Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 2, pp. 413–422
DOI : 10.47974/JIOS-1558

Enhanced credit card fraud detection in online transactions using Long Short Term Memory (LSTM) and RFM analysis with ADAYSN oversampling

Bhumika Rana ^
Tanu Tomar †
Amandeep Kaur §
Parsan Kaur ‡
Simmi Singh @
Department of Information Technology
Guru Tegh Bahadur Institute of Technology
New Delhi 110064
India

Puneet Singh Lamba #
Department of Computer Science
Sri Guru Tegh Bahadur Khalsa College
University of Delhi
Delhi 110007
India

Achin Jain $
Department of Information Technology
Bharati Vidyapeeth's College of Engineering
New Delhi 110063
India

Harsh Taneja *
Department of Computer Science & Engineering
Graphic Era (Deemed to be University)
Dehradun
Uttrakhand
India

^ *E-mail:* **bhumikarana983@gmail.com**
† *E-mail:* **tomartanu647@gmail.com**
§ *E-mail:* **amandeep.gtbit@gmail.com**
‡ *E-mail:* **parsank6@gmail.com**
@ *E-mail:* **simmirooprai@gmail.com**
# *E-mail:* **singhs.puneet@gmail.com**
$ *E-mail:* **achin.mails@gmail.com**
* *E-mail:* **harshtaneja.cse@gmail.com (Corresponding Author)**

Abstract

The expansion of online transactions, particularly online credit card transactions, has revolutionized the field of e-commerce and streamlined electronic payment systems. However, this growth has also given rise to a significant challenge in the form of credit card fraud. To combat this issue, banks and financial organizations have recognized the need for robust credit card fraud detection applications. Machine learning (ML) approaches have emerged as a valuable tool in this regard, as they offer the potential to accurately detect and prevent fraudulent transactions. Long Short-Term Memory (LSTM), a recurrent neural network, is used in this study's evaluation of ML approaches for detecting credit card fraud (CCFD) in online transactions. The most effective LSTM architecture is chosen after a thorough examination based on its capacity to identify credit card fraud with high accuracy and precision. The suggested method makes use of LSTM and RFM analysis to comprehend customer behavior and ADASYN sampling to address class imbalance. The findings show that the selected LSTM architecture, in combination with RFM analysis and ADASYN, delivers great efficiency and efficacy in identifying credit card fraud, hence promoting safe online transactions.

Subject Classification: *94-XX, 91-XX.*

Keywords: *Credit card fraud detection, ADASYN algorithm, LSTM model, Class imbalance, Temporal dependencies, Sequential data.*

1. Introduction

The Federal Trade Commission (FTC) reported 5.1 million reports in 2022, of which frauds accounted for 46% of the total [1]. Credit card fraud made up 43.7 percent of identity thefts, while other identity thefts, which includes email and social media fraud, online shopping and payment account fraud, and other thefts, made up 28.1 percent. It is crucial to remember that fraud detection is an insurmountable battle since fraudsters constantly alter their methods to take advantage of advancing technologies and security vulnerabilities. They shift their focus from older, well-known targets to exploit the weaknesses in new technologies and take advantage of the changing volume and characteristics of legitimate transactions. Recent years have seen the emergence of machine learning (ML) technologies as a possible means of addressing the credit card fraud detection (CCFD) issue. ML algorithms have the potential to analyze large volumes of transactional data, identify patterns, and distinguish between genuine and fraudulent transactions. This research intends to establish a thorough foundation for credit card fraud detection, allowing financial institutions to strengthen security measures and safeguard customers from fraudulent transactions. The main contributions of our proposed fraud detection method are:

1. **RFM Analysis Integration**: It offers valuable insights into customer behavior based on previous transactions.
2. **Handling Class Imbalance with ADASYN**: Used to the class imbalance problem, inherent in CCFD.
3. **Determination of Optimal LSTM Architecture**: The study evaluates several LSTM layer architectures, which include hidden units, dropout layers, and additional dense layers. Through experimentation, the optimal LSTM architecture is determined, maximizing the accuracy and precision of fraud detection.

2. Literature Review

Scientists are always exploring fresh ways to identify and stop fraud. Approaches based on machine learning (ML) have proven to be successful in spotting credit card fraud. The paper [2] introduces APATE, an innovative approach designed to detect fraudulent credit card transactions. Utilizing the Recency-Frequency-Monetary (RFM) paradigm, it makes use of intrinsic qualities that are obtained from the characteristics of incoming transactions and the purchasing patterns of customers. Authors in [3] explores the use of machine learning (ML) and neural networks (NN) to identify potential fraudsters and compares the approach with other ML algorithms such as Naive Bayes, Random Forest Regression, Logistic Regression, and Support Vector Machine. However, the study lacks consideration for class imbalance, potentially impacting model accuracy. Long Short-Term Memory (LSTM) networks, a machine learning model that can assess transaction sequences and capture previous purchasing behavior of credit card holders, are used in the paper [4] for a credit card fraud detection system. In [5], the investigators classified credit card transactions as fraudulent or valid using a deep neural network. To deal with the problem of imbalanced data, they employed a method known as the synthetic minority oversampling technique (SMOTE). They managed a 99.3% accuracy rate. The study [6] employs data-level resampling strategies, including both under-sampling and oversampling methods, to address the issue of data imbalance. By addressing the difficulties of imbalanced datasets and presenting proof that machine learning models, coupled with data resampling techniques, can result in enhanced performance in credit default prediction, it contributes to the field of credit risk prediction. Credit card fraud detection using an artificial neural network [7] contrasted with various machine learning techniques such as the artificial neural network (ANN), support vector machine, and k-Nearest Neighbor. The study [8] focuses on credit card fraud detection (CCFD) using a combination of machine learning (ML) and deep learning (DL) techniques. The proposed model is called the OCSODL-CCFD technique, which stands for "oppositional cat swarm optimization-based feature selection model with a deep learning model for CCFD."

The study [9] tries to resolve the difficulties brought on by the extremely skewed distribution of fraudulent instances in the dataset as well as the profiles of legitimate and fraudulent behaviors in credit card transactions that are frequently changing. The issue of unbalanced datasets in credit card fraud (CCF) detection systems is examined in the paper [10]. It examines current developments in deep reinforcement learning (DRL) and machine learning (ML) algorithms applied to unbalanced datasets, taking fraud and non-fraud labels into account. To handle the imbalance problem in the dataset, two resampling techniques are used: SMOTE and ADASYN. The paper [11] provides a unique method for detecting credit card fraud that makes use of attention-based LSTM deep recurrent neural networks to represent data sequentially.

3. Methodology

3.1 *RFM analysis*

RFM analysis is a valuable technique employed in credit card fraud detection research to analyze past transactions and customer behavior. RFM models are swift to implement and effectively capture customer characteristics, making them widely applicable in customer analysis and segmentation in various industries [16]. In this context, the analysis focuses on grouping customers based on three key dimensions: recency (R), frequency (F), and monetary (M).

3.2 *ADASYN (Adaptive Synthetic Sampling)*

An adaptive oversampling method called ADASYN was created to address class imbalanced problems. Based on each class's unique distributions, the ADASYN algorithm generates synthetic data samples for the minority class in an adaptable manner [13]. Specifically, for a minority sample , its *k* nearest samples are first derived and its imbalance degree is defined as $(\Delta m)/k$, where Δm is the number of majority samples out of the k nearest samples of xm [15] .The objective of ADASYN is to enhance the model's overall performance in identifying instances of the minority class by adaptively modifying the sampling approach [12].

3.3 *LSTM*

Recurrent neural network (RNN) layers called Long Short-Term Memory (LSTM) layers are excellent for identifying these long-term dependencies in sequential data. Because of their distinctive memory cell mechanism, LSTMs can effectively model and learn complex patterns within sequences and retain knowledge over long periods of time. The issue of vanishing and exploding gradients that can be seen during the training of conventional RNNs led to the development of LSTMs [14] [17] [18]. The model architecture combines two LSTM layers. These LSTM layers enhance the model's ability to capture a complex picture of data, enabling it to detect hierarchical and complex relationships among layers. To ensure optimal training stability and feature comparability, a normalization layer is applied at the beginning of the LSTM model as shown in Figure 1. In summary, the incorporation of LSTM layers, along with normalization and global layers, within the credit card fraud detection framework addresses the inherent temporal dependencies in transaction data. By leveraging the memory cell mechanism, the model effectively captures long-term patterns and dependencies, contributing to accurate and robust detection of fraudulent activities.

```
: model.summary()

Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 1, 31)            63
 n)

 lstm (LSTM)                 (None, 1, 128)            81920

 lstm_1 (LSTM)               (None, 1, 128)            131584

 global_average_pooling1d (G  (None, 128)              0
 lobalAveragePooling1D)

 dropout (Dropout)           (None, 128)               0

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 213,696
Trainable params: 213,633
```

Figure 1

LSTM Architecture

4. Implementation

By integrating ADASYN, which mitigates class imbalance by oversampling the minority class, with LSTM's ability to capture temporal dependencies, this hybrid approach aims to enhance fraud detection performance. The basic ADASYN + LSTM fraud detection model is shown in Figure 2. The workflow is as follows:

1. **Data preprocessing:** This includes feature engineering, normalization, and encoding categorical variables as well as application of RFM to understand customer behavioral patterns.
2. **Dealing with an Unbalanced Dataset:** We have used the ADASYN technique, which involves creating synthetic instances of the minority class (i.e., fraudulent transactions), to correct this imbalance.
3. **Feature Selection**: In feature selection, weighted techniques assign importance to features by assigning weights, reflecting their impact on the target variable or the overall model performance.
4. **LSTM Model**: The four-layered architecture, once applied to the preprocessed dataset, serves as a robust framework for credit card fraud detection. By leveraging its design, the model can effectively identify anomalies and detect fraudulent activities within the transaction data.

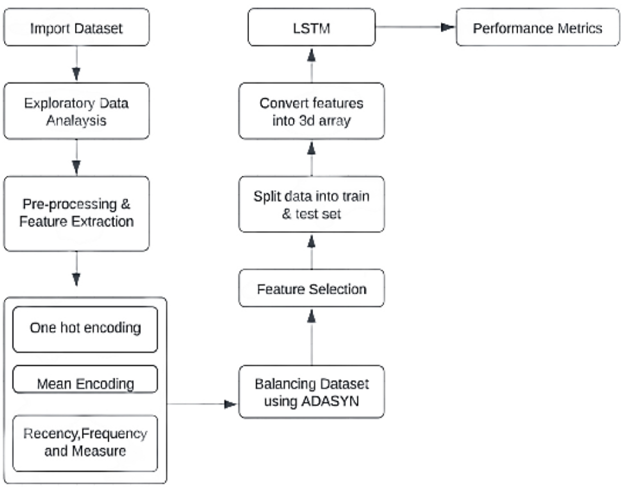

Figure 2

Proposed Approach

5. Experimental Setup

The experiment is conducted using different window widths of 1, 7, and 30 days to capture recent transaction behavior. The ADASYN algorithm was executed to focus on the minority group by oversampling it based on its difficulty level. This technique produced synthetic occurrences for the minority group, thereby mitigating the class imbalance predicament. To verify its validity, other balancing algorithms SMOTE and SMOTEENN were used to process training data and perform comparison. The model is trained using the Adam optimizer, a popular optimization algorithm for data sets and complex models. To evaluate model performance and prevent overfitting, a validation set takes up 20% of the dataset. This validation set examines the performance of the model, which contains hidden data that is not visible during training.

6. Results

The experimental results demonstrate that our proposed approach, which leverages LSTM layers within a broader ML framework, achieves superior performance in terms of efficiency and effectiveness for credit card fraud detection. The comparison table (Table 1) given below, shows that the ADASYN modeling approach leads to a significant improvement in the performance of the model. It achieves a higher precision, recall, and AUPRC scores compared to other sampling techniques, namely SMOTE and SMOTEENN. ADASYN has significantly elevated precision by approximately 0.189 compared to SMOTE and 0.879 compared to SMOTEENN. It has elevated recall by approximately 0.673 compared to SMOTE and 0.154 compared to SMOTEENN. The AUPRC score of 0.991 showcases ADASYN's ability to balance precision and recall effectively.

As the model is trained on the data set, ideally the value of the loss function decreases over successive periods. Decreasing losses indicate that the model is improving its predictions and getting closer to the true values.

Table 1

Comparison Of Various Over Sampling Approaches

Metrics	SMOTE	SMOTEENN	ADASYS
AUPRC	0.56	0.47	0.991
Precision	0.80	0.10	0.989
Recall	0.31	0.83	0.984

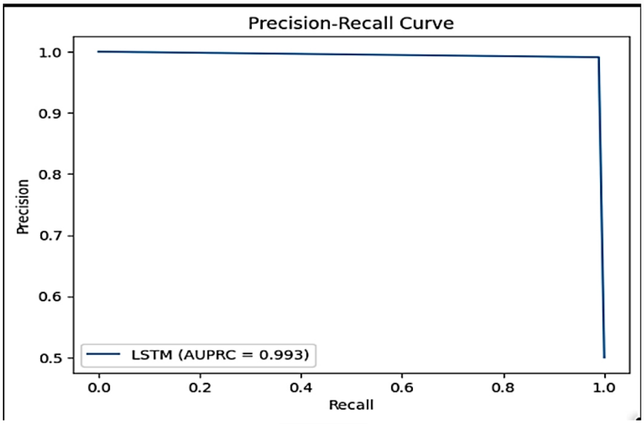

Figure 3

AUPRC Curve

The rapid and continuous decrease in the value of the loss function indicates that the model learns effectively from the data and adjusts its parameters to minimize prediction error. It also helps to see how an optimization algorithm (such as the ADAM optimizer used in this study) refines the model parameters well. Figure 3 shows the AUPRC (Area Under the Precision-Recall Curve) of the ADASYN-LSTM model. The AUPRC score of the ADASYN-LSTM model is reported to be 0.993. A high AUPRC score close to 1 indicates that the model has good performance in handling an imbalanced dataset, exhibiting a strong ability to recover good enough models (fraudulent correlations) while maintaining high accuracy, showing effective reduction of false positives. Overall, the AUPRC curve and the high AUPRC score of 0.993 demonstrate the effectiveness and reliability of the ADASYN-LSTM model in detecting credit card fraud.

7. Conclusion

The proposed approach combines RFM analysis, ADASYN sampling, and LSTM networks to create a comprehensive solution for credit card fraud detection. Overall, this integrated approach can be utilized to improve the accuracy of credit card fraud detection, ultimately enhancing the security and reliability of online payment systems.

For the future advancement of Credit Card Fraud Detection (CCFD), a promising direction involves adopting hybrid models that combine the strengths of diverse techniques. Ensemble methods, including stacking and boosting, can enhance the robustness of the system by integrating traditional machine learning algorithms with deep learning models like LSTM.

References

[1] The statistics on credit card fraud have been sourced from the following website.https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime.

[2] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B., APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38-48 (2015). https://doi.org/10.1016/j.dss.2015.04.013.

[3] Azhan, M., & Meraj, S., Credit Card Fraud Detection using Machine Learning and Deep Learning Techniques. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 514-518 (2020). Thoothukudi, India. doi: 10.1109/ICISS49785.2020.9316002.

[4] Benchaji, I., Douzi, S., & Ouahidi, B., Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks. *Journal of Advances in Information Technology*, 12(2), 113-118 (2021). https://doi.org/10.12720/jait.12.2.113-118.

[5] Khatib, K., Al-Aiad, A., Hussien, M., & Elayan, O., Credit Card Fraud Detection Based on Deep Neural Network Approach. 153-156 (2021). doi: 10.1109/ICICS52457.2021.9464555.

[6] Alam, T. M., et al., An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. IEEE Access, 8, 201173-201198 (2020). doi: 10.1109/ACCESS.2020.3033784.

[7] Asha, R. B., & Suresh Kumar, K. R., Credit card fraud detection using artificial neural network. Global Transitions Proceedings, 2(1) (2021). doi: 10.1016/j.gltp.2021.01.006.

[8] Prabhakaran, N., & Nedunchelian, R., Oppositional Cat Swarm Optimization-Based Feature Selection Approach for Credit Card Fraud Detection. *Computational Intelligence and Neuroscience*, 2023, Article ID 2693022, 1-13 (2023). doi: 10.1155/2023/2693022.

[9] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A., Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Network-

ing and Informatics (ICCNI), pp. 1-9 (2017). Lagos, Nigeria. doi: 10.1109/ICCNI.2017.8123782.

[10] Dang, T. K., Tran, T. C., Tuan, L. M., & Tiep, M. V., Machine Learning Based on Resampling Approaches and Deep Reinforcement Learning for Credit Card Fraud Detection Systems. Applied Sciences, 11(21), 10004 (2021). https://doi.org/10.3390/app112110004.

[11] Benchaji, I., Douzi, S., El Ouahidi, B., et al., Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. *Journal of Big Data*, 8, 151 (2021). https://doi.org/10.1186/s40537-021-00541-8.

[12] Dang, T. K., Tran, T. C., Tuan, L. M., & Tiep, M. V., Machine Learning Based on Resampling Approaches and Deep Reinforcement Learning for Credit Card Fraud Detection Systems. *Applied Sciences*, 11(21), 10004 (2021). https://doi.org/10.3390/app112110004.

[13] He, H., Bai, Y., Garcia, E. A., & Li, S., ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328 (2008). Hong Kong. doi: 10.1109/IJCNN.2008.4633969.

[14] Hochreiter, S., & Schmidhuber, J., Long Short-Term Memory. Neural Computation, 9(8), 1735-1780 (1997). doi: 10.1162/neco.1997.9.8.1735.

[15] Lin, T.-H.; Jiang, J.-R. Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest. *Mathematics*, 9, 2683 (2021). https://doi.org/10.3390/math9212683.

[16] Chen, A.H.-L.; Gunawan, S. Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining. Appl. Sci. 13, 10057 (2023). https://doi.org/10.3390/app131810057.

[17] Singh P, Singh DP. Comparative Analysis of Machine Learning Classifiers for Heart Disease Prediction in Cloud Environment. In2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 552-556 (2023 Mar 15). IEEE.

[18] Joshi A, Rauthan MM, Verma R. Human Mental Health Investigation in Internet Era. In2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), pp. 1-5 (2023 Sep 8). IEEE.