Suspicious Activity Detection Model in Bank Transactions using Deep Learning with Fog Computing Infrastructure

Girish Wali[1], Dr. Chetan Bulla[2*]

[1]Business Intelligence, Citi Bank, India
[2] Business Intelligence, Synechron, India
*bulla.chetan@gmail.com

Abstract. The banking sectors are facing several challenges in detecting and preventing different types of cyber attacks. The main challenge is to find the suspicious activities in money transactions. The majority of Financial institutions are commercial banks suffers lot due to these cyber attacks. The time critical applications requires very small latency in providing services and Cloud computing infrastructures are not well-suited for time-sensitive applications as it takes more latency. Thus, fog computing, an innovative computing paradigm, is utilized to reduce communication latency. The tradition, statistical and machine learning methods effectively identified suspicious activity, but accuracy and trade off between recall and precision is very less. IN this paper a novel suspicious activity detection model is proposed using deep learning with nature-inspired algorithm, to improve the accuracy. The proposed approach analyses transactional patterns in historical data and classify the suspicious and non-suspicious actions. The simulation model is developed using Python programming language and the Google Colab framework to evaluate proposed model. The simulation results show improved accuracy compared to existing state of art works

Keywords: Cyber attack, Suspicious activity detection, Machine learning. Deep learning, Temporal data, Bio-inspired algorithm

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_29

1 Introduction

In the current digital finance landscape, the rise in online transactions has also resulted in the development of advanced methods for fraudulent activities in the banking industry. Advanced technology is required to maintain the safety and authenticity of financial transactions [1]. A conventional rule-based and statistical system unable to capture fraudulent activities as these activities involves new tools and techniques. In this paper, an advanced machine learning models are used to find the suspicious activities with higher accuracy rates in spotting suspicious activity in financial transactions. One practical method for quickly and effectively detecting suspicious activity in financial transactions is to employ deep learning technologies as it gives high accuracy with good efficiency. These also have the ability to reveal intricate patterns and correlations within massive datasets.

Most of financial institution and organizations are migrating to cloud computing as it provides scalability, economic cost, pay-as-you-go, easy to manage, elasticity and many more. But it takes more latency in giving results of complex computational tasks and not suitable for time critical applications such as medical and financial applications. for example, require very low latency and cloud computing is not a good fit because of the substantial communication delays in these areas. The goal of cloud providers in reducing latency is to increase the number of regions and make sure each one has all the features that users need. Consequently, we introduce Fog Computing, a new paradigm in computing that makes use of processing units placed adjacent to data sources [3]. Therefore, it decreases the amount of time it takes for communication to occur. Figure 1 depicts the implementation of fog computing in the banking industry.

It is very essential to find and report suspicious activity to safeguard for bankers as well customer money. Further, these fraud activities affect the trust and reputation of banking firms. Once, the official report say that particular bank has suffered from any fraud activity from hacker or fraudlers, then customer will not invest in that and recover their investment money. So its affecting trust and reputation. Further, if any such cases coming more in market, then its also affect to country economy as foreign investors will not invest and recover their investment if already done. So it is very essential to reduce the fraudulent activity. To reduce fraudulent activity, the basic function called suspicious activity detection is very import to avoid above mentioned affects.

The increasing occurrence of cyber assaults and the advanced nature of fraudulent operations need a departure from conventional methods [5]. In order to proactively and actively prevent harmful actions, these models employ a neural network model

known as the LSTM (Long Short Term Memory) Model [6]. The LSTM Model have capability to learn the pattern of data and adjust the hyperparameters to improve the accuracy. The objective of this proposed model is to find the suspicious activities in the bank transaction with good accuracy and notify the user and bank in stipulated time to avoid the frauds.

The LSTM architecture is a deep learning framework built from the Recurrent Neural Network Model. The system is designed to detect potentially questionable actions and uncover significant trends within the dataset. The main emphasis is on examining temporal data, specifically time-series data. Utilizing a bio-inspired approach [7][12] to optimize hyperparameters and their associated values is crucial for maximizing the performance of the LSTM model. The goal is to attain a well-rounded compromise between recall and accuracy. The related works are discussed in Section 2. Section 3 presents a working principle of proposed model. The experiment and its evaluation are discussed in Section 4. Section 6 provides a summary and suggestions based on the research findings.

2. Literature Review

Logistic regression is a machine learning method employed for the purpose of predicting binary categorization values. This technique does not need the explanatory variables to be correlated or follow a normal distribution [16]. The results of these models are subjective in nature. A significant number of scholars have been employing logistic regression to detect insolvency in the banking industry. A data-driven multi-agent-based anomaly detection technique has been described to find abnormalities in the sensors placed in the fog- computing infrastructure [15]. The LSTM algorithm is employed for anomaly detection in sensor readings. The results of the experiment indicate that the proposed model improves accuracy and recorded 96.2% as highest accuracy.

Decision trees is used to classify a dataset into increasingly smaller subgroups bases on features of data and its values. The data is divided into sub parts based on a specific set of criteria [17] using complex mathematical calculation such as entropy and information gain. The proposed model identifies associations by recursively performing computations, which requires more computational resources. The accuracy of the model's clothing decreases when the data is refreshed. As stated in reference [18], random forests introduce an additional level of uncertainty to the bagging technique. In a random forest, the node division is done by using the best predictor selected randomly from a subset for each node. Bagging and random forest models were constructed using the Random forest package [19]. These combinations provide optimal solution by measuring the significance of each feature in relation to the training da-

taset. Random forest can be employed for video segmentation, picture classification for pixel analysis, and complex biological data processing in the field of bioinformatics. The random forest consumes are computational resource so it is not optimal to use in critical applications. The paper's contribution is Develop preprocessing models to convert data to curated form and developing an LSTM model to identify suspicious activities using past data with highest accuracy.

3 Methodology

The proposed approach is divided into three separate processes. The approach consists of three primary stages: data collection, data pre-processing, and suspicious activity detection. The input data is an historical data that contains various features of bank transactions like sender, receiver, time , place, type of transaction etc. along with two more features suspicious and non-suspicious transactions. The dataset contains all related information of past and current transactions with flagged as suspicious and valid. Figure 1 displays the several phases of the suggested approach for detecting suspicious activities.

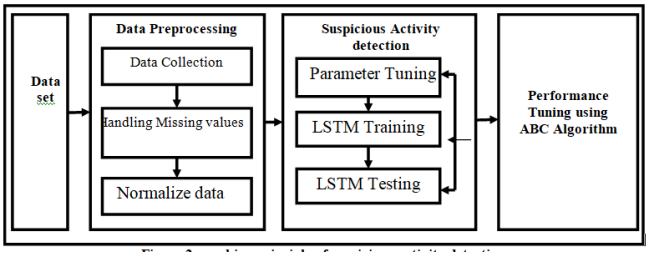

Fig 1: Stages of suspicious activity detection Model

The data is inputted into the model using two methods: historical load and incremental load. The deep learning model utilizes the transactional data for training and testing purposes to detect suspicious actions by analyzing past data patterns. The data is crucial for precisely identifying questionable transactions. Therefore, the data reprocessing method is employed to raw data to convert it into curated data by resolving missing values, outliers, and null values. An LSTM learning model's performance is enhanced using a bio-inspired technique that selects appropriate values for specific hyperparameters.

Data Collection: The data is collected from many streaming probes and stored in a landing zone in as-is format or in raw form. There are two forms of data load in ETL process: full load and incremental load [9]. The full load involves retrieving all data from the historical data store at once, whereas the incremental load updates or adds new data to the existing dataset based on timestamp such as hourly, daily and weekly updates etc. Since the model is specifically designed for bank transactions, we are now

setting up the daily data loading process. The data is loaded at end of the day to process further and depending upon application requirement we can adjust this value [14].

Data Preprocessing: Once the data is collected in landing zone then it will preprocess the raw data and convert it to curated data, which includes tasks such as data cleaning, de-duplication, and normalization[3][23][10]. If quality of data is compromised then it will leads to inaccurate results. So it is essential to clean the raw data to avoid inaccuracy. The proposed model utilizes three data processing approaches: (i) handling missing data and (ii) normalizing the data [22].

Suspicions Activity Detection Model

A Long Short-Term Memory (LSTM) network model is trained on preprocessed data to improve accuracy in detecting potential fraudulent activities. This approach utilizes historical data patterns to identify and highlight potentially suspicious behaviors. Features are extracted from historical data to help recognize potential fraud indicators. The indicators are compared to current transactional trends for analysis. The suspicious activities are identified based the data pattern in temporal dataset. The LSTM model able store longer memories and resolve the vanishing gradient problem [11]. Using the LSTM network model is best choice as compared to other deep learning model to detect specific pattern such as suspicious activities and risk identification. Figure 2 shows LSTM unit and its components .

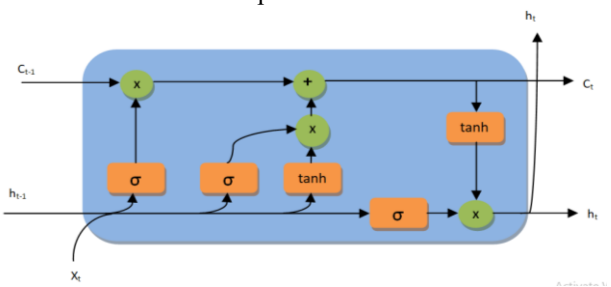

Fig 2: LSTM Model

The LSTM unit has three crucial components: the input gate, forget gate, and output gate. These gates have a specific function and regulating the flow of information within to achieve the goal. The input gate read the data from dataset and filters the required data and passes to enter the cell. The forget gate again remove the irrelevant information that reads from previous cell. Finally, output gate determines the output values produced by the LSTM cell and passed to next LSTM cell.

At time t, the input value xt is fed into the network. Input gate read the values and keeo the relavant data only. Other two gates evaluates incoming data and regulates it

to next gate and cell respectively using following formulae

$$f_t = \sigma(Wt_f.[h_{t-1,}X_t] + bs_f \quad (6)$$

$$i_t = \sigma(Wt_i.[h_{t-1,}X_t] + bs_i \quad (7)$$

$$\sim C_t = tanh(Wt_c.[h_{t-1,}X_t] + bs_c \quad (8)$$

$$C_t = f_t * C_{t-1} + i_t * \sim C_t \quad (9)$$

$$O_t = \sigma(Wt_o.[h_{t-1}, x_t] + bs_o) \quad (10)$$

$$h_t = O_t * tanh(C_t) \quad (11)$$

$$.y_t = O_n(y_{t,}y_{t+1,}y_{t+2}....y_{t+n-1}) \quad (12)$$

The LSTM network receives these sequences and extracts the features and learns the data patterns (suspicious and normal transaction) from test data and return the result in binary form. If the yt is 1 then it is suspicious activity and otherwise it is considered as normal transaction

Hyperparameter optimization

To improve or optimize the learning behaviour, it is essential to tune the performance parameters. So the bio-inspired method ABC algorithm that mimics the foraging behaviour of bee colonies [13]. ABC distinguishes three distinct sorts of bee colonies. Worker bees employ their memory to forage for food in close proximity and share this information with other bees. Through the use of this data, observer bees may discern nectar of superior quality. Upon the discovery of novel food sources, scout bees underwent a transformation and assumed the role of worker bees. In order to enhance precision, the recommended model fine-tunes its hyperparameters utilizing the ABC technique [15]. The fundamental objective of ABC optimization techniques is to identify the most effective combination of hyperparameter values that can enhance the accuracy of detecting suspicious activities. The LSTM model utilizes ABC foraging behaviour to identify the optimal hyperparameters for obtaining the highest degree of accuracy in multistep suspicious activity detection. \

4 Results and Discussions

In this section, simulation setup, dataset used for simulation, and performance parameters are discussed. The proposed model is compared to the state of art related works on suspicious activity detection methods [8] [19][20]. The Python programming language in Google Colab framework is used to design and develop the simulation model for suspicious activity detection. Python libraries for deep learning called Keras and matplotlib are used to implement the simulation model [22].

The dataset utilized consists of transaction details, client information, and book information. The score column is computed using different criteria if the transaction value is more than or equal to 10,000. The risk level also plays a role in calculating the score and determining outcomes based on predefined rules, typically as part of the ETL logic. The "is_Alerted" column indicates a transaction that has been created as an alert by the ETL system, whereas the "is_Suspicious" column indicates a transaction that has been reported as a genuine suspicious activity by a Compliance Person.

To assess the proposed model, various standard performance parameters are considered, namely, accuracy, recall, and F1-score. These are most common and standard performance parameters to evaluate the machine learning and deep learning models. The Root Mean Square Error (RMSE) is the most suitable choice to calculate and reduce the error

Results

The first result, achieved using default hyperparameters, shows an accuracy of 96.60% and a loss rate of 0.4% in 106 seconds. The accuracy is improved when the bio-inspired algorithm is used: 0.1. The proposed model requires many training sessions whenever new data is introduced. As a result, it consumes a significant amount of resources in order to achieve high accuracy. The table shows simulation results for different values of hyperparameters.

Table 1: Sample experiment results for tuning hyperparameters

Sl. No	Sliding-window size	LSTM units	Dropout rate	Regularizer rate	Optimizer
1	10	30	0.4	L2	RMPro
2	20	30	0.3	L1	Adagrad
3	30	40	0.28	L1	Adam
4	40	40	0.2	L2	Adagrad
5	50	40	0.25	L1	Adam
Epochs	**Activation function**	**Learning rate**	**Accuracy**	**Loss**	
100	ReLu	0.1	96.6	0.35	
200	Sigmoid	0.2	96.9	0.31	
300	Tanh	0.18	97.4	0.25	
400	ReLu	0.23	96.2	0.37	
500	Sigmoid	0.2	95.9	0.4	

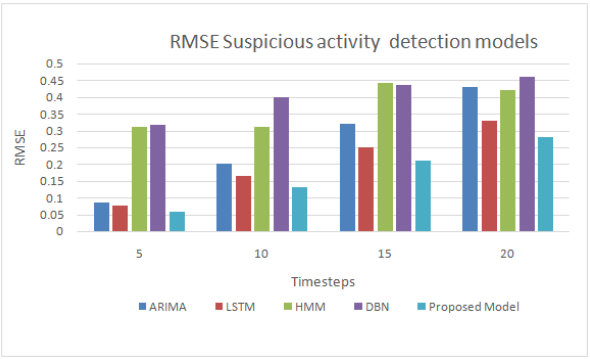

Fig 3: RMSE values for existing models and proposed model.

Figure 5 shows a comparison of the standard deep learning accuracy parameters such as recall, accuracy, and F1-score The proposed models' conclusions are compared with other related and existing state of art works: auto-encoders [22], LSTMs [5], and RNNs. The proposed model achieves higher accuracy, recall, and F1-score as compared to other models.

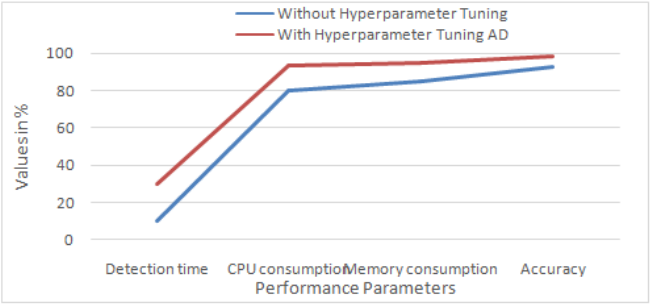

Fig 4 : Normal SAD Model versus Hyperparameter Tuner SAD Model

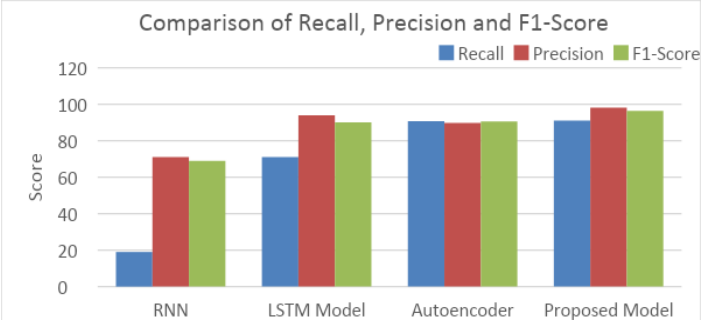

Fig 5: Comparison of the existing and proposed system' accuracy parameters

Table 2 shows the combination of hyperparameters that resulted in a 99.78% accuracy rate. The experimental results show that hyperparameter optimization gives good accuracy as compared to other existing models

Table 2: Highest accuracy using hyperparameter tuning in simulation result

sliding-window-size	LSTM-units	dropout-rate	regularize	Regularizer-rate	optimizer
20	50	0.3	L2	0.02	RMSProp
epochs	activation function	learning-rate	accuracy	Loss	
100	ReLu	0.45	99.78	0.21	

5 Conclusion

Detecting suspicious behaviour in bank transactions is crucial for minimizing financial losses for both banks and their consumers. This study introduces a data-driven algorithm for detecting suspicious activity, aiming to uncover questionable behaviors that occur in bank customers' transactions. To reduce the latency of response, a fog computing model is used and to detect suspicious activity, a deep learning model, called LSTM model, is used. The LSTM model reads the relevant data and find the data pattern for suspicious and normal transaction. To improve the accuracy, a bio-inspired model called Artificial Bee colony is integrated. The simulation results demonstrated that the proposed model improves accuracy compared to state of art works in literature. The future work involves risk assessment and compliance report generation using GenAI Models.

References

[1] Santomero, A.M. (1999). Risk Management in Banking: Practice Reviewed and Questioned. In: Galai, D., Ruthenberg, D., Sarnat, M., Schreiber, B.Z. (eds) Risk Management and Regulation in Banking. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-5043-3_3

[2] Chetan M Bulla, Satish S Bhojannavar, Vishal M Danawade,"Cloud Computing: Research Activities and Challenges",Volume 2, Issue 5, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2013.

[3] Resul Das, Muhammad Muhammad Inuwa,"A review on fog computing: Issues, characteristics, challenges, and potential applications",vo. 10, Telematics and Informatics Reports,2023.https://doi.org/10.1016/j.teler.2023.100049.

[4] Shi, S., Tse, R., Luo, W. et al. Machine learning-driven credit risk: a systemic review. Neural Comput & Applic 34, 14327–14339 (2022). https://doi.org/10.1007/s00521-022-07472-2

[5] Fang Dao, Yun Zeng, Jing Qian,"Fault diagnosis of hydro-turbine via the incorporation of bayesian algorithm optimized CNN-LSTM neural network",vol 290, Energy,Elsevier, 2024.

[6] Al-Sawwa, Jamil, and Mohammad Almseidin. "A Spark-Based Artificial Bee Colony Algorithm for Unbalanced Large Data Classification." *Information*, vol. 530, no. 11, 8 Nov. 2022, https://doi.org/10.3390/info13110530..

[7] Albatul Albattah & Murad A. Rassam. (2022) A Correlation-Based Anomaly Detection Model for Wireless Body Area Networks Using Convolution Long Short-Term Memory Neural Network. Sensors 22:5, pages 1951.

[8] Bulla, Chetan M., and Mahantesh N. Birje. "A Multi-Agent-Based Data Collection and Aggregation Model for Fog-Enabled Cloud Monitoring." *IJCAC* vol.11, no.1 2021: pp.73-92. http://doi.org/10.4018/IJCAC.2021010104

[9] Madhavi, K. Reddy, K. Suneetha, K. Srujan Raju, Padmavathi Kora, Gudavalli Madhavi, and Suresh Kallam. "Detection of COVID-19 using X-ray Images with Fine-tuned Transfer Learning." (2023). Journal of Scientific & Industrial Research (JSIR), National Institute of Science Communication and Information Resources (NISCAIR) by CSIR, Govt of India, 82,2, pp.241-248, 2023

[10] Pioli, L., Dorneles, C.F., de Macedo, D.D.J. et al. An overview of data reduction solutions at the edge of IoT systems: a systematic mapping of the literature. Computing 104, 1867–1889 (2022). https://doi.org/10.1007/s00607-022-01073-6

[11] Hongjian Li, Liangjie Liu, Xiaolin Duan, Hengyu Li, Peng Zheng, Libo Tang,"Energy-efficient offloading based on hybrid bio-inspired algorithm for edge–cloud integrated computation",vol 42,Sustainable Computing: Informatics and Systems,, elsevier, 2024.

[12] Bulla, C., Birje, M.N. (2022). Anomaly Detection in Industrial IoT Applications Using Deep Learning Approach. In: Fernandes, S.L., Sharma, T.K. (eds) Artificial Intelligence in Industrial Applications. Learning and Analytics in Intelligent Systems, vol 25. Springer, Cham. https://doi.org/10.1007/978-3-030-85383-9_9

[13] Reddy Madhavi, K., A. Vinaya Babu, G. Sunitha, and J. Avanija. "Detection of concept-drift for clustering time-changing categorical data: An optimal method for large datasets." In Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, pp. 861-871. Springer Singapore, 2020.

[14] Aysha Shabbir, Maryam Shabir, Abdul Rehman Javed, Chinmay Chakraborty, Muhammad Rizwan, Suspicious transaction detection in banking cyber–physical systems, Computers & Electrical Engineering, Volume 97,2022,Elsevier

[15] Avanija, J., K. E. Kumar, Ch Usha Kumari, G. Naga Jyothi, K. Srujan Raju, and K. Reddy Madhavi. "Enhancing Network Forensic and Deep Learning Mechanism for Internet of Things Networks." (2023). Journal of Scientific & Industrial Research (JSIR), National Institute of Science Communication and Information Resources (NISCAIR) by CSIR, Govt of India, Vol. 82, May 2023, pp. 522-528,2023,DOI: 10.56042/jsir.v82i05.1084

[16] Alghuried, A. A Model for Anomalies Detection in Internet of Things (IoT) Using Inverse Weight Clustering and Decision Tree., Technological University, Dublin, 2017.

[17] P. Zhang and Y. Liu, "Application of An Improved Artificial Bee Colony Algorithm," IOP Conf. Ser.: Earth Environ. Sci., vol. 634, no. 1, p. 012056, Feb. 2021.

[18] C Bulla, MN Birje."Improved data-driven root cause analysis in fog computing environment", 8 (4), 359-377, Journal of Reliable Intelligent Environments, Springer, 2022.