IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 9, 2022, accepted March 24, 2022, date of publication March 30, 2022, date of current version April 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3163246

CBiLSTM: A Hybrid Deep Learning Model for Efficient Reputation Assessment of Cloud Services

REEM AL SALEH[1], MAHA DRISS[2,3], (Senior Member, IEEE), AND IMAN ALMOMANI[2,4], (Senior Member, IEEE)

[1]IS Department, College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia
[2]Security Engineering Laboratory, CCIS, Prince Sultan University, Riyadh 12435, Saudi Arabia
[3]RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba 2010, Tunisia
[4]CS Department, King Abdullah II School of Information Technology, The University of Jordan, Amman 11942, Jordan

Corresponding author: Maha Driss (maha.idriss@riadi.rnu.tn)

ABSTRACT The cloud market is characterized by fierce rivalry among cloud service providers. The availability of various services with identical functionalities on the market complicates the selection decision for service requesters. Although objective trust measurements can be used to evaluate the trustworthiness of services, they are not always available and are static in nature. Subjective approaches are not always viable because they often require repeated service invocations to collect client feedback. To overcome these limitations, we propose, in this paper, a reputation-based trust assessment approach that combines the Net Brand Reputation (NBR) measure with a deep learning-based sentiment analysis model using online user reviews. CBiLSTM is the name of the proposed deep learning model that hybridizes the Convolutional Neural Networks (CNN) and the Bidirectional Long Short-Term Memory (BiLSTM) layers. The CNN layers deal with text inputs' high dimensionality, while the BiLSTM layer explores the context of the extracted features in both forward and backward directions. CBiLSTM was trained on a new dataset named CLOSER-DREAM, containing more than 13,000 reviews relating to several emerging cloud services to classify these reviews and assess the overall reputation of the cloud services providers. The results of the series of experiments that were conducted have shown that CBiLSTM outperforms the classic deep learning models with 98% of precision, 99% of recall, 98% as an F1-score, and 99.7% of accuracy. Also, CBiLSTM offered a reasonable training time of about 519ms with the CLOSER-DREAM dataset. The classification obtained by applying CBiLSTM was proven to be an effective method to calculate the NBR measure used for the reputation assessment of cloud service providers. The proposed technique yielded an NBR score of 98.3% for Google cloud services, which is close to the real/actual NBR of 96.25%.

INDEX TERMS Cloud services, service selection, reputation-based trust assessment, sentiment analysis, deep learning, CBiLSTM.

I. INTRODUCTION

Cloud computing is a robust model that enables the delivery of on-demand computing resources over the internet on a pay-as-you-go basis. This technology has been on a rapid upward trajectory in recent years. According to a recent Gartner[1] report, global public cloud spending would approach 45% of total company IT spending by 2026, up from less than 17% in 2021 [1]. The growing cloud services market is also characterized by fierce competition between service providers [2]. Each service provider claims to offer services that best satisfy user requirements regardless of the latency, privacy, security, or trust-related issues the service may have [3]–[5]. Although the intense competition is a sign of a healthy market, it complicates the selection decision for service requesters.

To make informed decisions, potential users need to assess the trust of the candidate cloud service providers. In this context, trust refers to the provider's confidence level that reflects the provider's capabilities, reliability, and honesty [5], [6]. To assess the trust, the users may refer to the published formal quality measures specified in the Service-Level Agreement (SLA) established between the service provider, the service user, and the standard audit reports provided by third parties [7]. These objective trust methods evaluate service conformance with promised Quality-of-Service (QoS)

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

[1] https://www.gartner.com/en

attributes like response time, availability, security, robustness, and scalability [8]–[10]. However, the values of these attributes are not always accessible and have a static nature, which may fail to reflect fluctuations in the service performance. Service clients may utilize the subjective approaches that employ user feedback and ratings to gradually assess the reputation-based trust of services [5], [11]. Such methods often rely on specific acquisition mechanisms to collect feedback on QoS attributes. However, the feasibility of the acquisition mechanism might be a concern.

In most cases, these subjective approaches depend on users' repeated invocations of services and the users' willingness to provide feedback on the invoked services. Thus, these approaches are challenged by data sparsity and cold start problems. Moreover, these approaches generally neglect qualitative factors that affect subjective trusts, such as aesthetics, affordability, and usability [5], [12].

On the other hand, there is a plethora of user feedback on various cloud services available on the Internet. These reviews represent a useful source of information that may be utilized to solve the issues outlined above. Recently, several research studies have employed sentiment analysis techniques to automatically transform unstructured customers' reviews into structured data, which can be particularly useful for service reputation management. In this context, sentiment analysis is used as a procedure for assessing customers' satisfaction toward invoked cloud services by classifying their related reviews. Existing approaches that employed sentiment analysis for reputation assessment can be categorized into three main classes: 1) statistics-based, 2) fuzzy-logic-based, and 3) traditional data mining-based approaches. Even though these approaches provide efficient methods for estimating services' reputation, they present several limitations, including:

- These approaches are not scalable since they don't enable analyzing newly added reviews.
- They are time-consuming and require high computing resources to analyze a large number of customers' reviews.
- They are domain-specific, and the obtained results are highly tied to the data context and features used in the experimentations. They need reengineering adjustments to ensure the reputation assessment of entities other than those considered in the experiments.
- They do not provide a concrete score that helps to assess the overall reputation.
- They are not validated through different performance metrics.
- They are often dependent on specific QoS information and constrained by their associated acquisition and analysis processes.
- They do not consider subjective, trust-based qualitative factors affecting the overall services' reputation score.

This study aims to answer the following question: What techniques may be employed to overcome the staticity, infeasibility, and inefficiency challenges associated with most current trust assessment approaches?

To address the limitations of the existing solutions for assessing the reputation of the next generation of IT services, we propose a novel approach that employs deep learning-based sentiment analysis and Net Brand Reputation (NBR) techniques. This approach introduces a novel hybrid deep learning model that hybridizes the Convolutional Neural Networks (CNN) and the Bidirectional Long Short-Term Memory (BiLSTM) layers. It is named Convolutional BiLSTM Deep Learning Model (CBiLSTM), and it allows effective review classification. This work also provides a new dataset of cloud service reviews. The collected dataset serves as input for the CBiLSTM classifier, which then feeds into NBR to compute the overall reputation score. The following points are a summary of the key contributions of this study:

- The proposed approach tackles data scarcity issues and cold start by exploiting the knowledge provided by customers' feedbacks available on the Internet. In addition, when processing and classifying this data, our approach considers qualitative criteria that substantially impact subjective trusts, such as aesthetics, affordability, and usability.
- A new dataset, named ClOud SErvices Reviews Dataset for REputation AssessMent (CLOSER-DREAM), is collected, cleaned, and labeled. This dataset contains more than 13,000 textual reviews related to various cloud services.
- An efficient and novel hybrid deep learning model, CBiLSTM, is proposed for reviews classification. In this model, CNN layers are used to deal with the high dimensionality of text inputs, in contrast, BiLSTM layers are used to investigate the context of the retrieved features in both forward and backward directions. According to the experiments' results, CBiLSTM outperforms the classical deep learning models used for sentiment classification. It provides 98% precision, 99% recall, 98% F1-score, and 99.7% accuracy on the CLOSER-DREAM dataset. Furthermore, the extensive experiments that were carried out have demonstrated that CBiLSTM requires less time for training than the other deep learning models considered in the comparative study.
- This work proposes a procedure for applying the NBR formula based on CBiLSTM outputs. It also generates a concrete NBR score to assess the overall services' reputation.
- To validate the proposed approach, Google cloud services' reputation is assessed based on CBiLSTM classification and compared to the real reputation score value. The obtained results show that CBiLSTM is effective for assessing the reputation of service providers.

This paper is organized as follows. Section II introduces the background on the main theoretical concepts of this study and discusses the related work. Section III presents the proposed approach. Section IV describes the experiments that were carried out and discusses the achieved outcomes. Section V summarizes the main contributions and findings and future research directions that will be investigated to extend this study.

II. BACKGROUND AND RELATED WORK

A. BACKGROUND

1) CLOUD COMPUTING

According to the National Institute of Standards and Technology (NIST) [13] ''cloud computing is a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction''.

Cloud computing offers services in three primary forms [14]: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). IaaS virtualizes the data centers' processing power, storage, and network access. PaaS provides a development platform with a range of services supporting the design, development, testing, deployment, monitoring, and hosting of applications on the cloud. SaaS presents software to the end-users as on-demand services, usually using a browser. Google Apps,[2] Microsoft Azure,[3] and Amazon Elastic Compute Cloud (EC2)[4] are some examples of SaaS, PaaS, and IaaS, respectively.

Cloud computing brings many benefits such as ease of use, cost-efficiency, flexibility, elasticity, on-demand scalability, and economies of scale [4], [5]. However, cloud services raise numerous concerns about latency, privacy, security, and trust [3]–[5]. These key aspects pose a challenge for cloud services that must be addressed to build trust among cloud stakeholders, including cloud service users, cloud service providers, and third parties [5].

2) DEEP LEARNING

Deep learning [15] is a representation learning approach that uses artificial neural networks to progressively learn representations like patterns and relationships from a large amount of raw data. Each neural network is a series of biologically inspired algorithms that transform the model at one level into a higher and more abstract level of representation through dynamic adjustment of weight values [15], [16]. The learned representations are eventually used for detection or classification tasks [15].

A neural network can be visualized as a set of connected artificial nodes or neurons. Each neuron receives input values or patterns from other neurons, performs some processing operations, and then produces outputs [16]. Neurons in deep neural networks are generally organized into three types of layers: input layer, hidden layers, and output layer. These layers are connected to allow communication between neurons [16].

Recently, deep learning has received a lot of traction in a variety of sectors and applications [17]–[19]. For sentiment classification, specific deep learning models are often used. These include CNN, Recurrent Neural Network (RNN) extensions like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and BiLSTM [20].

B. RELATED WORK

1) OVERVIEW OF RELEVANT STUDIES

Several studies in the literature have addressed the trust and reputation assessment of cloud services. They have adopted different classification models, techniques, quality features, and data sources.

Fan *et al.* [21] proposed a multi-dimensional trust-based mechanism for selecting cloud services using Evidential Reasoning techniques. It combines perception-based trust value that an active user acquired from direct service interactions and reputation-based trust value derived from other users' interactions.

Noor *et al.* [22] described the design and implementation of a management framework for the reputation-based trust called CloudArmor. This framework provides a set of functionalities and features to offer Trust-as-a-Service (TaaS): i) anonymization techniques are utilized to guard users against privacy breaches, ii) several metrics for detecting the feedback collusion and Sybil attacks are suggested, and iii) load balancing techniques are adopted to distribute the workload and maintain a desired level of availability through the deployment of the Trust Management Services (TMS).

Ding *et al.* proposed in [23] a ranking prediction model for a personalized selection of cloud service that considers the expectation and attitude of customers towards the quality of service. To enhance the service ranking prediction's accuracy, the proposed technique employed an enhanced Kendall Rank Correlation Coefficient (KRCC) measure that integrates Jaccard's coefficient to reduce the effect of negative customers' reviews in calculating ranking similarity. It also used a customer satisfaction function named Cloud Service Ranking Prediction (DSRP) to find the preference values on pairs of services.

Mao *et al.* in [16] employed Particle Swarm Optimization (PSO) to find the optimal initial connection weights between layers in networks to reduce the impact of initial parameter settings and ultimately achieve more accurate trust prediction of cloud services. The proposed neural network based on PSO techniques demonstrated its efficiency over both basic classification methods (e.g., bayesian network and decision tree) and traditional Back-Propagation Neural Networks (BPNN).

Somu *et al.* in [24] proposed a multi-level Hypergraph Coarsening-based Robust Heteroscedastic Probabilistic Neural Network (HC-RHRPNN) for cloud service

[2] https://workspace.google.com/

[3] https://azure.microsoft.com/en-us/

[4] https://aws.amazon.com/ec2/

trustworthiness prediction. The proposed model utilizes pruning, a dimensionality reduction technique, to identify good-conditioned samples to be then used for HRPNN training. The proposed model made improvements concerning prediction accuracy and execution time.

In [4], Deshpande *et al.* proposed an Evidence-Based Trust Estimation Model (EBTEM) for cloud services' adaptive and dynamic trust assessment. EBTEM employed evidence factors of various QoS attributes of cloud services derived from the direct interaction between cloud users and the services. The computed cumulative trust value, on which the user's decision to use or not use the service is based, represents the core of the proposed dynamic trust prediction approach.

Liu *et al.* in [25] proposed a method that combines clustering-based techniques and trust-based Collaborative Filtering (CF) approach to improve the prediction accuracy and recommendation quality. The clustering-based technique incorporated explicit textual information, rating information, and implicit context information to identify similar users and provide personalized services. The trust-aware CF approach merges local and global trust values to address user unreliability issues.

Rizvi *et al.* [5] suggested a Fuzzy Inference System (FIS) that returns a quantitative security index for Cloud Service Users (CSUs). For coherent analysis of security, the system addressed the multiple possibilities or uncertainties that CSUs may have when assessing the reliability of a cloud service provider. The overall security assessment depended on CSUs' evaluation of four main factors: compliance, access controls, auditability, and encryption.

Li *et al.* in [26] proposed a framework, named FASTCloud, that facilitated the selection of trustworthy cloud services by Potential Cloud service Consumers (PCCs) and enhanced the feasibility of the acquisition of QoS information. The model collects information related to QoS attributes of different cloud services (i.e., Service Level Objectives (SLOs) and Actual Monitoring Values (AMVs)) from Cloud Service Providers (CSPs) and Cloud Service Consumers (CSCs), respectively. The trustworthy cloud services selection component of FASTCloud receives the information and evaluates the cloud service trust level. Also, the deviation maximization-based weight assignment method is utilized for an objective determination of QoS attributes' weights.

2) RELATED STUDIES COMPARISON

Table 1 summarizes the relevant studies reviewed in the previous subsection. The first column presents the previously mentioned related works. The "Classification" column provides a high-level classification of the reviewed reputation and trust approaches; this classification is inspired by the work of Wahab *et al.* [11]. The "Techniques" column lists the specific techniques employed by each work. The "Assessed Quality Features" column specifies the attributes that are considered in the assessment process. The "Outcomes" column highlights the added values provided by each work. Finally, the "Limitations" column points to the studies' drawbacks or constraints.

3) DISCUSSION

Based on Table 1, it is noticeable that many studies rely on users' feedback for reputation and trust assessment. However, it is not always practical to request many users to rate services against fine-grained criteria for an overall view of community opinion as users can be reluctant or unmotivated to spend time evaluating services. As a result, the data sparsity problem becomes problematic when relying on such an approach. This explains why some studies assessed using prediction methods. Furthermore, some approaches rely on user history information and presume that QoS status monitoring tools/services are available to service customers. Thus, the feasibility of QoS information acquisition mechanisms might be a concern. Considering the performance and security of service platforms, the monitoring techniques used for QoS information acquisition apply only to individual service users. Monitoring platforms by multiple clients using the same or different services is not supported. Furthermore, the actively gathered QoS information of service providers from open sources can be incomplete and inaccurate due to inconsistent updates made by service providers [5], [26].

As shown by Table 1, most studies investigate the quantitative factors, i.e., QoS attributes of service trust assessment, such as performance, availability, and response time. The proposed approaches have the advantage of linking subjective user feedback to specific QoS attributes or providing objective trust assessment. However, they overlook the qualitative factors that considerably affect subjective trusts, such as aesthetics, affordability, and usability [5], [12].

Furthermore, the environment of the new generation services has a dynamic nature, where new services with unpredictable QoS attributes emerge continuously. This can make the service selection problem more complex. Prediction-based trust assessment approaches can be plausible in solving this issue, especially when the trustworthiness of a newly emerging service needs assessment with minimal knowledge of the QoS characteristics of the service [24].

Because of its self-learning capabilities in modeling complicated and arbitrary relationships, the deep learning-based approaches have outperformed traditional methods in trust prediction [24], [27].

This work utilizes the deep learning-based approach for service reputation-based trust assessment. Instead of soliciting users' comments at every service invocation, the proposed approach takes advantage of the rich information resources accessible online in the form of free-text user reviews to address the issues previously highlighted. It is worth mentioning that having a vast number of reviews assessed can ensure that various use cases are tested and reduce the effect of unauthentic or misleading reviews.

This work aims to improve the existing research in this area. Our primary purpose is to develop a comprehensive, trustworthy, and novel approach that employs deep

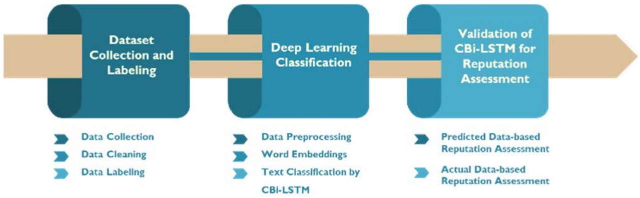
Dataset Collection and Labeling
Deep Learning Classification
Validation of CBi-LSTM for Reputation Assessment
Data Collection
Data Cleaning
Data Labeling
Data Preprocessing
Word Embeddings
Text Classification by CBi-LSTM
Predicted Data-based Reputation Assessment
Actual Data-based Reputation Assessment

FIGURE 1. The proposed approach's pipeline.

learning-based sentiment analysis and NBR techniques to ensure reputation assessment for cloud services. This work introduces a hybrid deep learning model named CBiLSTM for reliable review classification. It also offers a new dataset of cloud service reviews. The collected dataset serves as input for the classifier, which in turn generates input information for NBR. The NBR score reflects the Quality of Experience (QoE), i.e., the overall user acceptance of service based on subjective perception [9], [10], [28]. The proposed approach has the following advantages:

- It is feasible as it does not require direct user intervention to get user feedback on services.
- It is dynamic as CBiLSTM is capable of classifying any newly added service reviews.
- It is time-saving because CBiLSTM uses existing web reviews to accomplish classification in a short period of time.
- It is effective for reputation assessment as it generates NBR scores closer to the actual/real scores.
- It deals with more authentic and rational feedback as it employs a large number of published reviews.
- It delivers a comprehensive reputation assessment that considers all subjective trust-based factors.
- It is generalizable since it may be used to measure the reputation of various entities other than cloud services.

III. METHODOLOGY AND PROPOSED APPROACH

This work aims to adopt deep learning-based sentiment analysis for effective and efficient reputation assessment of cloud service providers. Considering that, this paper follows a pipeline of three main phases:

1. Dataset collection and labeling phase.
2. Deep learning classification phase.
3. Deep learning model validation phase.

These phases are illustrated in Figure 1 and detailed in the following paragraphs.

A. DATA COLLECTION AND LABELLING PHASE

This work introduces a new dataset consisting of English reviews on a range of cloud services shown in Table 2. The dataset is named CLOSER-DREAM, ClOud SErvices Reviews Dataset for REputation AssessMent. It contains 13,178 reviews including 12,567 (95.37%) "Positive" reviews, 260 (1.97%) "Negative" reviews, and 350 (2.66%) "Neutral" reviews. Figure 2 shows the unbalanced distribution of the dataset based on the number of instances per class.

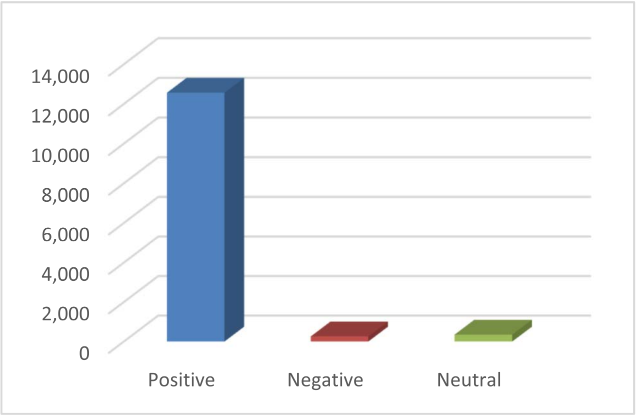

FIGURE 2. CLOSER-DREAM distribution.

In addition, the reviews in CLOSER-DREAM are relatively long, with a maximum review length of 818 words, a minimum review length of 3 words, and a median review length of 64 words. The number of unique words is estimated at 15,218 words. Most of the gathered reviews discuss the benefits and drawbacks of the reviewed services.

The reviews in CLOSER-DREAM are scraped from multiple review websites. They are also cleaned by removing duplicates and noises. Manual labeling of this dataset is impractical due to the large number of collected reviews that are handled. Furthermore, the widely available ratings on review websites are inconsistent, as clients with similar concerns may rate the same service differently. As a result of these limitations, we suggest a two-stage labeling technique. First, a sentiment analysis tool is used to label the dataset automatically. Second, minority classes reviews are manually checked and re-labeled depending on specific features.

B. DEEP LEARNING CLASSIFICATION PHASE

This phase employs deep learning to classify texts in terms of sentiment. This section discusses the preparation activities required to convert the textual reviews into a machine-understandable format. In addition, it introduces a novel hybrid deep learning model for sentiment classification.

1) DATA PREPROCESSING

When applying Deep Learning to text, several preprocessing steps are required to convert texts into appropriate formats and sizes:

1. Data is filtered to uniform text and remove unneeded characters.
2. Sentences are tokenized, i.e., divided into smaller units such as words.
3. A word-to-index dictionary can be created by mapping each vocabulary to a unique integer value based on word frequency. Text sentences are then converted to sequences of integers where each number matches up to the corresponding words in the index by following a typical process called sequencing.

TABLE 1. Comparison table of relevant studies related to the trust and reputation assessment of cloud services.

Criteria Work	Classification	Techniques	Assessed Quality Features	Inputs for Trust Assessment	Outcomes	Limitations
[Fan et al., 2015]	Evidence and statistics-based approach	Evidential reasoning	Security and privacy, operational performance, and other QoS attributes	Multi-dimensional trust ratings (trust value from active users based on direct service interactions + trust value based on other users' interactions)	• Real-time trust value of aggregated multi-dimensional trust ratings • Consistently updated trust evidence	• Dependent on subjectively given weights • Vulnerable to data sparsity and "cold start" problems • Feedback authenticity is critical if data is insufficient • Does not consider subjective trust-based qualitative factors
[Noor et al., 2016]	Feedback and statistics-based approach	Several mathematical equations and algorithms	QoS attributes	Users' feedback through distributed TMS interfaces	• A set of functionalities provided as TaaS	• Vulnerable to data sparsity and "cold start" problems • Feedback authenticity is critical if data is insufficient • Does not consider subjective trust-based qualitative factors
[Ding et al., 2017]	Feedback and statistics-based	KRCC, Jaccard Coefficient, and CSRP	High utility QoS attributes	Users' feedback from a third-party platform	• Predicted ranking of cloud services matching customers' preferences	• Inefficient at ranking enormous numbers of cloud services • No method for computing the overall trust score • CSRP requires historical records of service candidates • Does not consider subjective trust-based qualitative factors
[Mao et al., 2017]	Data mining-based (Neural networks-based)	PSO-driven neural networks	QoS attributes	Public QoS dataset (QWS dataset) + Population size of swarm + Basic parameters of PSO and BPNN	• Optimized parameters settings for the neural network for accurate and effective trust prediction of cloud services	• Feasibility concerns of QoS information acquisition • Uses a Web services dataset for performance evaluation • Requires tuning of PSO parameters • Inefficient as it needs multiple times training of many basic neural networks • Does not consider subjective trust-based qualitative factors
[Somu et al., 2018]	Data mining-based (Neural networks-based)	Multi-level HC-RHRPNN approach	QoS attributes	Public QoS dataset (QWS dataset)	• Improved trustworthiness prediction accuracy and minimized runtime • Proven effectiveness of pruning on classifier performance	• Feasibility concerns of QoS information acquisition mechanisms • Uses (Web services) dataset for performance evaluation • lower accuracy than that of state-of-the-art techniques • Does not consider subjective trust-based qualitative factors
[Deshpande et al., 2018]	Evidence and statistics-based	EBTEM	QoS attributes	Inputs from the interactions between users and the considered service (Set of cloud service attributes + Number of time instances + Threshold (minimum expected) trust + Cumulative historical trust value + Number of positive and negative ratings)	• Dynamic trust prediction (present and cumulative trust values) based on service evidence factors	• Requires a long and difficult process to acquire valid and credible QoS evidence • Assumes existence of valid services/tools for service extraction, monitoring, and related functionalities • Vulnerable to data sparsity and "cold start" problems • Does not consider subjective trust-based qualitative factors
[Liu et al., 2019]	Data mining-based (Cluster-based and collaborative filtering)	K- medoids	QoS attributes	Users explicit textual feedback and rating + Implicit context information	• Personalized QoS value prediction and reliable recommendation (using local and global trust values) of top cloud services	• Less effective performance under the condition of multi-tasks in large-scale data sets • Does not consider subjective trust-based qualitative factors
[Rizvi et al., 2020]	Feedback and fuzzy-logic-based	Fuzzy inference system offering fuzzification, Inference engine, and defuzzification	Security attributes	Numerical (crisp) input or fuzzy expressions provided by CSUs + Inference rules	• Quantitative security index based on CSUs' evaluation of specific top-level factors with consideration of the emotional aspects related to the computation of the subjective trust	• Feasibility concerns of security information and acquisition mechanisms • Evaluates a limited number of security sub-factors • Puts the same emphasis (weight) on all sub-factors • Requires a regular update of fuzzy logic rules • Less effective than machine learning and neural network approaches
[Li et al., 2020]	Feedback and statistics-based	Several mathematical equations and algorithms	QoS	Information related to QoS attributes (i.e., SLOs and AMVs) of cloud services collected from CSPs and CSCs	• Ranked list of CSPs with matching QoS requirements to PCCs' requests	• Assumes CSPs provide QoS status monitoring tools/services for CSCs • Requires frequent submission of SLOs by CSPs and frequent submission of AMVs by CSCs • Does not consider subjective trust-based qualitative factors

TABLE 2. Services reviewed in CLOSER-DREAM.

Google Cloud Services			
Google Storage	Google Cloud Platform	Google Compute Engine	Google App Engine
Google Cloud Functions	Kubernetes Engine	Google Cloud AI platform	Google Drive
BigTable	Google Fire Store	Google Cloud SQL	Google Cloud Scanner
Google Cloud ML Engine	IoT Google Cloud	Google BigQuery	Google Cloud Console
Google Cloud Operations	IAM		
AWS Services			
Amazon Lambda	Amazon RDS	Amazon Chime	Amazon Elastic Compute Cloud EC
AWS Managed Services	Amazon Auto Scaling	Amazon Services	
Microsoft Azure Services			
Microsoft Azure	Azure Storage	Azure DevOps	Azure DevOps Server
Azure Backup	Azure Machine Learning		
Services Provided by Other Providers			
IBM Cloud Bare Metal Servers	IBM Cloud Database	Salesforce Platform	Heroku Platform
Oracle MySQL Cloud Service	Oracle Database	Digitalocean Droplet	Airtab
Just Cloud			

4. As neural networks require inputs with the same length and dimension, padding is applied to consider a threshold number of words for all sequences.

If a sequence is shorter than the threshold, extra 0s are added, and sequences are truncated if they are longer than the threshold. Padding is used to increase the computational efficiency and the performance of the neural network model. Most CLOSER-DREAM reviews are long, with a maximum review length of 818 words, a minimum review length of 3 words, and a median review length of 64 words. The padding threshold is usually set to the maximum length of the longest sentence in the training set, which is equal to 818 characters in our experiments. This choice is generally made by the mostly conducted works related to applying deep learning models for sentiment analysis classification [29], [30]. We have chosen to use the 'post' padding, which means that our sentence sequence numeric representations corresponding to word index entries will appear at the left-most positions of our resulting sentence vectors. In contrast, the padding characters ('0') will appear after our actual data at the right-most positions. Table 3 illustrates a sample review's tokenization, word indexing, sequencing, and padding.

TABLE 3. Review conversion to padded sequences.

Text	it is an excellent cloud platform for visual studio users I like it.
Tokens	‘it’, ‘is’, ‘excellent’, ’cloud’, ‘platform’, ‘for’, ’visual’, ‘studio’, ‘users’, ‘i’, ‘like’, ‘it’, ‘.’
Word_to_ Index	{'it': 1, 'is': 2, 'an': 3', excellent': 4, 'cloud': 5, 'platform': 6, 'for': 7, 'visual': 8, 'studio': 9, 'users': 10, 'I': 11, 'like': 12}
Sequence	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1]
Padded Sequence	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 0, 0, 0, 0, 0, 0, 0, 0 ,0 ,0 ,0,0 ,0 ,0 ,0 ,0, 0, 0, 0, …, 0]

2) WORD EMBEDDINGS

An index value represents each token in the previous preprocessing steps. However, these indices do not reflect any relationship between the tokens, i.e., indices' numerical order does not have much conventional meaning. Therefore, an extra encoding step, known as embedding, is required to create a dense representation of each preprocessed token to reflect their relationships. The preprocessed tokens serve as input to the word embedding layer, which is the first layer in the proposed model. This layer converts the inputs to vector representations that capture the semantic meanings of words and reflect the relationship among them.

3) HYBRID CONVOLUTIONAL BiLSTM DEEP LEARNING MODEL (CBi-LSTM) FOR SENTIMENT CLASSIFICATION

CNN and RNN are typical models for sentiment classification [16]. The difference between the two models is that CNN can extract local features by examining the spatial relationship within the data but cannot learn sequential correlations [31]. On the other hand, RNN looks for the temporal relationship and can extract global features [32]. While RNNs are suitable for sequential relationships, traditional RNNs are susceptible to gradient explosion or vanishing when exposed to long data sequences. LSTM [31] is an extension of RNN that can prevent these problems. Also, it can remember long-term dependencies with chains of memory cells as hidden units. BiLSTM [33] enhances LSTM by combining two LSTM layers to process information sequences in forward and backward directions in parallel.

We propose a novel hybrid model that combines CNN with BiLSTM layers in this work. This model is named CBi-LSTM. At the top of CBi-LSTM, three CNN layers are added to extract the most important n-gram features from text vectors and, thus, reduce the dimensionality. The successive BiLSTM layer is fed with the extracted features from the top CNN layers and investigates their contexts to capture phrase-level patterns. In addition, a batch normalization layer is added to standardize the inputs from the BiLSTM layer and stabilize the learning process without changing vector dimensions. The output of the batch normalization layer is passed to a global Max pooling layer. This layer facilitates the transition to the output prediction layer, also named the dense layer, by downsampling each representation vector to a single value. Figure 3 illustrates the architecture of the proposed CBi-LSTM.

In a sentence like ''azure is a reliable, affordable cloud platform'', the words ''reliable'' and ''affordable'' express a positive sentiment about the cloud service provider. At the top layers of CBi-LSTM, CNN filters capture the features from sequential groups of words (phrases). Therefore, the positive sentiment in the keywords ''reliable'' and ''affordable'' could be predicted correctly by a single CNN. However, in sentences like ''Do not miss out on Google platform'' and ''Do not waste your time with Google platform'', the two phrases ''do not miss'' and ''do not waste'' convey different opinions. As CNN extracts word-level patterns, it could classify both sentences as negative comments, although the former implies a positive sentiment. Adding a BiLSTM layer enables CBi-LSTM to remember the past and forward contexts to detect the phrase- level patterns, which could help predict both sentences' classes correctly. BiLSTM could also effectively predict complex sentences with dependencies between features like: ''Although this is a good platform, its functions appear to have some delay in revealing data''.

C. REPUTATION ASSESSMENT AND MODEL VALIDATION PHASE

The goal of this phase is to apply the NBR formula by using the proposed deep learning model results to assess the reputation of cloud service providers [34]. It also aims to validate the effectiveness of using the proposed deep learning model for reputation assessment.

NBR is the net value of a brand reputation estimated from published reviews. It employs sentiment analysis to measure clients' satisfaction levels. The NBR index focuses more on the positive feedback from brand promoters than on the negative ones. The output of NBR can be any value in the range [−100,100]. Higher values mean that more positive reviews are considered. The NBR equation is illustrated by Eq. 1.

$$NBR = \left(\frac{Positive\ Reviews - Negative\ Reviews}{Positive\ Reviews + Negative\ Reviews}\right) \times 100 \tag{1}$$

To substitute the positive reviews and negative reviews values in Eq. 1, the confusion matrix of the proposed deep learning model is used. The confusion matrix is a performance measure that reports the number of ''True Positive'' (TP), ''True Negative'' (TN), ''False Positive'' (FP), and ''False Negative'' (FN) values. The TP value substitutes the positive reviews' value in NBR, whereas the TN value substitutes the negative reviews' value [35]. TP represents the truly predicted labels as ''Positive'', whereas TN denotes the truly predicted labels as ''Non-Positive''. The latter includes both the ''Negative'' and ''Neutral'' labels.

To validate the effectiveness of using the proposed deep learning model for reputation assessment, the resulting NBR score is compared to the real/actual data-based NBR score. The NBR score uses the total numbers of positive and non-positive labels counted from the original dataset. The total number of positive reviews in the dataset substitutes the positive reviews variable in NBR, while the total number of negative plus neutral reviews substitutes the negative reviews variable.

IV. EXPERIMENTATION AND VALIDATION

This section focuses on the experiments conducted to validate the proposed approach. First, CBiLSTM classification performance is tested and compared to other baseline models. Second, computations are performed to compare the CBiLSTM-based NBR to Google Cloud's actual/ real data-based NBR.

A. VALIDATION OF CBiLSTM CLASSIFICATION PERFORMANCE

This subsection presents an experimental study conducted on CLOSER-DREAM to evaluate the classification performance of CBiLSTM.

1) DATASET PREPARATION

The reviews are extracted using the Web Scraper[5] extension offered by Google Chrome. Multiple review websites are scraped, including Capterra,[6] g2,[7] Gartner,[8] TrustRadius,[9] Software Advice,[10] GetApp,[11] Trust Pilot,[12] and Spiceworks.[13] Because the reviews are gathered from several websites, some reviewers can submit the same review on more than one website. This causes a data redundancy problem in the dataset. To solve this issue, Python code is implemented to remove duplicates from CLOSER-DREAM.

[5] https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn?hl=en

[6] https://www.capterra.com/

[7] https://www.g2.com/

[8] https://www.gartner.com/reviews/home

[9] https://www.trustradius.com/

[10] https://www.softwareadvice.com/

[11] https://www.getapp.com/

[12] https://www.trustpilot.com/

[13] https://www.spiceworks.com/

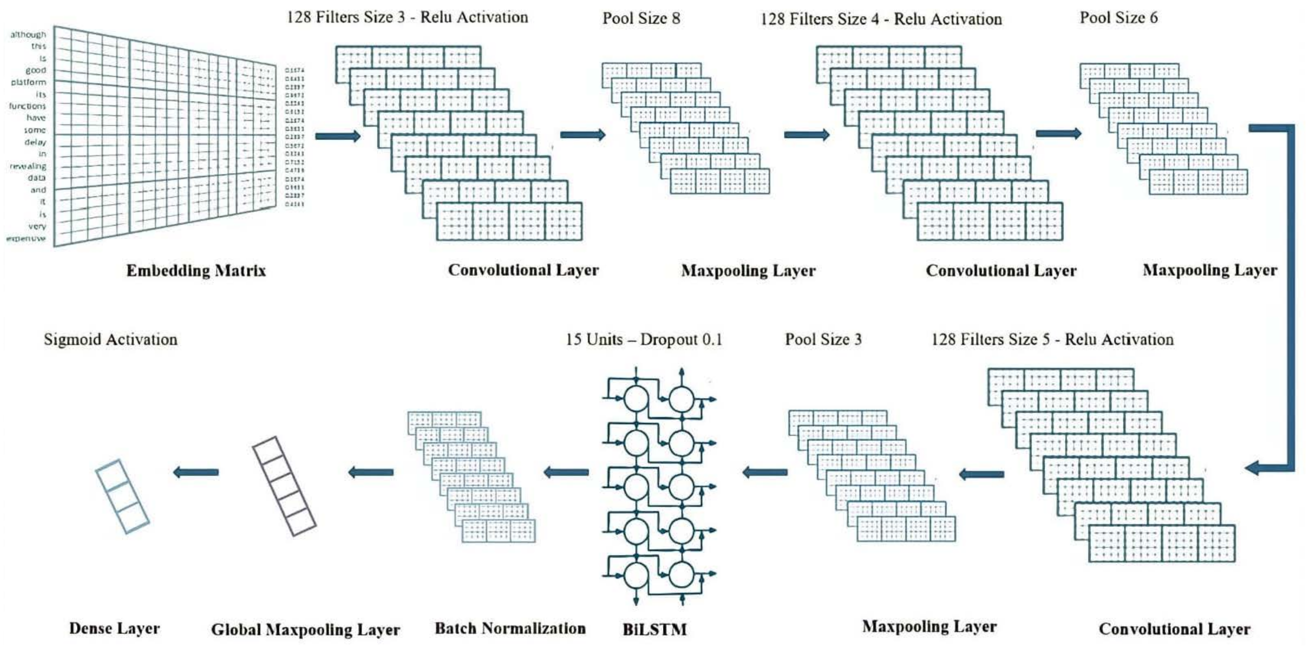

FIGURE 3. CBi-LSTM architecture.

In addition, some reviews contain noise that must be cleaned. For example, some reviews end or start with sentences like: "This review was collected by g2 website", "show more show less", or "Published on 9/4/2020". Such sentences are useless for categorizing the reviews; thus, they are removed. After cleaning noise, the dataset is automatically labeled using Valence Aware Dictionary for sEntiment Reasoning (VADER) [36], a sentiment analysis Python package. VADER is an open-source lexicon and rule-based tool for sentiment analysis. It determines reviews polarity and classifies them in multiple sentiment analysis classes. Vader extracts sentimental words and their corresponding intensity from sentences. It returns a polarity score between −4 and 4 of each word. The closer the score to −4, the more intense the word's negativity is, and the closer the score to 4, the more intense its positivity is. The word scores are then normalized to obtain an overall statement sentiment score, known as Compound Score. This score reflects statement polarity and corresponding intensity, and it falls in the range between −1 and 1. The compound score's formula is given in Eq. 2, where x is the sum of polarity scores of constituent words and α is a normalization constant, the default value is 15.

$$Compound\ Score = \frac{x}{\sqrt{x^2 + \alpha}} \tag{2}$$

Because VADER's results are not totally accurate, labels of minority classes, neutral and negative, are checked and updated manually. Also, labels with a compound score lower than 0.7 are checked and updated manually. Other reviews with higher compound scores (0.8, 0.9, 1.0) have more intense polarity, thus, are more likely to be classified correctly.

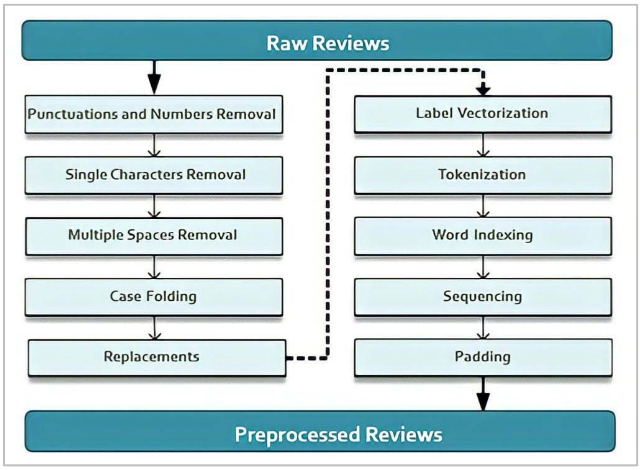

FIGURE 4. Pre-processing steps.

Before implementing CBiLSTM using the CLOSER-DREAM dataset, some preparation tasks are required to convert the textual reviews into a machine-understandable format.

Several preprocessing steps, illustrated in Figure 4, are followed in this work. First, punctuations, numbers, single characters, and multiple spaces are removed, and all characters are converted into lower case characters. Then, some replacements are made to get more uniformity in text, such as replacing "'ll" with "will" and "I 'm" with "I am". Also, all three labels are vectorized in such a way that "positive" corresponds to "1, 0, 0", "negative" to "0, 1, 0", and "neutral" to "0, 0, 1". Sentences are then tokenized, a word-to-index

TABLE 4. CNN structure.

Layers	Output Dimensions	Parameters	Activation/ Dropout
Input	(None, 818)		
Embedding	(None, 818, 100)		
Convolution	(None, 816, 128)	(128 filters of size 3)	Relu
MaxPooling	(None, 136, 128)	(Pool size 6)	
Convolution	(None, 133, 128)	(128 filters of size 4)	Relu
Max Pooling	(None, 44, 128)	(Pool size 3)	
Convolution	(None, 40, 128)	(128 filters of size 5)	Relu
Global Max Pooling	(None, 128)		
Dense	(None, 128)		Relu
Dense	(None, 3)		Sigmoid

TABLE 5. BiLSTM structure.

Layers	Output Dimensions	Parameters	Activation/ Dropout
Input	(None, 818)		
Embedding	(None, 818, 100)		
Bi- LSTM Layer	(None, 818, 30)	(15 units)	0.1
Global Max Pooling	(None, 30)		
Dense	(None, 3)		Sigmoid

TABLE 6. GRU structure.

Layers	Output Dimensions	Parameters	Activation/ Dropout
Input	(None, 818)		
Embedding	(None, 818, 100)		
Bidirectional (LSTM)	(None, 818, 15)	(15 units)	0.1
Global Max Pooling	(None, 15)		
Dense	(None, 3)		Sigmoid

TABLE 7. CBiLSTM structure.

Layers	Output Dimensions	Parameters	Activation/ Dropout
Input	(None, 818)		
Embedding	(None, 818, 100)		
Convolution	(None, 816, 128)	(128 filters, size 3)	Relu
MaxPooling	(None, 102, 128)	(Pool size 8)	
Convolution	(None, 99, 128)	(128 filters, size 4)	Relu
Max Pooling	(None, 16, 128)	(Pool size 6)	
Convolution	(None, 12, 128)	(128 filters, size 5)	Relu
Max Pooling	(None, 4, 128)	(Pool size 3)	
Bi- LSTM Layer	(None, 4, 30)	(15 units)	0.1
Batch Normalization	(None, 4, 30)		
Global Max Pooling	(None, 30)		
Dense	(None, 3)		Sigmoid

dictionary and padded sequences are created. In this work, the maximum sequence length is set to 818.

For the word embeddings, this work exploits GloVe [37] representation, which is an unsupervised learning algorithm that leverages word co-occurrence frequencies.

2) BASELINE MODELS

This work compares the classification performance of CBiLSTM to three deep learning models. These are as follows:

- CNN: relies on convolution and pooling layers and applies convolutional filters to capture local features.
- BiLSTM: combines two opposite LSTM layers for context analysis.
- GRU: GRU stands for Gated Recurrent Unit. It is an RNN variation with a simpler architecture than LSTM. It has no internal memory and has fewer gates than LSTM.

Tables 4, 5, 6, and 7 provide summaries of the structures of the experimented models.

3) EXPERIMENTAL SETUP

This work conducts various experiments to benchmark the performance of CBiLSTM to the other three baseline models. In addition, it carries out some experiments to present the effect of using different resampling techniques with CBiLSTM to handle the dataset's imbalanced nature. All the experiments were performed on Google Colab[14] using Python 3.7.10[15] and Keras 2.4.3.[16] The proposed approach was implemented using a desktop computer that has the following configuration: an 11th Gen Intel® Core™ i9-11900H@ 2.50 GHz processor and a 32 GB RAM. An NVIDIA GeForce RTX 3080 Ti 16 GB graphics card is used to facilitate the smooth training of the proposed classifier.

CLOSER-DREAM is split into 70% for model training and 30% for validation and testing. The shuffle feature is disabled so that Google Cloud-related reviews remain in the validation and testing portions of CLOSER-DREAM. The testing portion includes only reviews related to Google cloud services. This portion is used to assess the reputation of Google as a cloud services provider. Table 8 shows the number of samples in each subset.

Moreover, 400,000 pre-trained vectors in the "glove.6B.100d.txt" file [38] are used to prepare the embed-

[14] https://colab.research.google.com/notebooks/intro.ipynb

[15] https://www.python.org/downloads/release/python-3710/

[16] https://keras.io/

TABLE 8. Number of samples in training, validation, and testing subsets.

Number of Training Samples	8224
Number of Validation Samples	1000
Number of Testing Samples	3954

ding layer. The parameters used for the embedding layer include 20,000 as the maximum vocabulary size, 818 as the maximum sequence length, and 100 as the embedding dimension for all the models. For all models' training, the batch size is set to 128, and the number of epochs is 70. The gradient descent algorithm is employed to set up the optimal hyperparameters (i.e., batch size, epochs, optimizer, momentum, and weight decay) of the different models deployed for the experimentation. This technique is extensively employed to minimize the cost/loss function to develop machine learning and deep learning-based applications. Gradient descent [39] is an iterative first-order optimization algorithm that identifies a local minimum/maximum function. In the gradient descent technique, we start with random model parameters and calculate the error for each learning iteration, then continuously changing the model parameters to get values closer to the values that result in the lowest cost. Because this is the steepest descent, the objective is to take repeated steps in the opposite direction of the function's gradient (or approximation gradient) at the current position. Stepping in the direction of the gradient, on the other hand, will result in a local maximum of that function; this is known as gradient ascent.

Furthermore, to enhance the obtained performance results, all minority classes in the training subset, negative and neutral, are oversampled before classifiers' implementation.

4) RESULTS AND DISCUSSION

Five metrics are used in this work to evaluate models' performance: precision, recall, F1-score, accuracy, and the confusion matrix [40]. Precision is the fraction of results that the model accurately predicts. The recall is the fraction of the model's relevant results correctly predicted. F1-score is a balanced metric that reflects the harmonic mean of both precision and recall. Accuracy evaluates "How good is a model's performance?". It reflects how regularly the model's predicted label is right. Eq. 3, 4, 5, and 6 illustrate the equations of the performance metrics mentioned above.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + TN} \tag{4}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

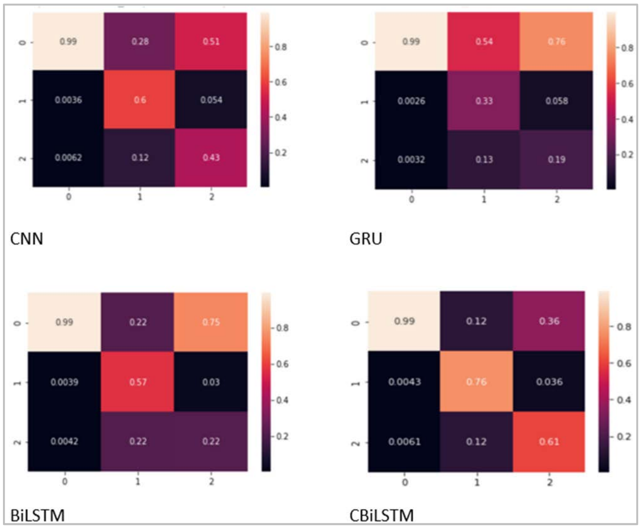

FIGURE 5. Normalized confusion matrices of CNN, BiLSTM, GRU, and CBiLSTM models.

As this work deals with a multiclass classification on an imbalanced dataset, both the macro-averaged and weighted average scoring metrics of recall, precision, and F1-score are considered.

The macro-averaged measure is the arithmetic mean of all class scores. It treats all classes equally regardless of their proportions in the dataset. For example, the macro-average precision is the mean of the precision scores of classes positive, negative, and neutral. On the other hand, the weighted average measure is the average of weighted class scores. It multiplies each class score by its corresponding class proportion.

The remaining part of this section provides relevant illustrations about the performance of all DL models that are compared.

Table 9 shows a detailed classification report of each class per model. CBiLSTM achieves the highest precision, recall, and F1 scores for all classes (i.e., positive, negative, and neutral). For positive reviews, CBiLSTM guarantees 100% of recall. For negative reviews, it achieves a precision of 76%. Finally, GRU provides the highest recall of 60% for neutral reviews, whereas our model performs better for classifying this class in terms of precision and F1-score, 76% and 54%, respectively. Tables 10 and 11 show the experimented models' macro-averaged and weighted scores, respectively. CBiLSTM has the highest macro-average precision and F1-score, while GRU achieves the highest recall. CBiLSTM outperforms all the other models for the weighted average by ensuring an overall precision of 98%, a recall of 99%, and an F1-score of 98%. Table 12 shows the training times of each model. It indicates that CNN requires the least training time, followed by CBiLSTM. The training time of our proposed classifier remains reasonable compared to the training time of GRU and BiLSTM models.

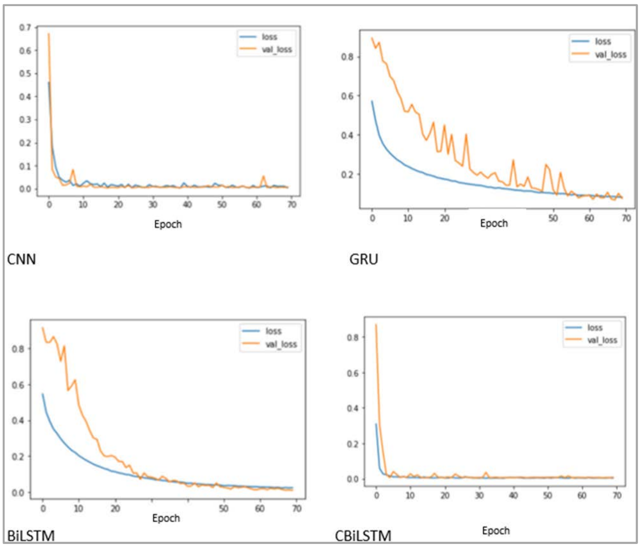

FIGURE 6. Training and validation loss learning curves of CNN, BiLSTM, GRU, and CBiLSTM models.

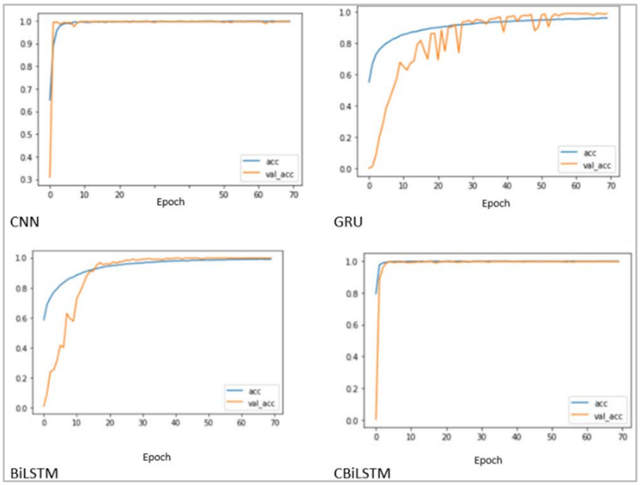

FIGURE 7. Training and and validation accuracy learning curves of CNN, BiLSTM, GRU, and CBiLSTM models.

Figure 5 depicts the models' confusion matrices, normalized by predictions. The diagonal of the CBiLSTM matrix shows the lightest colors and the highest accurate predictions per class. In Figure 5, it is clear that our proposed classifier outperforms the other models considered in these experiments by offering the highest accuracies for classifying the different reviews' classes.

All models' training and validation loss and their training and validation accuracy are depicted in Figures 6 and 7. As shown in these figures, the learning curves of CBiLSTM and CNN present good fits. In case of a good fit, the training and validation losses decline to a stable point with a minimum gap between the two curves at the end. In comparison to CNN, our proposed classifier's loss and accuracy learning curves have shown fewer fluctuations, as demonstrated by Figure 8, which presents all models' validation accuracy learning curves in one line chart. This conclusion is also confirmed by Figure 9, which provides a closer look into the differences between the CNN and CBiLSTM validation accuracy learning curves.

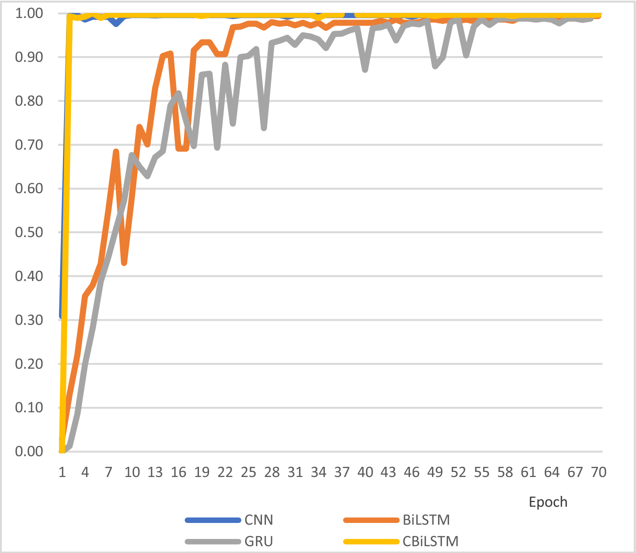

FIGURE 8. Validation accuracy learning curves of CNN, BiLSTM, GRU, and CBiLSTM models.

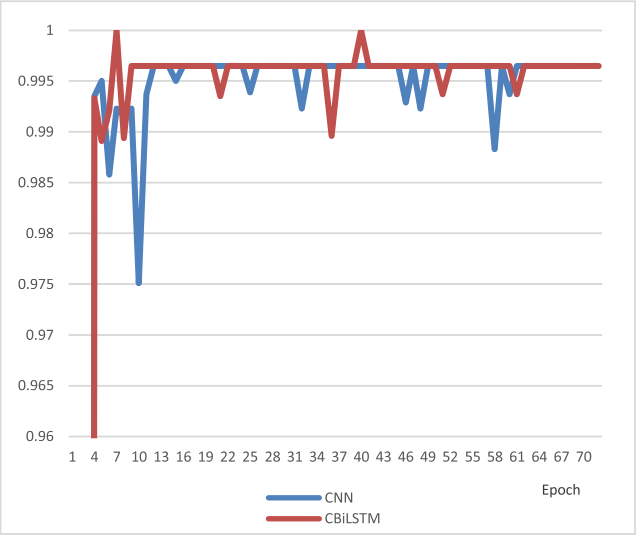

FIGURE 9. Validation accuracy learning curves of CNN vs. CBiLSTM.

5) VALIDATION OF CBiLSTM FOR REPUTATION ASSESSMENT

To validate the performance of CBiLSTM in reputation assessment, the NBR score of Google cloud is calculated two times, and the results are compared. First, it is generated based on the confusion matrix of CBiLSTM. Second, it is generated based on the original dataset reviews numbers.

To calculate the NBR equation provided in Eq. 1, the confusion matrix of the CBiLSTM model applied on the testing set is used. The TP value substitutes the positive reviews' value in NBR, whereas the TN value substitutes the negative reviews' value. However, CBiLSTM classifies reviews into three classes, and the results in its confusion matrix are presented in a 3*3 matrix. To derive the TP and TN values from this multiclass confusion matrix, we need to transform

TABLE 9. Values of the main classification metrics per class.

DL Model	Classes	Precision (%)	Recall (%)	F1-score (%)
CNN	Positive	99	99	99
	Negative	60	48	54
	Neutral	43	37	40
BiLSTM	Positive	99	98	99
	Negative	57	42	48
	Neutral	22	51	31
GRU	Positive	99	97	98
	Negative	33	42	37
	Neutral	19	**60**	29
CBiLSTM	Positive	**99**	**100**	**99**
	Negative	**76**	**42**	**54**
	Neutral	**61**	40	**48**

TABLE 10. Macro-average measures of precision, recall, and F1-score for CNN, BiLSTM, GRU, and CBiLSTM models.

	CNN	BiLSTM	GRU	CBiLSTM
Precision (%)	67	59	50	**79**
Recall (%)	62	64	**66**	60
F1-score (%)	64	59	55	**67**

TABLE 11. Weighted average measures of precision, recall, and F1-score for CNN, BiLSTM, GRU, and CBiLSTM models.

	CNN	BiLSTM	GRU	CBiLSTM
Precision (%)	98	98	98	**98**
Recall (%)	98	97	96	**99**
F1-score (%)	98	97	97	**98**

TABLE 12. Training time for CNN, BiLSTM, GRU, and CBiLSTM models.

	CNN	BiLSTM	GRU	CBiLSTM
Training Time (ms)	**445.37**	985.41	613.98	519.43

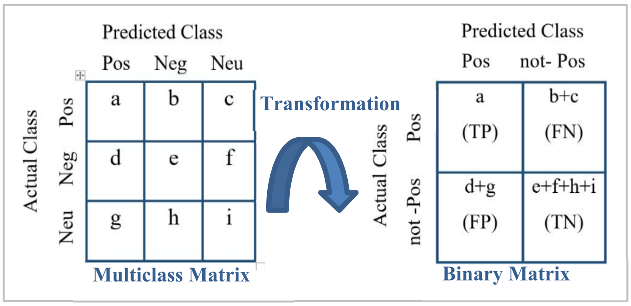

FIGURE 10. Transformation of a multiclass confusion matrix into a binary matrix.

the obtained confusion matrix into a binary confusion matrix [47]. The transformation process is illustrated in Figure 10.

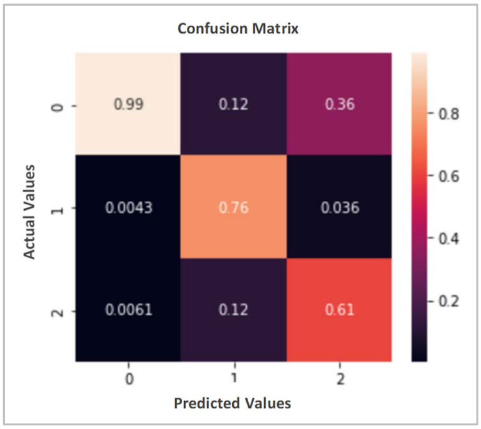

FIGURE 11. CBiLSTM's normalized confusion matrix.

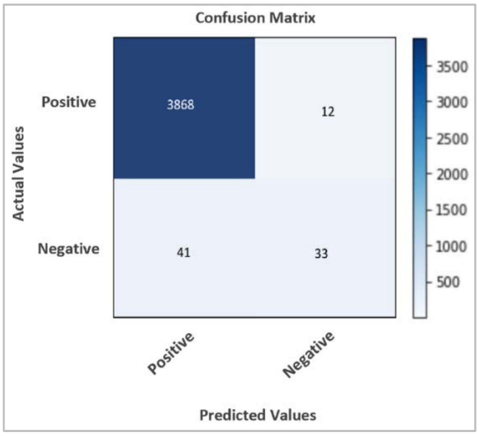

FIGURE 12. CBiLSTM's binary matrix.

Figure 11 shows the normalized confusion matrix of the CBiLSTM model.

The result of transforming the multi-class confusion matrix provided in Figure 11 into a binary confusion matrix is shown in Figure 12. The latter figure is generated through implementation. It classifies the reviews into "Positive" and "not-Positive" classes. Based on the CBiLSTM results, the resulting classification is utilized to determine the NBR score.

Based on the obtained binary matrix, the number of positive reviews is 3868, which is the TP value, whereas the number of negative reviews is 33, which is the TN value. According to the result obtained by Eq.7, the NBR score of Google Cloud services is estimated at 98.3 %.

$$NBR = \left(\frac{3868 - 33}{3868 + 33}\right) \times 100 = 98.3\% \tag{7}$$

The testing dataset contains 3,880 positive reviews, 31 negative reviews, and 43 neutral reviews. 3,880 substitutes for the positive reviews in NBR, whereas the sum of the negative and neutral reviews, 31 plus 43, substitutes for the negative reviews in NBR. The reputation score of Google Cloud is estimated as 96.25% based on 3954 reviews as calculated by Eq. 8.

$$NBR = \left(\frac{3880 - 74}{3880 + 74}\right) \times 100 = 96.25\% \tag{8}$$

Comparing the CBiLSTM-based NBR score of Google Cloud to the NBR score generated from the original dataset, the two values are close. This indicates that CBiLSTM can be considered a reliable technique for the reputation assessment of service providers.

V. CONCLUSION

In recent years, the number of competing services in the cloud services industry has increased. Although this has numerous advantages, certain difficulties occur when clients must choose amongst a range of services that provide the same functionality. The literature discusses several objective and subjective measurements of service trust. However, the existing approaches are challenged by staticity, acquisition feasibility, data sparsity, and cold start issues. This paper presents a novel approach to dealing with these issues. It develops a new deep learning model to classify cloud service-related reviews based on sentiments derived from the examined reviews. The proposed deep learning model, named CBiLSTM, is a hybrid model that combines CNN and BiLSTM layers. The CNN layers handle the high dimensionality of text inputs by extracting word-level features, and the BiLSTM layer investigates the context of the formerly extracted features in backward and forward directions simultaneously. The CBiLSTM's classification results are utilized to compute the overall reputation score using the NBR formula. Multiple experiments were carried out to validate the proposed approach. First, the performance of CBiLSTM is compared to that of CNN, BiLSTM, and GRU models, and the findings show that CBiLSTM surpasses these models. Second, experiments using Google Cloud reviews indicated that CBiLSTM is a reliable method for assessing service providers' reputations. The goal of this work is to provide a reputation score for service providers based on user QoE, which is represented by categorizing reviews as "Positive", "Negative", or "Neutral", and provides an overall assessment of user sentiments regarding services providers.

Despite the numerous contributions made by this study, it presents a number of shortcomings. Mainly, it needs to perform more in-depth and refined research aiming at investigating the multimodal sentiment analysis techniques for reputation assessment. Indeed, multimodal content has evolved from text content to multimedia material including videos and images as a medium for user expression on the web today. Textual material has given way to multimedia data including films and photographs in multimodal content. For a variety of decision-making applications, these multimodal forms of expression have become the standard information resource. Although these new forms of expression provide more affluent and more expressive information resources, their dispersion in terms of multimodal emotional expressions needs a more complex analysis to extract relevant and valuable data.

As future work, we plan to extend our approach beyond text-based sentiment analysis techniques and make significant contributions by deploying the promising multimodal sentiment analysis techniques to ensure the effective assessment of services' reputation. Also, we intend to turn the multiclass classification problem into a multi-label problem to extract additional and more valuable characteristics from reviews. For example, depending on customers' subjective sentiments, we can categorize reviews to reflect service aesthetics, affordability, usability, security, and QoS attributes. Furthermore, we aim to enhance the proposed classifier to ensure the detection and classification of ironic or sarcastic reviews to increase the overall reputation assessment's accuracy. Finally, to address the unbalanced nature of the CLOSER-DREAM dataset, our future work will include the investigation of the resampling strategies to provide considerable improvements for the overall performance of the suggested CBiLSTM model

ACKNOWLEDGMENT

The authors would like to thank the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

REFERENCES

[1] (2021). *Gartner Says Four Trends are Shaping the Future of Public Cloud*. Accessed: Jan. 15, 2022. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2021-08-02-gartner-says-four-trends-are-shaping-the-future-of-public-cloud

[2] *Gartner Says Worldwide IaaS Public Cloud Services Market Grew 40.7% in 2020*. Gartner. Accessed: Jan. 15, 2022. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2021-06-28-gartner-says-worldwide-iaas-public-cloud-services-market-grew-40-7-percent-in-2020

[3] F. N. Nwebonyi, R. Martins, and M. E. Correia, "Reputation-based security system for edge computing," in *Proc. 13th Int. Conf. Availability, Rel. Secur.*, Aug. 2018, pp. 1–8.

[4] S. Deshpande and R. Ingle, "Evidence based trust estimation model for cloud computing services," *Int. J. Netw. Secur.*, vol. 20, no. 2, pp. 291–303, 2018.

[5] S. Rizvi, J. Mitchell, A. Razaque, M. R. Rizvi, and I. Williams, "A fuzzy inference system (FIS) to evaluate the security readiness of cloud service providers," *J. Cloud Comput.*, vol. 9, no. 1, p. 42, Dec. 2020.

[6] L. F. Bilecki and A. Fiorese, "A trust reputation architecture for cloud computing environment," in *Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2017, pp. 614–621.

[7] K. Papadakis-Vlachopapadopoulos, R. S. González, I. Dimolitsas, D. Dechouniotis, A. J. Ferrer, and S. Papavassiliou, "Collaborative SLA and reputation-based trust management in cloud federations," *Future Gener. Comput. Syst.*, vol. 100, pp. 498–512, Nov. 2019.

[8] Q. She, X. Wei, G. Nie, and D. Chen, "QoS-aware cloud service composition: A systematic mapping study from the perspective of computational intelligence," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112804.

[9] M. Driss, A. Aljehani, W. Boulila, H. Ghandorh, and M. Al-Sarem, "Servicing your requirements: An FCA and RCA-driven approach for semantic web services composition," *IEEE Access*, vol. 8, pp. 59326–59339, 2020.

[10] M. Driss, S. Ben Atitallah, A. Albalawi, and W. Boulila, "Req-WSComposer: A novel platform for requirements-driven composition of semantic web services," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 2, pp. 849–865, Feb. 2022.

[11] O. A. Wahab, J. Bentahar, H. Otrok, and A. Mourad, "A survey on trust and reputation models for web services: Single, composite, and communities," *Decision Support Syst.*, vol. 74, pp. 121–134, Jun. 2015.

[12] X. Liu, A. Kale, J. Wasani, C. Ding, and Q. Yu, "Extracting, ranking, and evaluating quality features of web services through user review sentiment analysis," in *Proc. IEEE Int. Conf. Web Services*, Jul. 2015, pp. 153–160.

[13] (2011). *The NIST Definition of Cloud Computing*. Accessed: Jan. 15, 2022. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-145/final#:~:text=Cloud%20computing%20is%20a%20model,effort%20or%20service%20provider%20interaction

[14] M. N. O. Sadiku, S. M. Musa, and O. D. Momoh, "Cloud computing: Opportunities and challenges," *IEEE Potentials*, vol. 33, no. 1, pp. 34–36, Jan./Feb. 2014.

[15] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020.

[16] C. Mao, R. Lin, C. Xu, and Q. He, "Towards a trust prediction framework for cloud services based on PSO-driven neural network," *IEEE Access*, vol. 5, pp. 2187–2199, 2017.

[17] S. Latif, M. Driss, W. Boulila, Z. E. Huma, S. S. Jamal, Z. Idrees, and J. Ahmad, "Deep learning for the industrial Internet of Things (IIoT): A comprehensive survey of techniques, implementation frameworks, potential applications, and future directions," *Sensors*, vol. 21, no. 22, p. 7518, Nov. 2021.

[18] S. Ben Atitallah, M. Driss, W. Boulila, and H. Ben Ghezala, "Randomly initialized convolutional neural network for the recognition of COVID-19 using X-ray images," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 1, pp. 55–73, 2022.

[19] S. Ben Atitallah, M. Driss, W. Boulila, A. Koubaa, and H. Ben Ghézala, "Fusion of convolutional neural networks based on Dempster–Shafer theory for automatic pneumonia detection from chest X-ray images," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 2, pp. 658–672, 2022.

[20] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[21] W. J. Fan, S. L. Yang, H. Perros, and J. Pei, "A multi-dimensional trust-aware cloud service selection mechanism based on evidential reasoning approach," *Int. J. Automat. Comput.*, vol. 12, no. 2, pp. 208–219, 2015.

[22] T. Noor, Q. Z. Sheng, L. Yao, S. Dustdar, and A. H. H. Ngu, "CloudArmor: Supporting reputation-based trust management for cloud services," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 367–380, Feb. 2016.

[23] S. Ding, Z. Wang, D. Wu, and D. L. Olson, "Utilizing customer satisfaction in ranking prediction for personalized cloud service selection," *Decision Support Syst.*, vol. 93, pp. 1–10, Jan. 2017.

[24] N. Somu, G. R. Gauthama, V. Kalpana, K. Kirthivasan, and S. S. Shankar, "An improved robust heteroscedastic probabilistic neural network based trust prediction approach for cloud service selection," *Neural Netw.*, vol. 108, pp. 339–354, Dec. 2018.

[25] J. Liu and Y. Chen, "A personalized clustering-based and reliable trust-aware QoS prediction approach for cloud service recommendation in cloud manufacturing," *Knowl.-Based Syst.*, vol. 174, pp. 43–56, Jun. 2019.

[26] X. Li, "FASTCloud: A framework of assessment and selection for trustworthy cloud service based on QoS," 2020, *arXiv:2011.01871*.

[27] O. Wahab, J. Bentahar, H. Otrok, and A. Mourad, "Towards trustworthy multi-cloud services communities: A trust-based hedonic coalitional game," *IEEE Trans. Services Comput.*, vol. 11, no. 1, pp. 184–201, Jan./Feb. 2018.

[28] P. Casas and R. Schatz, "Quality of experience in cloud services: Survey and measurements," *Comput. Netw.*, vol. 68, pp. 149–165, Aug. 2014.

[29] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar. 2017.

[30] M. Gimnez, J. Palanca, and V. Botti, "Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis," *Neurocomputing*, vol. 378, pp. 315–323, Feb. 2020.

[31] Y. Luan and S. Lin, "Research on text classification based on CNN and LSTM," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Mar. 2019, pp. 352–355.

[32] J. Du, C.-M. Vong, and C. L. P. Chen, "Novel efficient RNN and LSTM-like architectures: Recurrent and gated broad learning systems and their applications for text classification," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1586–1597, Mar. 2021.

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

[34] D. K. B. Kulevome, H. Wang, and X. Wang, "A bidirectional LSTM-based prognostication of electrolytic capacitor," *Prog. Electromagn. Res. C*, vol. 109, pp. 139–152, 2021.

[35] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020.

[36] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in *Proc. Int. Multiconference Eng. Comput. Scientists*, vol. 122, 2019, p. 16.

[37] N. A. Vidya, M. I. Fanany, and I. Budi, "Twitter sentiment to analyze net brand reputation of mobile phone providers," *Proc. Comput. Sci.*, vol. 72, pp. 519–526, Jan. 2015.

[38] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[39] G. Morse and K. O. Stanley, "Simple evolutionary optimization can rival stochastic gradient descent in neural networks," in *Proc. Genetic Evol. Comput. Conf.*, Jul. 2016, pp. 477–484.

[40] Y. Li, Q. Pan, T. Yang, S. Wang, J. Tang, and E. Cambria, "Learning word representations for sentiment analysis," *Cognit. Comput.*, vol. 9, no. 6, pp. 843–851, Dec. 2017.

REEM AL SALEH received the M.Sc. degree (Hons.) in information systems from the College of Computer Science and Engineering (CCSE), Taibah University, Saudi Arabia, in 2021. Her primary research interests include software engineering, data science, service computing, and artificial intelligence.

MAHA DRISS (Senior Member, IEEE) received the Engineering degree (Hons.) in computer science and the M.Sc. degree from the National School of Computer Science (ENSI), University of Manouba, Tunisia, in 2006 and 2007, respectively, and the Ph.D. degree conjointly from the University of Manouba and the University of Rennes 1, France, in 2011. From 2012 to 2015, she was an Assistant Professor in computer science at the National Higher Engineering School of Tunis, University of Tunis, Tunisia. From 2015 to 2021, she was an Assistant Professor of computer science at the IS Department, College of Computer Science and Engineering, Taibah University, Saudi Arabia. She is currently an Assistant Professor of computer science and a Senior Researcher with Prince Sultan University, Saudi Arabia. She is also a Senior Researcher with the RIADI Laboratory, University of Manouba. Her primary research interests include software engineering, service computing, distributed systems, cybersecurity, the IoT, the IIoT, and artificial intelligence. She served as a reviewer in several world-leading high-impact journals and she has chaired tracks and participated as a reviewer at a number of international conferences.

IMAN ALMOMANI (Senior Member, IEEE) received the bachelor's degree from United Arab Emirates, in 2000, the master's degree in computer science from Jordan, in 2002, and the Ph.D. degree in wireless network security from De Montfort University, U.K., in 2007. She is currently an Associate Professor in cybersecurity. She is the Associate Director with the Research and Initiatives Centre (RIC) and also the Leader with the Security Engineering Laboratory (SEL), Prince Sultan University (PSU), Riyadh, Saudi Arabia. Before Joining PSU, she worked as an Associate Professor and the Head of the Computer Science Department, The University of Jordan, Jordan. Her research interests include wireless networks and security, mainly wireless mobile *ad-hoc* networks (WMANETs), wireless sensor networks (WSNs), multimedia networking (VoIP), and security issues in wireless networks. She is also interested in the area of electronic learning (e-learning) and mobile learning (m-learning). She has several publications in the above areas in a number of reputable international and local journals and conferences. She is in the organizing and technical committees for a number of local and international conferences. She is also a Senior Member of IEEE WIE. Also, she serves as a reviewer and a member for the editorial board in a number of international journals.