World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:19, No:2, 2025

An Attentional Bi-Stream Sequence Learner for Credit Card Fraud Detection

Mohsen Hasirian, Amir Shahab Shahabi

Abstract**—**Modern societies, marked by expansive Internet connectivity and the rise of e-commerce, are now integrated with digital platforms at an unprecedented level. The efficiency, speed, and accessibility of e-commerce have garnered a substantial consumer base. Against this backdrop, electronic banking has undergone rapid proliferation within the realm of online activities. However, this growth has inadvertently given rise to an environment conducive to illicit activities, notably electronic payment fraud, posing a formidable challenge to the domain of electronic banking. A pivotal role in upholding the integrity of electronic commerce and business transactions is played by electronic fraud detection, particularly in the context of credit cards which underscores the imperative of comprehensive research in this field. To this end, our study presents an Attentional Bi-Stream Sequence Learner (AttBiSeL) framework that leverages attention mechanism and recurrent networks. By incorporating bidirectional recurrent layers, specifically bidirectional Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) layers, the proposed model adeptly extracts past and future transaction sequences, while accounting for the temporal flow of information in both directions. Moreover, the integration of an attention mechanism accentuates specific transactions to varying degrees, as manifested in the output of the recurrent networks. The effectiveness of the proposed approach in automatic credit card fraud classification is evaluated on the European Cardholders' Fraud Dataset. Empirical results validate that the hybrid architectural paradigm presented in this study yields enhanced accuracy compared to previous studies.

Keywords**—**Attention mechanism, credit card fraud, deep learning, recurrent neural network.

I. Introduction

SOCIETIES globally are now profoundly interconnected due to the expansion of the Internet and the emergence of electronic commerce applications. The speed, efficiency, and ease of access offered by electronic commerce have attracted enormous number of customers, making it a prominent channel for buying and selling. Within this context, electronic banking has witnessed remarkable growth in recent years, becoming a cornerstone of online activities, with millions of financial transactions being processed every moment. However, this rapid expansion has concurrently given rise to opportunities for fraudulent activities. As a result, electronic payment fraud has emerged as a critical challenge in the realm of electronic banking. Fraudsters exploit vulnerabilities inherent in electronic systems and target the weaknesses of users engaged in non-physical/virtual service utilization to achieve their illicit objectives. This evolving landscape has led to an escalation in the strategies employed by fraudsters and deceivers within the electronic domain [1].

The fraud detection in payment processes using electronic banking systems (such as credit card payments) have a significant impact on the security of transactions, and financial transfers [2]. With the aim of mitigating the risks associated with electronic payment in financial systems, researchers intend to examine different perspectives and propose strategies to minimize these risks. The primary approach involves the collection and analysis of data from banking account transactions, which will be utilized to detect instances of fraudulent activities. The outcome will be the provision of an impactful tool for addressing these challenges. An effective tool for aiding internet banking systems and detecting electronic fraud is the utilization of intelligent techniques based on Deep Learning (DL) [3], [4]. Utilizing these methods, patterns of online fraud in financial transactions can be extracted by a learner model, reducing the error in fraud detection caused by human factors. These electronic fraud detection systems, particularly in the context of credit cards, play a significant role in the perceived trustworthiness of electronic commerce and businesses. Hence, studying and researching this topic is of utmost importance.

A comprehensive examination of the existing literature in DL and fraud detection domains demonstrates that the utilization of DL models has yielded promising results in the field of fraud detection. Among various sequential models used for fraud detection, LSTM [5] and GRU [6] networks have gained substantial interest. While these networks have the capability to process sequences of varying lengths, utilizing them as feature extractors in a deep architecture lead to high-dimensional feature spaces. Another constraint of these networks is their identical emphasis on distinct features. To address the aforementioned issues, we present a novel framework called the Attentional Bi-Stream Sequence Learner (AttBiSeL) approach, which is based on the attention mechanism [7] and recurrent networks. The proposed AttBiSeL model employs bidirectional recurrent networks, specifically bidirectional LSTM and GRU networks, to extract past and future transaction sequences while taking into account the temporal flow of information in both directions. Moreover, the utilization of the attention mechanism is employed on the output of the recurrent networks in order to accentuate different transactions with varying levels of emphasis. Our experimental results have substantiated the notion that the incorporation of attention mechanisms can

Mohsen Hasirian is with Department of Electrical Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran (e-mail: mhhasirian@gmail.com)

Amir Shahab Shahabi is with Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran (e-mail: shahabi_amir@azad.ac.ir)

significantly improve the efficacy of deep neural networks [8]. Hence, the utilization of Bi-Stream Sequence Learner has the potential to enhance accuracy in comparison to prior methodologies. The efficacy of the proposed model has been assessed in the context of automated credit card fraud classification, a significant and widespread undertaking in the field of electronic banking analysis. The experiments were carried out using the European Card Holders Fraud Data [9]. Furthermore, the obtained results have been compared to prior research in the domain of fraud detection, specifically in terms of the accuracy attained by the proposed model.

The organization of the paper is as follows: Section II is devoted to the review of previous methods. Section III explains the proposed method and its constituent components. Section IV presents the implementation results and the evaluation of the results. Finally, Section V concludes the study.

II. Related Works

A multitude of studies have been conducted in the field of automatic credit card fraud detection. In the following sections, we will briefly delve into the methods previously researched, and present relevant insights and considerations.

A. Machine Learning-Based Methods

A wide spectrum of machine learning approaches has been employed to detect and classify credit card fraud, encompassing Unsupervised Learning (UL), and Supervised Learning (SL) [2]. UL techniques are utilized for detecting abnormal behavior within a system and identifying transactions that deviate from normal behavior, indicating potential instances of fraudulent activities [10]. These approaches can facilitate the detection of novel fraud patterns that have not been previously acknowledged. Supervised classification methods have proven highly effective in tackling this challenge, where pre-labeled historical transaction datasets are employed for training classifiers. A classifier constructs a model capable of predicting the fraudulent or legitimate nature of a new transaction. Conventional machine learning algorithms that are commonly used include Support Vector Machine (SVM) [11], Hidden Markov Model (HMM) [12], Decision Tree [13], Random Forest (RF) [14], K-Nearest Neighbors (KNN) [15], and AdaBoost [16].

Xuan et al. [17] employed two types of RF algorithms (i.e., tree-based and Classification And Regression Tree (CART)) for modeling fraudulent and legitimate (a.k.a. normal) transactions. Their results indicate that the first model exhibits superior performance compared to the second model in terms of accuracy, recall, and F1-score. Additionally, they stated that the second model performs better when the classes in the dataset are imbalanced compared to the first model. Makki et al. [18] conducted an empirical analysis of eight classification methods (including, artificial immune systems, naive Bayes, decision trees, and neural networks) under imbalanced dataset conditions and elucidated their pros and cons on real-world credit card datasets. The experimental findings suggest that the methods employed to address imbalanced data primarily improve the performance of classifiers. However, it is observed that these methods often result in a significant increase in the occurrence of false positives, which is considered undesirable.

In certain research studies, the utilization of clustering algorithms during the data preprocessing phase has been observed to improve the performance of classification models. To this end, Jiang et al. [19] clustered similar credit card holders based on historical transaction patterns into similar subsets. Subsequently, they utilized the sliding window concept to leverage the advantage of aggregating transactions within each subset. Furthermore, through the examination of behavioral patterns exhibited by individual cardholders, multiple classifiers were trained for each respective subset. Lastly, a feedback mechanism was utilized in order to tackle the issue of concept drift. According to the authors, their proposed methodology demonstrates superior performance compared to existing models in terms of average recall and accuracy.

Nevertheless, the detection of credit card fraud poses various obstacles that have caught the attention of academics in the field of artificial intelligence for a variety of reasons. One of the issues encountered with credit card fraud datasets pertains to their extremely imbalanced nature, when the count of legitimate transactions greatly surpasses that of fraudulent transactions. Consequently, several conventional classifiers have difficulties in accurately detecting instances belonging to the minority class inside imbalanced datasets [20]. However, conventional classifiers are designed to analyze transaction details such as location, amount, and date in order to identify potentially fraudulent transactions [21]. The primary issue with these methodologies lies in their failure to consider the precise sequential information that characterizes users' profiles. The abovementioned models are considered inadequate for credit card fraud classification task due to their omission of behavior of the consumer in purchasing. This behavior is crucial in identifying pertinent fraud trends that change over time as a result of seasonal variations and ever-evolving attack techniques [22].

B. Deep Learning-Based Methods

DL is widely recognized as a dominant and auspicious domain within the realm of machine learning, having demonstrated remarkable advancements in diverse real-world applications. Prior studies encompass a range of methodologies that employ Deep Neural Networks (DNNs), including Convolutional Neural Networks (CNNs) [23] and Recurrent Neural Networks (RNNs) [24]. RNN is a variant of dynamic DNN that demonstrates the capability of interpreting the evolving temporal patterns exhibited by diverse financial accounts. This is accomplished through precisely representing the dependencies that exist among consecutive transactions made by credit cardholders. The primary benefit of RNNs resides in their domain-agnostic methodology for detecting credit card fraud, surpassing the performance of CNNs by exhibiting superior results. Given the capacity of RNNs and their variants, namely LSTM and GRU, to effectively capture long-term dependencies, these models are extensively utilized in the field of credit card fraud detection. Recently, attention mechanisms have been employed to improve the efficacy of

DNNs by allowing them to concentrate on where to learn; for instance, attention-based methods have demonstrated notable success in the fields of machine translation [25] and image captioning [26].

Jurgovsky et al. [5] employed LSTM models to gather cardholders purchase profiles, aiming to improve the accuracy of fraud detection in upcoming transactions. The authors demonstrated that physical and virtual transactions exhibit distinct behaviors as a result of their sequential nature. They illustrated that when dealing with physical transactions, employing a LSTM model is adequate for capturing concealed sequential patterns and enhancing the modeling of transaction history. Furthermore, by employing LSTM, a distinct set of true positive instances will be present, which is notably different from the fraudulent cases detected by the RF algorithm. Roy et al. [6] conducted a series of empirical experiments to provide evidence that the performance of two LSTM and GRU networks outperforms that of the conventional RNN. They inferred that the sequential arrangement of transactions within an account possesses significant value in distinguishing between fraudulent and legitimate transactions. Yang et al. [27] introduced a novel attention-based network known as Hierarchical Attention Networks (HAN) to address the task of text classification. They utilized two levels of attention modules, one for word-level attention and another for sentence-level attention. These attention modules are applied on the outputs of sequence-based encoders utilizing a GRU model. These attentional networks are placed in succession on the results of GRU-based sequence learners. Lebichot et al. [28] proposed two novel approaches for domain adaptation, which is a specialized form of transfer learning, by incorporating transfer learning into the context of DL. The first methodology involved improving the accuracy of predictions by integrating features derived from e-commerce transactions with those of physical transactions. The second methodology involves training the model to acquire domain-specific features by optimizing the feature layer in order to predict fraud and domain labels. The researchers conducted an evaluation of their models using real-world credit card transaction datasets. Their findings indicate that these two methods demonstrate performance that is comparable to existing approaches in the field. Fu et al. [29] proposed a CNN-based credit card fraud detection framework due to CNNs' capacity in training on massive data volumes and ability to mitigate overfitting issues. Furthermore, the concept of transactional entropy was introduced in order to incorporate the influence of intricate customer behaviors on costs. Through the integration of transactional features and feature metrics within a CNN architecture, the researchers effectively showcased the superior performance of their model in comparison to other methodologies documented in the existing body of literature. Somasundaram et al. [30] proposed a novel approach named Transaction Window Bagging (TWB), which utilizes ensemble learning techniques for credit card fraud classification. Their parallelized ensemble technique demonstrates promising potential in effectively addressing the challenges associated with detecting and classifying fraudulent credit card transactions. The objective of this model is to effectively address the phenomenon of concept drift through the utilization of incremental learning technique. Furthermore, the authors assert that their model possesses the ability to address data imbalance through the reduction of training bias. The empirical experiments demonstrate that their proposed method has the potential to improve the efficiency of fraud detection systems, while concurrently decreasing latency.

III. Proposed Method

This paper presents a framework for credit card fraud detection, utilizing sequential data modeling techniques. It specifically leverages recurrent networks (i.e., LSTM and GRU networks) in conjunction with the incorporation of an attention mechanism. In contrast to prior studies, the key advantage of the proposed method is its acknowledgment of the sequential characteristics inherent in transaction data. This enables the model to discern the most crucial transactions within the input sequence, improving the prediction accuracy for fraudulent transactions. The main features of the proposed model are constructed by combining the attributes of three separate models. The combination of sequential LSTM and GRU models with an attention mechanism serves the purpose of uncovering temporal correlations among events that may be widely separated within the input sequence. This enhancement significantly improves the classification efficiency and results in improved detection of fraudulent transactions compared to traditional models. Furthermore, the attention mechanism plays a pivotal role in enhancing the performance of LSTM and GRU models by enabling varying degrees of emphasis on diverse transactions.

As depicted in Fig. 1, our proposed method is composed of five stages. In the first stage, data preprocessing techniques are applied in order to place the input data into a suitable space. In the second stage, an oversampling method is applied to the minority class of the training dataset to alter the class distribution. In the third stage, bidirectional LSTM and GRU networks are employed as a sequence learner to extract and capture the salient dynamic pattern of sequential dependencies between consecutive credit card transactions. In the fourth stage, the incorporation of the attention mechanism is aimed at tailoring the focus on output information emerging from the hidden layers of the LSTM and GRU networks. This mechanism empowers our model to discern pertinent fraud patterns and to more effectively pinpoint highly diverse transactions within consumer purchases. In the last stage, the softmax classifier is employed for classification of the features obtained from the attention layer. Further details about each constituent part of the proposed method will be elaborated in the following sections.

A. Data Preprocessing

Data preprocessing involves the preparatory steps performed on data before it is fed into a network for training. One practical technique in data preprocessing is standardization. Standardization is employed to transform the data in such a way that it follows a standard normal distribution with a mean of 0 and a standard deviation of 1. This transformation is often a

prerequisite for DL models, which require data to be standardized before input. Neglecting to standardize the data could potentially impact the model's performance. In this study, we have applied one of the widely used standardization methods, known as the z-score [31], which is defined by (1):

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \tag{3}$$

where x, μ and σ correspond to the value of the data point, the mean and standard deviation of the input data, respectively.

B. Class Imbalance Problem and Data Resampling

In the majority of real-world applications, an imbalanced class distribution is common, meaning that one class contains significantly more samples than another. This issue is particularly pronounced in certain applications, including fraud detection, where the number of fraud cases is exceedingly low in comparison to legitimate cases. Most machine learning algorithms do not perform well in the presence of imbalanced classes. More precisely, the predictive model tends to classify minority class samples as majority class samples. To address this problem, in this study, the Synthetic Minority Over-sampling Technique (SMOTE) [32] is employed as an oversampling method. SMOTE is one of the most well-known and widely used oversampling techniques employed for balancing dataset distribution. As illustrated in Fig. 2, the main objective of this method is to generate new synthetic samples for the minority class by interpolating between several neighboring minority samples, rather than oversampling with replacement. Consequently, it mitigates the risk of overfitting to the training data. Depending on the required level of oversampling, the closest neighbors of minority class samples are randomly selected.

C. Proposed Model

As depicted in Fig. 3, the proposed network is constructed with two branches for separately extracting features from input transactions, followed by a feature fusion layer (specifically, a concatenation layer) to combine the features extracted from each path. The architecture of these sequence learners is designed for feature learning and fusion on the input transactions, ultimately utilizing the fused features for more precise classification. The core layers of the network include a bidirectional GRU layer, a bidirectional LSTM layer, an attention mechanism layer, a fully connected (FC) layer, and the softmax layer. It is important to note that the incorporating bidirectional LSTM and GRU layers in the proposed model enables the proposed model to capture not only past but also future contexts by considering the temporal information flow in bilateral directions. Further details regarding each constituent component of the proposed network will be elaborated upon in the subsequent sections.

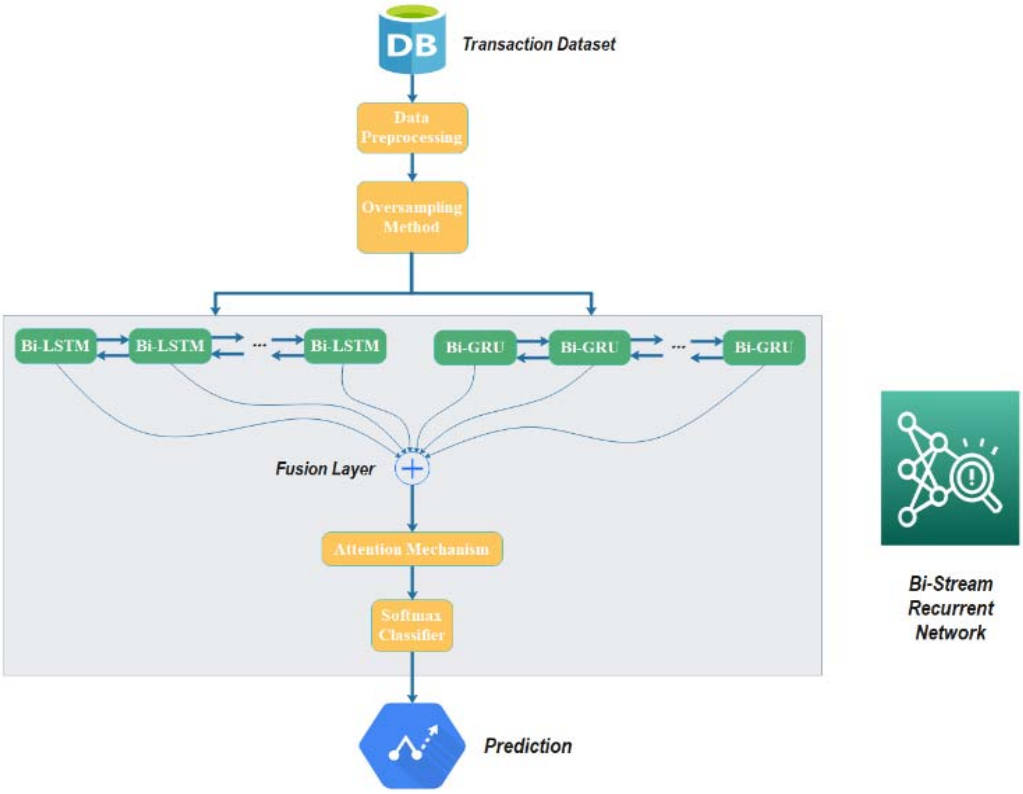


Fig. 1 The pipeline of proposed credit card fraud detection system

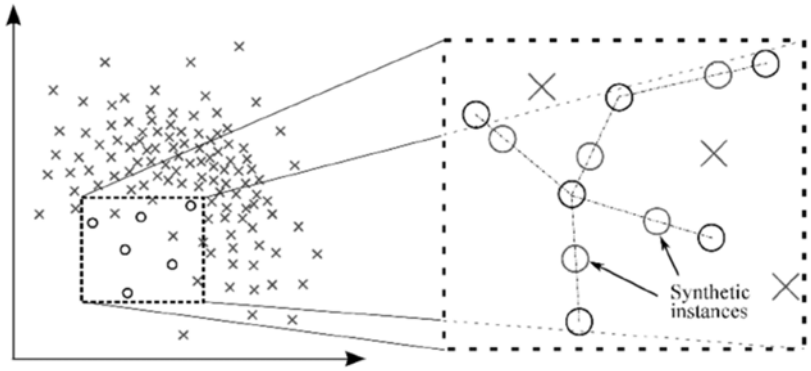


Fig. 2 The generation of synthetic samples using the SMOTE method [32]

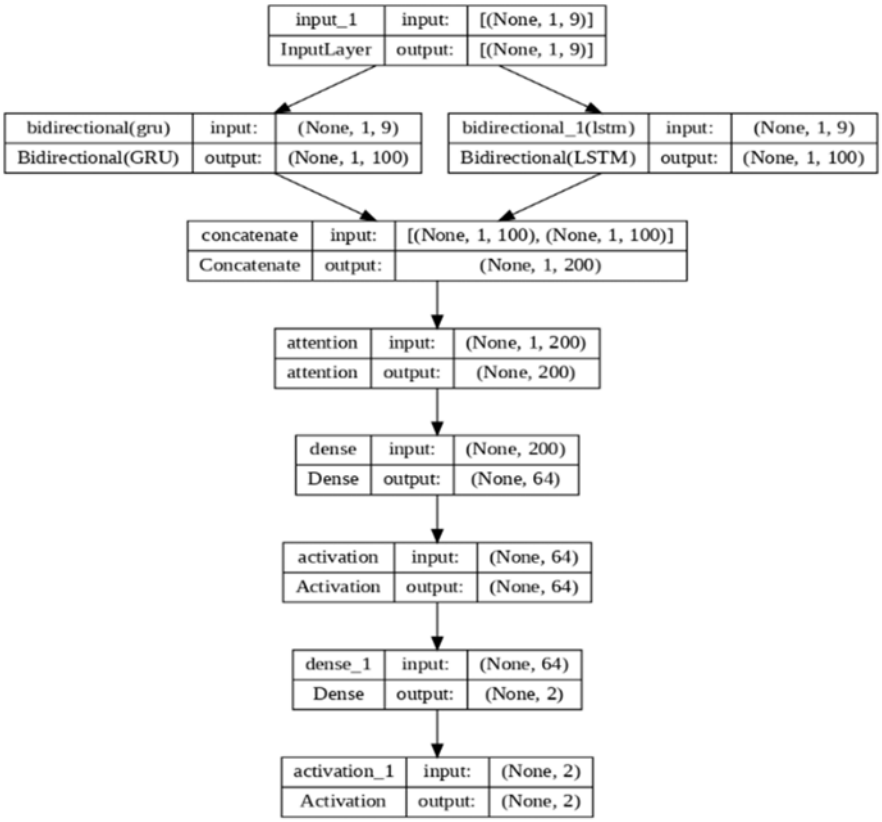


Fig. 3 The architecture of the proposed AttBiSeL network

1) Bidirectional LSTM Layer

LSTM is a variant of RNN designed to address the issue of vanishing or exploding gradients that traditional RNNs encounter. Similar to other varieties of RNNs, LSTM models produce their outputs by considering the input at the current time-step and the output from the previous time-step, subsequently passing the current output to the subsequent time-step. Furthermore, the LSTM network incorporates a memory element that allows it to retain important events that occurred in previous time-steps. In contrast to the conventional RNNs, which simply calculate the weighted sum of input signals and pass them through an activation function (as illustrated in Fig. 4), each LSTM unit comprises a memory cell and three gates. The memory cell (c_t) maintains its state over specified time intervals and the three non-linear gates include an input gate (i_t), a forget gate (f_t), and an output gate (o_t). These gates are designed to control the flow of information into and out of the memory cell [33]. The forget gate decides which information to discard and assigns a value in the range [0, 1] accordingly. The input gate determines which new information to store, and the output gate determines which portion of the cell state to output.

2) Bidirectional GRU Layer

The GRU network is another variant of RNNs, known for its improved computational efficiency and reduced computational cost when compared to the LSTM network. One of the core functionalities of RNNs is their capacity to consider long-term dependencies. In the GRU architecture, each unit is comprised of two critical gates. The first gate, the update gate (r), is

constructed by amalgamating the forget gate and the input gate. The second gate is the reset gate (z) [34]. The update and reset operations are executed using (4) and (5):

$$r_t = \delta(W_r h_{t-1} + U_r x_t + b_r) \quad (4)$$

$$z_t = \delta(W_z h_{t-1} + U_z x_t + b_z) \quad (5)$$

where δ stands for the logistic sigmoid function, while W and U signify weight matrices associated with gates or cells, representing input vectors (x_t) and hidden states (h_t), and b denotes the bias.

The reset gate plays a crucial role in deciding when to disregard the preceding hidden state, while the update gate is responsible for regulating the quantity of information that should be transmitted to the present state. The calculation of the hidden state is performed using (6) and (7):

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (6)$$

$$\tilde{h}_t = \tanh\left(W_{\tilde{h}_t}(h_{t-1} \odot r_t) + U_{\tilde{h}_t} x_t\right) \quad (7)$$

where ⊙ represents element-wise multiplication.

3) Attention Layer

By reviewing the literature on DL [25] and [35], it can be observed that the attention mechanism has proven to be an effective approach for achieving excellent results by selecting relevant information. The basic premise of this technique is to prioritize the most pertinent aspects of the information in lieu of the entirety of its content. Essentially, instead of encoding the entire input sequence into a fixed vector, this approach allows the model to determine how to generate the output vector at each step based on the current attentive parts and what it has learned so far.

In this paper, the attention mechanism is employed to assign different weights to pieces that contribute varying levels of information within a sample. One common approach to assigning varying weights to different pieces of information within a sample is the use of weighted combination (α_t) of all hidden states (see Fig. 6). The weighted combination has relationships expressed through (8) and (9):

$$\alpha_{tj} = \frac{\exp(v^T \cdot \tilde{h})}{\sum_t \exp(v \cdot \tilde{h})} \quad (8)$$

$$S_{A_w} = \sum_t \alpha_t h_t \quad (9)$$

where h and $\tilde{h}$ are determined by (6) and (7), and v serves as a trainable parameter.

4) FC Layer

The FC Layers are one of the most commonly used types of layers in DNNs. Their main function is to combine and transform inputs into a feature space for further processing in subsequent layers. The key feature of fully connected layers is that each neuron in the previous layer is connected to all neurons in the current layer. This way, feature information from inputs is fully propagated to the next layers. In the proposed architecture, a fully connected layer is employed before the final layer.

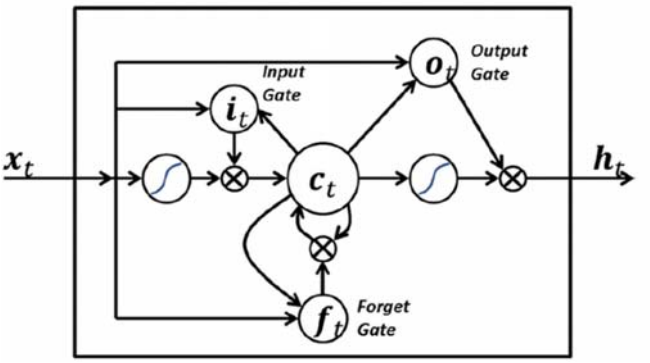


Fig. 4 Structure of an LSTM cell [33]

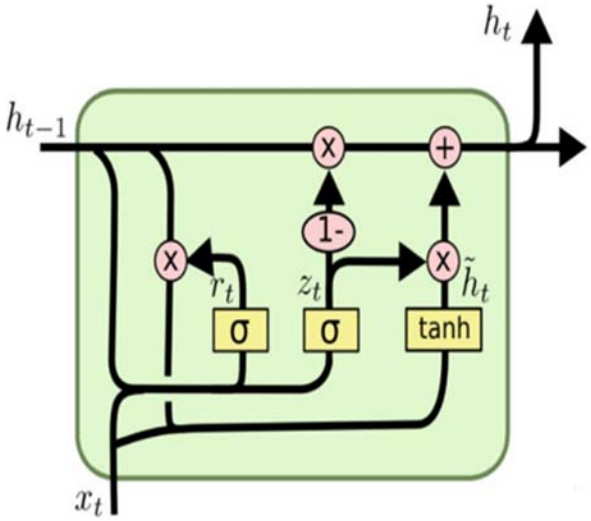


Fig. 5 Structure of a GRU cell [34]

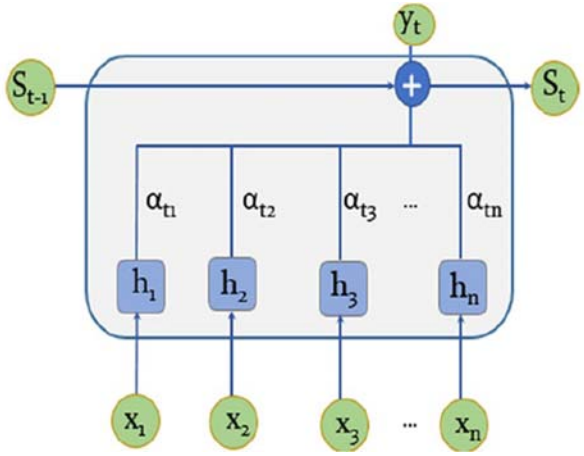


Fig. 6 The structure of attention mechanism [7]: x_n represents the input sequence vector, and S_t represents the hidden state

5) Dropout Layer

Dropout [36] is a regularization technique employed in deep networks to address the problem of overfitting in training data.

In this method, with a probability of p, certain neurons in the network are randomly dropped out, reducing the unnecessary information presented to the learning process. This action confines the network's reliance to specific and crucial features, preventing excessive overfitting. In this study, this layer is utilized to enhance the network's generalizability by focusing on essential and informative features. It can be argued that improved generalizability stems from the fact that an overly dominant neuron is not solely reliant on neighboring neurons' information. Consequently, the network is compelled to learn more robust and significant features from the inputs to accurately perform the task for which it was trained. This effectively encourages a deeper understanding of the fundamental and conceptual features inherent in the data. In this study, we employ Dropout after each sequence learner in order to mitigate the problem of high variance that can lead to overfitting.

6) Softmax Layer

In this study, the softmax function has been employed for classifying categories in the final layer of the proposed deep network. This function is a generalized version of the sigmoid function and is used to represent the probability distribution for n different classes with values ranging from 0 to 1. Therefore, for an input vector $z = [z1, z2, \ldots, zk]$ with k dimensions, it is defined as (10):

$$S(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, 1 \leq i \leq k \quad (10)$$

IV. Results

In this section, we will delve into the examination of the conducted experiments. The primary objective is the detection and classification of fraudulent activities in banking transactions using credit card data. Firstly, we elaborate the implementation details of the proposed method, the utilized database, and the evaluation criteria for conducting empirical tests. Subsequently, in the experiments section, the proposed approach is assessed, and the obtained results are presented.

A. Implementation Details

In this research, Python 3 programming language and the TensorFlow framework [37] were utilized for constructing, training, and testing the proposed architecture. Additionally, the Google Colab service was employed to fulfill the necessary hardware requirements. This service provides a Graphics Processing Unit (GPU) with a total memory of 15,109 MB, of which around 12,720 MB is accessible. This memory capacity satisfies the hardware requirements needed for conducting this study.

For all LSTM and GRU layers, the hyperbolic tangent (tanh) activation function was employed, while the softmax function was used for the classification layer. The cross-entropy loss function was used as the cost function for optimization. The Adam optimizer [38] was employed with a batch size of 25,000 for the training set and a batch size of 256 for validation in all experiments. The network's weights were initialized randomly and pre-trained models were not utilized for initialization.

B. Dataset

Datasets play a pivotal role in training and validating the proposed methodologies effectively within the research. In this study, we utilize the European Card Holders Fraud Data [9] dataset for training and evaluating the proposed model. This dataset was collected by European cardholder organizations and encompasses 284,807 transactions conducted over a span of two consecutive days, of which 492 cases were instances of fraud. One of the significant challenges posed by this dataset is its highly imbalanced nature, where the positive class (i.e., fraudulent transactions) accounts for a mere 0.172% of the entire dataset.

Another notable aspect of the dataset is that the Principal Component Analysis (PCA) transformation is applied to the data, resulting in only numerical input variables. Thus, the features v1, v2, ..., v28 are the primary PCA components. "Time" and "Amount" are the only features in this dataset that have not changed. The "Time" feature displays the elapsed time between each transaction and the first, while the "Amount" feature displays the monetary value of the transaction. In addition, the "Class" label with binary values "1" and "0" indicates whether a transaction is fraudulent or legitimate.

Fig. 7 depicts the correlation matrix of variables in the dataset, providing an interpretation of the pairwise correlation between each variable. This matrix illustrates that none of the principal components V1 to V28 are correlated with each other. However, upon closer inspection, it can be inferred that the response variable "Class" exhibits a certain level of positive and negative correlation with the principal components but lacks correlation with the "Time" and "Amount" variables.

C. Dataset Split

During the model selection process, the dataset is partitioned into three distinct subsets: the training set, validation set, and test set. Once the model is chosen, it undergoes training using the complete training dataset and is subsequently evaluated using unseen test data to make predictions. In this study, a hold-out method has been employed to ensure that the training and test sets have non-overlapping thematic coverage. Utilizing this approach, the majority of the data is allocated to the training set, while a portion is set aside for the test set. Specifically, we have dedicated 65% of the dataset for training, 15% for validation, and 20% for testing.

D. Data Balancing

One of the fundamental topics in machine learning is data balancing. In machine learning models, imbalanced learning occurs within classifiers. Imbalanced classification refers to a scenario where the number of observations in one class is significantly lower than in the other class [39]. In this research, a random oversampling method (i.e., SMOTE approach) has been utilized to balance the distribution of the dataset in each class by generating artificial instances for the minority class. This process continues until instances in the majority (legitimate) class and the minority (fraudulent) class are balanced. The main advantage of this approach is its potential

to enhance the accuracy of the proposed model. The distribution of each class in the dataset after applying the data balancing technique is depicted in Fig. 8.

E. Evaluation Metrics

To evaluate the model's performance, we use Accuracy, Precision, Recall, and F1-Score metrics. Specifically, the accuracy metric is used as the primary criterion for measuring model enhancement. A brief explanation of these metrics follows.

Accuracy is indicative of the ratio of correctly identified instances to the total number of instances and is calculated using (11):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

Precision signifies the ratio of correctly predicted transactions to the total predicted transactions and (12) demonstrates its calculation.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

Recall represents the ratio of correctly predicted fraudulent transactions to the total actual fraudulent transactions present in the dataset, and its calculation is demonstrated using (13):

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

F1-Score is the weighted average of precision and recall, computed according to (14):

$$F1-Score = 2 \times \frac{precision \times recall}{precision+recall} = \frac{2TP}{2TP+FP+FN} \tag{14}$$

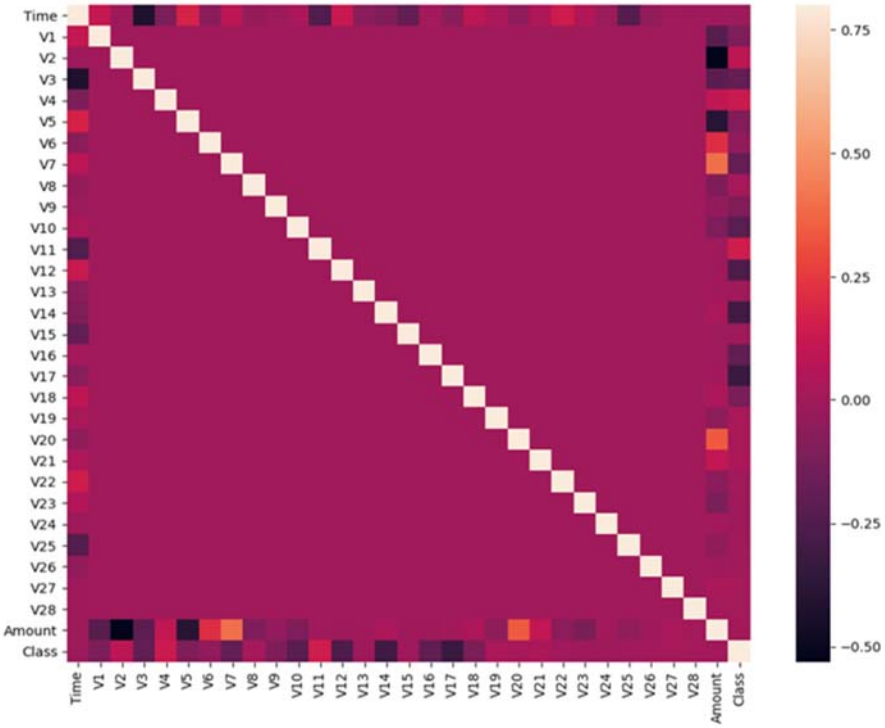


Fig. 7 The correlation matrix of the variables in the utilized dataset

F. Experiments

In this section, we delve into the empirical outcomes derived from the evaluation of the proposed methodology using the European Card Holders Fraud Data dataset. It is of utmost significance to underline that all the reported values presented herein are exclusively related to the test set.

Fig. 9 shows learning curves, which serve as insightful tools for understanding the performance and capabilities of a DL model. These curves effectively portray the trend of accuracy and cost throughout each training epoch, offering a clear picture of the model's behavior across both the training and validation sets.

Table I and Fig. 10 demonstrate a comprehensive depiction of the classification report and confusion matrix for the test set. As it can be seen from Table I, the incorporation of the attention mechanism has demonstrated a noteworthy enhancement in accuracy, corroborating its efficacy in focusing on relevant features while disregarding noise. This improvement signifies the model's ability to capture intricate patterns and relationships within the data, ultimately leading to more accurate predictions on the minority class. The intricate examination of the confusion matrix reveals the model's proficiency, capturing the majority of samples with remarkable accuracy. Furthermore, even in instances where misclassifications occurred, the proposed model exhibited its ability to accurately allocate samples to their respective categories.

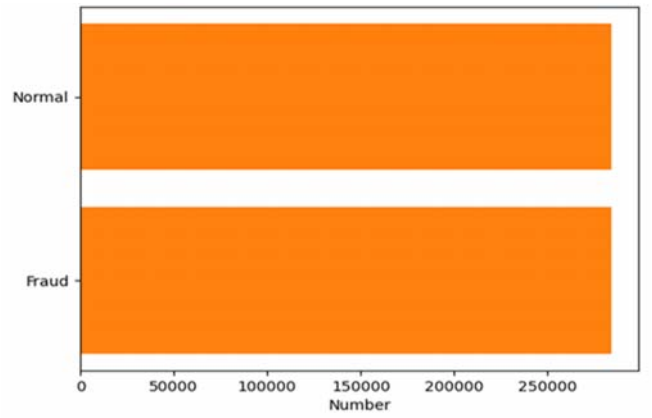


Fig. 8 The distribution of classes in the utilized database after applying the SMOTE balancing method

Table II brings together a diverse array of existing works, spanning from machine learning-oriented approaches to the more intricate domain of DL. The assessment entails a rigorous comparison of their respective accuracies against the proposed methodology. Empirical evidence unequivocally demonstrates that the proposed network excels in accuracy, substantially surpassing the performance of its predecessors. This improvement not only indicates the model's superior capability to distinguish between legitimate and fraudulent transactions but also reflects its potential for real-world deployment. This notable enhancement in accuracy is attributed to a twofold rationale: firstly, the incorporation of bidirectional layers within the LSTM and GRU recurrent networks, which effectively account for long dependencies within transactions, and secondly, the utilization of an attention mechanism, enabling the model to concentrate automatically on data elements that exhibit the most significant correlations with the classification task. This synergistic fusion of these two features culminates in a considerable elevation of predictive accuracy. A point of paramount importance pertains to the model's trainable parameters, which is remarkably lower than that of the assorted compared methods. This inherent feature positions the proposed network as an optimal and efficient choice, particularly in real-time applications where resource constraints and rapid processing are essential considerations.

To estimate the complexity of the model, it is often useful to determine the number of parameters that its architecture can accommodate. Table III provides a detailed breakdown of the parameters for each layer and their respective quantities within the proposed network. As evident from this table, the proposed network comprises a total of only 55,495 trainable parameters. This concise parameter count is indicative of the model's efficient design, striking a balance between complexity and performance. While other models might require considerably more parameters to achieve similar levels of accuracy, the proposed network demonstrates an impressive ability to capture the underlying patterns and relationships within the data while maintaining computational efficiency. The relatively low number of parameters is advantageous for various reasons. Firstly, it translates into faster training times, allowing the model to be trained on larger datasets or fine-tuned more quickly. Secondly, it enables the model to be deployed on resource-constrained environments without compromising performance. Finally, the reduced parameter count signifies the model's ability to generalize well, preventing overfitting and ensuring robust performance on unseen data.

TABLE I
THE CLASSIFICATION REPORT FOR THE TEST DATASET

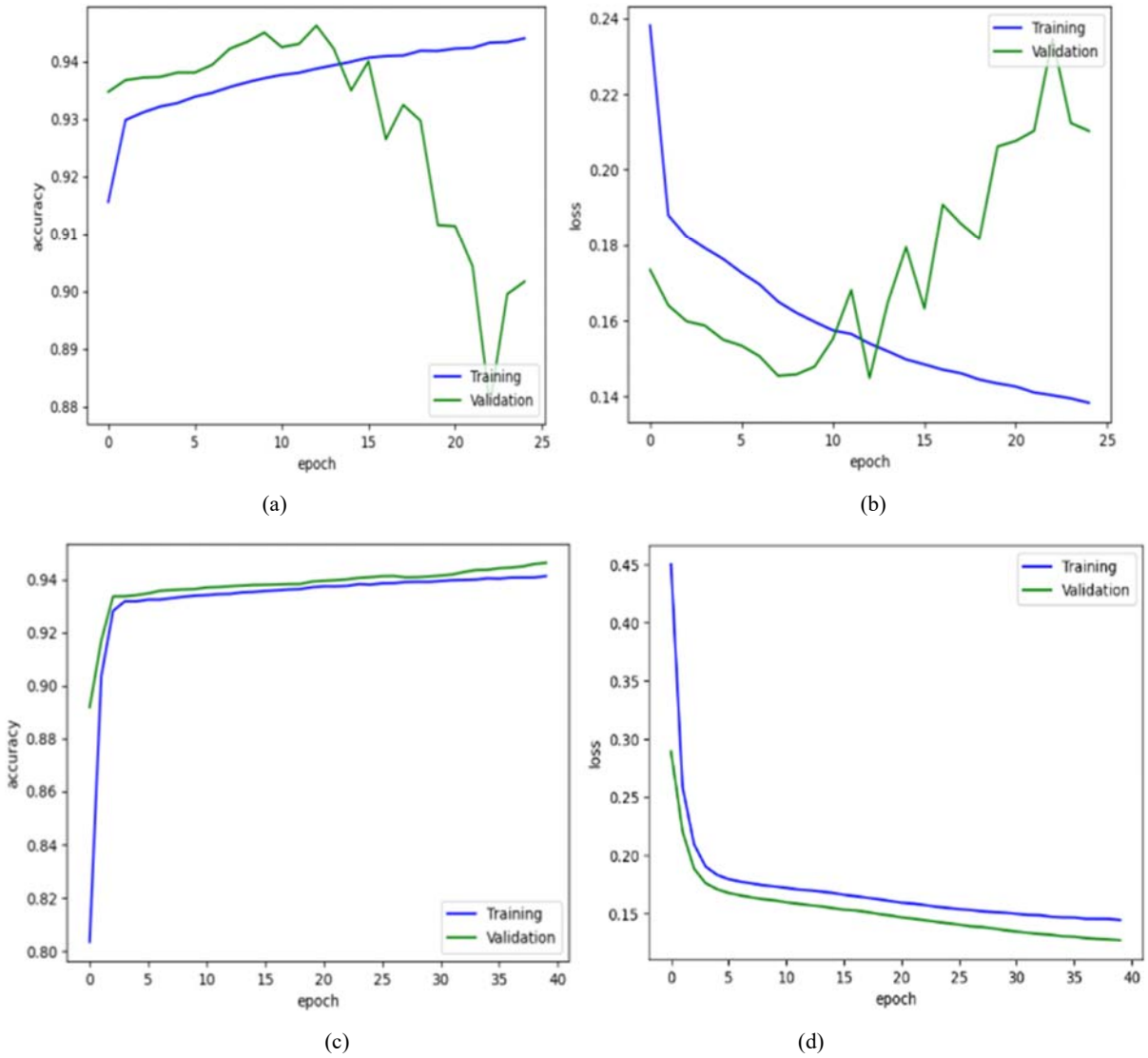
SMOTE	Attention	Class	Precision	Recall	F1-Score	Accuracy
×	×	Legitimate	99.96	99.98	99.97	87.49
		Fraudulent	84.62	75.00	79.52	
×	✓	Legitimate	99.97	99.98	99.98	90.47
		Fraudulent	85.00	80.95	82.93	
✓	×	Legitimate	90.17	98.46	94.13	93.82
		Fraudulent	98.28	89.18	93.51	
✓	✓	Legitimate	92.26	97.78	94.94	**94.76**
		Fraudulent	97.62	91.73	94.58	

TABLE II
COMPARISON OF RESULTS OF OUR PROPOSED FRAMEWORK WITH RECENT STUDIES

Method	Algorithm	Data Balancing	Attention Mechanism	Accuracy
[6]	GRU	✓		91.60
[11]	SVM	✓		93.49
	KNN			92.82
[14]	RF	✓		90.00
[15]	KNN	✓	×	94.40
[16]	AdaBoost	✓		97.00
[23]	CNN	×		93.50
[24]	LSTM	×		74.08
	GRU			72.08
Proposed Method	AttBiSeL	✓	✓	**94.76**

TABLE III
THE NUMBER OF PARAMETERS FOR EACH LAYER OF THE PROPOSED NETWORK

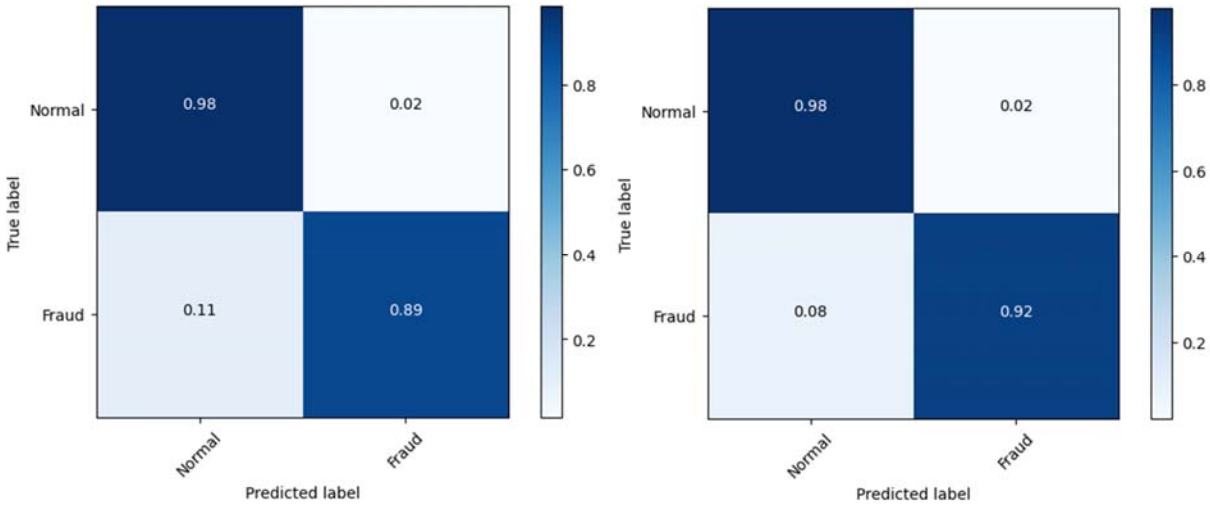
Layer Name	Output Shape	Parameters
Input Layer	[(None, 1, 9)]	0
Bidirectional GRU with Dropout	(None, 1, 100)	18300
Bidirectional LSTM with Dropout	(None, 1, 100)	24000
Concatenate	(None, 200)	0
Attention	(None, 200)	201
Dense	(None, 64)	12864
Dense	(None, 2)	130
Total parameters: 55,495		
Trainable parameters: 55,495		
Non-trainable parameters: 0		



(a) (b)

(c) (d)

Fig. 9 Learning curves of the proposed model: (a) and (b) accuracy and loss curves for training and validation sets without attention mechanism; (c) and (d) accuracy and loss curves for training and validation sets with attention mechanism



(a) (b)

Fig. 10 Normalized confusion matrix for the test set: (a) without the attention mechanism; (b) with the attention mechanism

V. Conclusion

In this study, we presented a framework for credit card fraud detection in order to enhance prediction performance when identifying fraudulent transactions. This achievement was attained through a combination of advantages from various methods. Specifically, the SMOTE oversampling technique was employed to address the class imbalance issue in the dataset, sequence learner networks (specifically, bidirectional LSTM and GRU networks) were utilized for the purpose of capturing and representing long-term dependencies within sequences of transactions. Additionally, an attention mechanism was employed to automatically prioritize and concentrate on the data elements that are most pertinent to the classification task. Therefore, the proposed approach can potentially identify meaningful consumption patterns that effectively contribute to distinguishing between fraudulent and legitimate transactions. For validating the performance of the proposed model, a benchmark dataset was utilized. Empirical results demonstrate its superior performance in classification of fraudulent transaction samples, which are crucial in this domain. Furthermore, the obtained results illustrate a significant improvement in accuracy compared to previous methods and highlight potential gains in terms of time and memory consumption. While this study establishes a strong foundation for combating credit card fraud, the landscape continues to evolve. The potential avenues for future research include the exploration of advanced attention mechanisms, real-time fraud detection, and the integration of additional data sources to further enhance the model's resilience and effectiveness.

References

[1] C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," *Computers in Human Behavior*, vol. 28, no. 3, pp. 1002–1013, Sep. 2010, doi: 10.1016/j.chb.2012.01.002.

[2] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, Jun. 2016, doi: 10.1016/J.JNCA.2016.04.007.

[3] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/J.EJOR.2017.11.054.

[4] Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, "Credit card fraud detection based on improved Variational Autoencoder Generative Adversarial Network," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3302339.

[5] J. Jurgovsky *et al.*, "Sequence classification for credit-card fraud detection," *Expert Systems with Applications*, vol. 100, pp. 234–245, Jun. 2018, doi: 10.1016/J.ESWA.2018.01.037.

[6] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," *2018 Systems and Information Engineering Design Symposium, SIEDS 2018*, pp. 129–134, Jun. 2018, doi: 10.1109/SIEDS.2018.8374722.

[7] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, Jun. 2017, doi: 10.48550/arxiv.1706.03762.

[8] H. M. Gomes, J. P. Barddal, A. F. Enembreck, and A. Bifet, "A Survey on Ensemble Learning for Data Stream Classification," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, Mar. 2017, doi: 10.1145/3054925.

[9] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pp. 159–166, 2015, doi: 10.1109/SSCI.2015.33.

[10] C. Sweetlin Hemalatha, V. Vaidehi, and R. Lakshmi, "Minimal infrequent pattern based approach for mining outliers in data streams," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1998–2012, Mar. 2015, doi: 10.1016/J.ESWA.2014.09.053.

[11] A. RB and S. K. KR, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/J.GLTP.2021.01.006.

[12] A. Srivastava, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection using Hidden Markov Model," *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 1, pp. 37–48, 2008, doi: 10.1109/TDSC.2007.70228.

[13] A. Dal Pozzolo, R. Johnson, O. Caelen, S. Waterschoot, N. V. Chawla, and G. Bontempi, "Using HDDT to avoid instances propagation in unbalanced and evolving data streams," *Proceedings of the International Joint Conference on Neural Networks*, pp. 588–594, Sep. 2014, doi: 10.1109/IJCNN.2014.6889638.

[14] M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," *2019 Proceedings of the 3rd International Conference on Computing and Communications Technologies, ICCCT 2019*, pp. 149–153, Feb. 2019, doi: 10.1109/ICCCT2.2019.8824930.

[15] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit Card Fraud Detection using Pipeling and Ensemble Learning," *Procedia Computer Science*, vol. 173, pp. 104–112, Jan. 2020, doi: 10.1016/J.PROCS.2020.06.014.

[16] Y. F. Zhang, H. L. Lu, H. F. Lin, X. C. Qiao, and H. Zheng, "The Optimized Anomaly Detection Models Based on an Approach of Dealing with Imbalanced Dataset for Credit Card Fraud Detection," *Mobile Information Systems*, vol. 2022, no. 1, p. 8027903, Jan. 2022, doi: 10.1155/2022/8027903.

[17] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," *ICNSC 2018 - 15th IEEE International Conference on Networking, Sensing and Control*, pp. 1–6, May 2018, doi: 10.1109/ICNSC.2018.8361343.

[18] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[19] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3637–3647, Oct. 2018, doi: 10.1109/JIOT.2018.2816007.

[20] I. Benchaji, S. Douzi, and B. El Ouahidi, "Novel learning strategy based on genetic programming for credit card fraud detection in big data," *Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019*, pp. 3–10, 2019, doi: 10.33965/BIGDACI2019_201907L001.

[21] N. Mahmoudi and E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2510–2516, Apr. 2015, doi: 10.1016/J.ESWA.2014.10.037.

[22] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, Aug. 2018, doi: 10.1109/TNNLS.2017.2736643.

[23] M. Z. Mizher and A. B. Nassif, "Deep CNN approach for Unbalanced Credit Card Fraud Detection Data," *2023 Advances in Science and Engineering Technology International Conferences, ASET 2023*, 2023, doi: 10.1109/ASET56582.2023.10180615.

[24] J. Forough and S. Momtazi, "Ensemble of deep sequential models for credit card fraud detection," *Applied Soft Computing*, vol. 99, p. 106883, Feb. 2021, doi: 10.1016/J.ASOC.2020.106883.

[25] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, Accessed: May 21, 2023. Online. Available: https://arxiv.org/abs/1409.0473v7

[26] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2048–2057, Feb. 2015, Accessed: May 21, 2023. Online. Available: https://arxiv.org/abs/1502.03044v3

[27] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," *2016 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, pp. 1480–1489, 2016, doi: 10.18653/V1/N16-1174.

[28] B. Lebichot, Y.-A. Le Borgne, L. He-Guelton, F. Oblé, and G. Bontempi, "Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection," in *INNS Big Data and Deep Learning conference*, 2020, pp. 78–88. doi: 10.1007/978-3-030-16841-4_8.

[29] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks," pp. 483–490, 2016, doi: 10.1007/978-3-319-46675-0_53.

[30] A. Somasundaram and S. Reddy, "Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance," *Neural Computing and Applications*, vol. 31, no. 1, pp. 3–14, Jan. 2019, doi: 10.1007/S00521-018-3633-8/METRICS.

[31] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of Microarray Data Using Z Score Transformation," *The Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, May 2003, doi: 10.1016/S1525-1578(10)60455-2.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[33] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/J.NEUCOM.2019.01.078.

[34] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, "Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 34–44, Aug. 2016, doi: 10.1109/MCI.2016.2572539.

[35] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[36] H. G. K. A. S. I. S. R. S Nitish, "Dropout: a simple way to prevent neural networks from overfitting," *J Mach Learn Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arxiv.org*, Online. Available: https://arxiv.org/abs/1603.04467

[38] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, Accessed: Jan. 09, 2022. Online. Available: https://arxiv.org/abs/1412.6980v9

[39] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowledge-based systems*, vol. 212, p. 106631, 2021.