Double-Blind Peer Reviewed Refereed International Journal ©www.abjmi.com
DOI No. : 10.5958/2394-9309.2022.00068.3
ISSN (Print) : 0975-7139
ISSN (Online) : 2394-9309

FINANCIAL FRAUD PREVENTION WITH SYNTHETIC DATA GENERATION USING GAN

Rashi Jaiswal* and Brijendra Singh**
*ICT Research Lab, Department of Computer Science, University of Lucknow, Lucknow, U.P., India.
**ICT Research Lab, Department of Computer Science, University of Lucknow, Lucknow, U.P.,India.
E-mail : rashijaiswal.rj95@gmail.com, drbri_singh@hotmail.com

ABSTRACT

In the real world, online transactions are increasing rapidly to facilitate the users with better comfort services. At present, it is essential to increase privacy to make the digital world safe and secure and to analyze transactional data timely to prevent transactional fraud. The financial fraud can be prevented by detection and prediction from transactional data to secure future transactions. Various techniques are available for financial fraud detection. However, the major issue is analyzing the user's real data which is too risky and difficult to access. The organizations do not have any right to provide their customers with data due to authenticity and privacy concerns. Therefore, it required to generate the synthetic temporal data. This paper proposes a model for fraud detection with synthetic data generation using GANs. We have illustrated the model on synthetic data to perform the experiment. Results obtained from experiment shows that the proposed model outperforms with synthetic data and visualize the reliability of the generated synthetic Metadata over original data. The proposed model provides the facility to deal with less data availability by providing the best solution on transactional data for fraud detection.

Keywords : Fraud Detection, Prediction, Synthetic data generation, GAN, Temporal data, Financial Data
Mathematics Subject Classification 2020: 53C15, 53C25, 53C55.

I. INTRODUCTION

Online transaction services are rapidly increasing with the liberty to manage daily tasks pleasantly. Online transactions increase the risk concern about the user's account safety and security. In the financial world, various organizations facilitate online transactions for different purposes such as marketing, banking, shopping, sale and purchase, business, etc. where the users do transactions with easiness to pay bills [1]. After the covid19 lockdown situation, users give priority to needfulness for cashless, touchless, and carefree transactions. However, the fraud cases also rapidly growing over time. Various organizations working on safety and privacy concerns to provide safe transactions with the best services to users, where they do a timely survey based on user's transaction history or users data to detect the fraud and prevent the users. There are various techniques provided by the researchers to identify the anomalies in the transactional history of the users. Transactional data is temporal and multivariate type data. Various researchers have discussed the different anomaly detection techniques over temporal data for fraud detection tasks [2] [3][4]. The major problem with financial fraud detection is less data availability due to

authenticity and privacy [5]. This issue creates a challenge for researchers and data scientists to do fraud detection with reliable results.

This paper provides the solution for the above-mentioned problem of fraud detection by generating synthetic data to increase the reliability of fraud detection and prediction. We have proposed a model for fraud detection with synthetic data generation using TimeGAN [6] to detect the fraud and prevent the users accordingly. Synthetic data generation is used to generate sufficient data based on original data to perform a reliable analysis of transaction data with reliable, precise, and accurate results. The illustration of the proposed model has been done on the credit card dataset to detect the fraudsters to prevent the next unsafe transactions. The results show that the proposed model outperforms on Metadata (fusion of original and synthetic data) than the original fewer data. The importance of the proposed model is with the reliability and usability by providing the synthetic data to increase the training data, handling the temporal data generation by using TimeGAN, and prevent from fraud by anomaly detection. The highlights of this paper are as follows:

1. Proposed a Fraudster detection model for the transactional dataset for fraud prevention
2. Handling the problem of imbalanced data in the financial dataset
3. Synthetic data generation by using TimeGAN to train the model with sufficient data
4. Metadata generation by integration of original and synthetic data
5. Comparative analysis to validate the model based on performance score
6. Illustrate the proposed model on the credit card dataset and obtained approximately 2-3% more accuracy.

This paper has organized into six sections: Section I introduce the Problem of fraud detection in transactions and discuss the issues with literature details. Section II elaborates the synthetic data generation and Time GAN. The proposed model for fraud detection using synthetic data generation discusses in Section III. Section IV illustrates the proposed model for fraud detection on the credit card dataset. Section V provides the results and discussion based on experiments. Section VI concludes the paper.

II. SYNTHETIC DATA GENERATION USING GANS

The synthetic data generation has been performed by different Generative Adversarial Network-based Models to solve the problem of less data availability [7]. The GANs came to generate the fake images to expose the computer vision tasks. After that, there are some other GAN-based Model architectures have been proposed for various purposes in different domains such as Seq GAN[8], Info GAN[9], Health GAN[10], CGAN [11], and WGAN [12], etc. Synthetic data generation for the time series data is a challenging task because of its temporal properties.

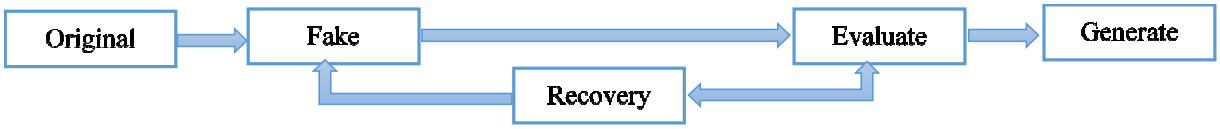

Figure 1. Time GAN Working Process

Time GAN was proposed by Yoon et al. [6] to generate the synthetic time series, which is used in this paper, to generate transactional data for fraud prevention. The synthetic data provides the facility to train the model on a sufficient amount of data to learn accurately and increase the robustness of the training model by providing reliable results. In this paper, we used Time GAN for generating the synthetic transactional data. The Generator generates the synthetic data based on original data and the discriminator checks the data pattern based on original patterns with temporal correlation and dynamic sampling. The process of Time GAN has shown in Figure1.

The synthetic transactional data needs to manage the lacking of data and is useful for prediction and fraud detection. It helps and makes the detection model capable to provide precise results. The precise outcomes define the performance of the model for fraud prevention on the past transactional data and find out the outliers to identify the fraudsters. The proposed Model's detailed descriptions have provided in the next section.

PROPOSED MODEL FOR FRAUD DETECTION USING TEMPORAL SYNTHETIC DATA

The proposed model has been designed with synthetic data generation using GAN to detect fraud from financial data and prevent it. In this paper, proposed model is useful on following problems:

1. The credit card transactional dataset is insufficient, biased, and imbalanced data. The fraudsters are less than normal persons which is the reason for biasness in the subject class.
2. The financial temporal data has the temporal property due to the time dependency of the attributes that create the complexity to compute and detect the fraud from them.
3. The generation of the synthetic multivariate temporal data to create fake data from the original data.

We have proposed a model for fraud prevention by integrating the synthetic data generation and the binary classification technique to detect the fraud and prevent it in the future on financial (transactions) data. Figure 2 shows the architecture of proposed fraud detection model.

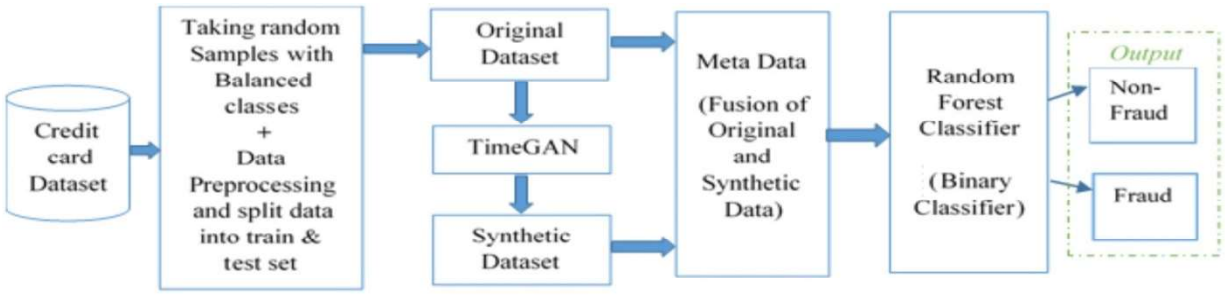

Figure 2. Architecture of Proposed Fraud Detection Model

The proposed model provides the solutions for the above-mentioned problems are as, (a) the balanced dataset samples created to handle the imbalanced transactional data which helps to remove the biases of results, and (b)the time sliding window used to handle the temporal data preprocessing by time indexing, and (c) temporal synthetic generation by using Time GAN. The methodology for implementing the designed architecture has given below:

1. The dataset selection and take as an input and prepare it by data preprocessing and split the data into train and test parts with the ratio of 75:25 respectively.
2. Apply the sliding window technique to convert the data for supervised learning.
3. The input data is assigned as original data and then generate the synthetic data based on original data by using the Time GAN.
4. Generate the Meta Data by fusion of Synthetic and original data.
5. Train the model for outlier detection and classification to classify the real or fraud data from the Meta Data.
6. Performance evaluation of the proposed model based on performance metrics (Precision, Recall, F1-score, and Accuracy) Metrics.

The details of illustration of the proposed model based on the above methodology on transaction data have given in the next section.

ILLUSTRATION OF THE PROPOSED MODEL

This paper has designed the experimental setup based on the proposed model by using the Anaconda IDE with python programming and hardware requirement as 12 GB RAM and i5 processor and the software requirement as windows operating system, Python rich libraries as SK-learn, Pandas, Tensor Flow (Keras). The details of the selected Dataset and model, which have been chosen for the illustration of the proposed model for the fraud prevention task, have given below.

The experiments have been done on the credit card dataset to illustrate the proposed model for fraud detection in transactional data. (*code will be available as per request to author**) The dataset has been taken from the open-source UCI Machine Learning repository [13]. The dataset has 258456 instances, but here, we have taken 1061(569 ones class + 492 zeros class) instances as samples from them for balanced classes in computation to reduce the biases. The selected dataset is multivariate with 31 features/attributes. The target class is represented with 0's and 1's for the non-fraud and fraud respectively. These target classes help to classify the fraud data in the dataset.

As mentioned in the previous section methodology has been followed to do the experiments, where the selected data is taken as input original data with small balanced subsamples. Then after, preprocessing and splitting the data into training and testing parts and train the model with the selected classifier (Random forest classifier [14]) to classify both classes and fetch the outliers for anomaly detection [15].We have chosen a Random Forest classifier based on its popularity in fraud detection and fast computation power with imbalanced classes [16]. The Time GAN was used for synthetic data generated from the original data for both target classes with 600 new samples for each. The fusion of original data and the synthetic data has been done to create Metadata for further process by the selected classifier (Random Forest) for binary classification based on target class of dataset and calculate the fraudsters.

Illustrate the proposed model for fraud prevention by binary classification on selected dataset (credit card). The performance evaluation of the model concerning the credit card data set has been done by

using the performance metrics (precision, recall, f1-score, and Accuracy) for the proposed model and on the base model. Where the precision indicates the positive predictive values, recall indicates the sensitivity, f1_score is the harmonic mean of precision and recall, and the accuracy represents the accurate model measurement. The results of the experiments have detailed discussed in the next section.

RESULTS AND DISCUSSION

Results obtained from experiments shows that the transactional dataset has the biasness in form of imbalanced classes that create a challenge in taking 1061 samples from the dataset. Therefore, we have applied random sampling to take the equalized samples based on target classes. After that, the original data is insufficient so we have applied the Time GAN to generate the synthetic transactional data with 1216 instances based on the original selected samples. The Metadata as final input data is prepared which has given to the proposed model for fraud detection through binary classification task by applying the random forest classifier to detect the fraud based on the target classes' computation and its volume (2277 instances) in the Meta dataset. The experimental study was performed on original and synthetic data for performance evaluation based on performance metrics.

The experiment results shows that the proposed model outperforms for the fraud detection on temporal financial (credit card) data. Where we found that the precision, recall, f1-score, and accuracy increases on Metadata with approx. 2%, 3%, 3%, and 3% respectively. The results have been shown in Table 1 and a comparative analysis on that has also been done based through results visualization on original and Metadata in Figure 3.The comparative analysis clearly shows the importance of the proposed fraud detection model with synthetic data generation where the metadata perform better than the original data. In the future, this proposed model can be upgraded by apply the other advanced technique for the same task in different domains for different detection tasks.

Table 1. Experimental Results on credit card original and Meta dataset

Dataset credit card	Samples Size	Precision	Recall	F1-Score	Accuracy
Original Dataset	1061	0.95008679	0.94835681	0.94817962	0.94835681
Meta Dataset	1216	**0.97172079**	**0.97149123**	**0.97148917**	**0.97149123**

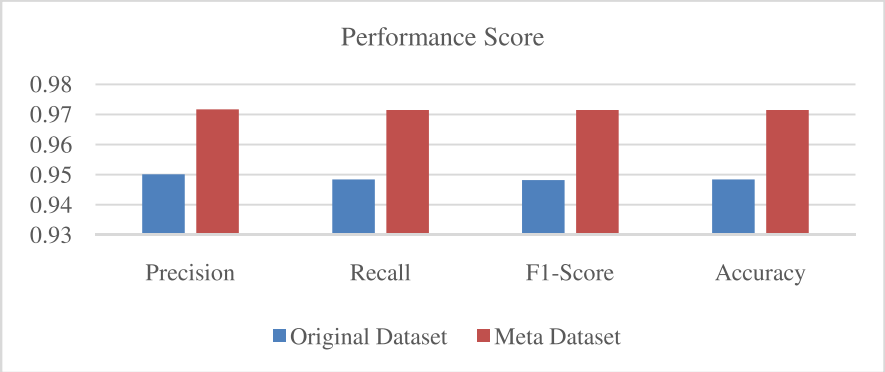

Figure 3. Comparative Analysis visualization on original and Meta Data

CONCLUSION

In this paper, we have proposed a model for fraud prevention by integrating synthetic data generation and binary classification to classify the fraudsters. It help to prevent the next fraud by solving the problem

of less data availability. The illustration of the proposed model has been done through experiments where we obtained that the proposed model outperforms for selected transactional (credit card) data than the original data. The overall performance of the proposed model increases with approx. 2-3% than the base model and original dataset. In the future, we can apply the other techniques to deal with the imbalanced transactional dataset and can use the other advanced model to compute fast and accurately.

REFERENCES

1. *S. S. Mandava, "A Survey of Credit Card Fraud Detection using Supervised Machine Learning Algorithms".*
2. *W. Hilal, S. A. Gadsden and J. Yawney, "A Review of Anomaly Detection Techniques and Applications in Financial Fraud," Expert Systems with Applications, p. 116429, 2021.*
3. *B. Twala, "Multiple classifier application to credit risk assessment," Expert systems with applications, vol. 37, p. 3326-3336, 2010.*
4. *M. Albashrawi, "Detecting financial fraud using data mining techniques: a decade review from 2004 to 2015," Journal of Data Science, vol. 14, p. 553-569, 2016.*
5. *S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," in 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), 2011.*
6. *J. Yoon, D. Jarrett and M. Van der Schaar, "Time-series generative adversarial networks," Advances in Neural Information Processing Systems, vol. 32, 2019.*
7. *A. Langevin, T. Cody, S. Adams and P. Beling, "Generative adversarial networks for data augmentation and transfer in credit card fraud detection," Journal of the Operational Research Society, vol. 73, p. 153-180, 2022.*
8. *M. Alzantot, S. Chakraborty and M. Srivastava, "Sensegen: A deep learning architecture for synthetic sensor data generation," in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2017.*
9. *X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," Advances in neural information processing systems, vol. 29, 2016.*
10. *S. Dash, A. Yale, I. Guyon and K. P. Bennett, "Medical time-series data generation using generative adversarial networks," in International Conference on Artificial Intelligence in Medicine, 2020.*
11. *K. E. Smith and A. O. Smith, "Conditional GAN for timeseries generation," arXiv preprint arXiv:2006.16477, 2020.*
12. *K. E. Smith and A. Smith, "Time Series Generation using a One Dimensional Wasserstein GAN," in ITISE 2019. Proceedings of papers. Vol 2, 2019.*
13. *Data_source, "Datatset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud," 2022.*
14. *R. Jaiswal and B. Singh, "A Study on Classifiers for Temporal Data," in International Conference on Communication Systems and Network Technologies (CSNT-2022) , M.P., India, 2022.*
15. *S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random forest for credit card fraud detection," in 2018 IEEE 15th international conference on networking, sensing and control (ICNSC), 2018.*
16. *P. Verma and P. Tyagi, "Credit Card Fraud Detection Using Selective Class Sampling and Random Forest Classifier," ECS Transactions, vol. 107, p. 4885, 2022.*