Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 51, No. 1, 2025, pp. 27-37

DOI: https://doi.org/10.25195/ijci.v51i1.547
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Hybrid LSTM–Seq2Seq Models: Improved Patient Interaction for Healthcare Chatbots

[1*]Saba .A .Ali
Informatics Institute of Postgraduate Studies, Iraqi Commission for Computers and Informatics
Darbandikhan Technical Institute, Sulaimani Polytechnic University, Sulaimani, KRD,IRAQ
ms202220725@iips.edu.iq

[2] Prof. Dr. Suha Mohammed Hadi
Informatics Institute of Postgraduate Studies, Iraqi Commission for Computers and Informatic
Baghdad Iraq
dr.suhahadi@gmail.com

[3] Dr. Mustafa Musa
Informatics Institute of Postgraduate Studies, Iraqi Commission for Computers and Informatic
Baghdad Iraq
Mustafa.musa@iips.edu.iq

ARTICLE INFO
Article History
Received: 22/02/2025
Accepted: 06/03/2025
Published: 24/04/2025

ABSTRACT

Healthcare chatbots play a critical role in improving communication between patients and healthcare providers by offering accurate and timely responses. A novel approach is proposed, which leverages a deep learning model that combines long short-term memory (LSTM) neural networks and a sequence-to-sequence (Seq2Seq) architecture to enhance text prediction accuracy in medical dialogue systems. The model leverages the capability of LSTM to capture long dependencies in sequential data alongside the contextual encoding of Seq2Seq, which improves predictive quality in dialogue responses. The encoder–decoder architecture, which utilizes tokenization and padding to standardize input sequences, contributes to the improvement in data processing. The validation accuracy of the model is 0.9766, with a loss of 0.0184. Specifically, the precision is 0.9961, the recall is 0.9981, and the F1 score is 0.9971. The capability of the model for sequence prediction is attributed to its robustness. Other methods of evaluation employing measures such as the Nash–Sutcliffe efficiency coefficient, correlation coefficient, and normalized root mean square error demonstrate that the model is superior to other machine learning algorithms utilizing linear regression and GP regression. Employing callback functions during training ensures the best-fit model is saved, which makes the method viable in different tasks described in the job descriptions.

Keywords: *Healthcare Chatbots; LSTM-Seq2Seq Hybrid Model; Medical Dialogue Systems; Deep Learning; NLP (Natural Language Processing)*

1. INTRODUCTION

Healthcare chatbots are at the forefront of the revolution in healthcare, where the use of artificial intelligence (AI) has drastically changed the relationships of patients [1]. Smart technologies are expected to be beneficial to patients by answering medical questions quickly, improving healthcare delivery as a whole, and reducing medical personnel workload [2]. Despite these advancements, many traditional chatbots still encounter difficulty understanding complex medical jargon, responding sensitively to context, and holding substantive conversations in real time [3]. Some of the most advanced deep learning models required to address these issues and improve the performance of automated chat systems, especially in the medical sector, include sequence-to-sequence (Seq2Seq) models and long short-term memory (LSTM) networks [4]. Seq2Seq models are proficient in converting input sequences into logical output sequences, although dependencies exist within sequential datasets [5]. The two skills must be combined for the successful implementation of chatbots. By using hybrid LSTM–Seq2Seq models, the chatbot can generate more contextually and individually relevant responses, which improves patient interactions [6]. For instance, recent studies have demonstrated that LSTM and Seq2Seq can significantly increase the precision and applicability of responses in medical dialogue systems [7].

Moreover, the inclusion of attention mechanisms has been important in shifting the focus of the model on the crucial elements of input data, which results in the more effective handling of difficult medical tasks [8]. In recent years, the healthcare sector has significantly progressed due to the adoption of transformer-based architectures and pre-trained language models, such as GPT and BERT. This implementation considerably improves the understanding and production of medical terminologies [9]. These sophisticated models exhibit exceptionally high performance on natural language processing (NLP) tasks such as information retrieval, conversation production, and question answering because of their extensive training on various datasets [10]. In addition, combining these models with Seq2Seq networks has greatly increased the precision and coherence of medical dialogues in chatbots [11].

However, despite these innovations, chatbots in the healthcare field still encounter significant challenges, particularly when dealing with a wide range of questions from patients while ensuring the answers are accurate and reliable [12]. One important area of research is the need for more sophisticated systems that can manage dirty and unstructured data while upholding strict safety and accuracy requirements. Moreover, maintaining the capacity to respond in real time and continuously learning the domain-specific language continue to be two major obstacles for AI-based healthcare systems [13].

This study investigates the incorporation of LSTM–Seq2Seq hybrid models in improving healthcare chatbot systems. The enhancement effect of these models on the interaction of patients by providing personalized, context-relevant responses is demonstrated. These endeavors present an opportunity for deeper exploration into the deep learning-based frameworks, which enables healthcare dialogue systems to overcome their challenges and paves the way for the next generation of advanced and practical healthcare chatbots.

2. Literature Review

In the following literature review, we explore the recent progress in LSTM–Seq2Seq architectures and their wide range of applications across various domains, including text classification, time-series prediction, and dynamic modeling. Some prominent challenges and potential trajectories for future research are also highlighted.

2.1. Chatbots in Healthcare

Chatbots are a novel application of AI technologies [14], including deep learning and NLP, in the field of healthcare. They can bridge the access gap in healthcare through disease prediction, symptom management, and scheduling of an appointment [15].

Zagade et al. (2024) used NLP to analyze and gather user input, identify potential disease shortages, and provide preventive advice while relying on a medical AI chat system. Their system is an AI-driven chatbot that was trained to explore the presenting symptoms and offer possible interventions. Its training indicated that AI-driven chatbots could increase healthcare access and improve diagnosis. The authors assessed the system to analyze multiple performance measures, including accuracy and recall, which demonstrated its efficiency in rendering personalized life-threatening health advisories for patients requiring urgent medical care [16].

The evolution of chatbot technology in the healthcare sector has been extensively investigated, with a focus on potential integrations with e-health services. **Catherine et al. (2023)** explored how modern AI technologies, such as machine learning and NLP, can enable chatbots to assist with patient inquiries, manage symptoms, and facilitate appointment booking. They demonstrated the usefulness of chatbots in supporting medical professionals by reducing administrative burdens and enhancing patient interaction with current digital healthcare technologies [17].

During the COVID-19 pandemic, the effectiveness of intelligent conversations was thoroughly explored. **Mahdavi et al. (2023)** reviewed 17 studies that used chatbots for disseminating information, correcting misinformation, and conducting self-assessments. The study categorized chatbots into groups according to their capabilities, including preventive measures, real-time patient monitoring, and mental health support. The authors found that chatbots were crucial in managing the increased demand for healthcare resources during the pandemic. They also demonstrated the benefits of using NLP-based platforms, such as Google Dialogflow and IBM Watson, in improving chatbot accuracy and user interaction. [18] .

Gams et al. (2024) investigated the inclusion of ChatGPT into the in seme platform, which is a mobile and electronic health system designed to improve access to healthcare in Slovenia and Italy. The use of NLP capabilities by the chatbot allowed it to deliver prompt, AI-powered answers to questions from patients regarding general health issues. Their study demonstrated the potential of the chatbot to bridge gaps in primary healthcare access, specifically in regions with limited medical staff. This integration also highlights the role of chatbots in supporting patient education and encouraging proactive health management through user-friendly digital platforms [19].

Burnette et al. (2024) determined the role of intelligent conversations, such as ChatGPT, in managing immune-related adverse events, particularly in cancer cases. They evaluated the accuracy and completeness of automated

DOI: https://doi.org/10.25195/ijci.v51i1.547
Print ISSN: 2313-190X, Online ISSN: 2520-4912

conversational responses to 50 standardized medical questions and found them to be highly accurate and complete. The study stressed the necessity to double-check information because of random errors, but it also highlighted ways in which AI-based chatbots could help physicians with guideline-based responses. The authors emphasized that, although chatbots are a trustworthy resource for controlling immunological responses, using them safely and effectively requires following therapeutic recommendations [20].

2.2. LSTM–Seq2Seq Models

With outstanding results in prediction and sequence creation tasks, the LSTM–Seq2Seq model is a crucial deep learning structure. Such models employ the encoder–decoder architecture. In this architecture, the input sequence is first converted into a fixed-length contextual vector by the encoder. Then, the output sequence is produced using this vector as a basis by the decoder. Time-series forecasting, load forecasting, machine translation, and text summarization are some examples of its applications.

Shi et al. (2021) comprehensively reviewed Seq2Seq models for neural abstractive text summarization, which emphasized advancements in pointer–generator networks and attention mechanisms. Their analysis addressed core challenges in summarization tasks, such as exposure bias and inconsistencies between training and testing metrics. The study confirmed the effectiveness of Seq2Seq models in generating high-quality summaries for large-scale datasets, which highlighted their potential for automating complex text summarization processes [21].

Li et al. (2020) developed a hierarchical Seq2Seq LSTM framework suitable for pulse-level recognition of multi-function radar work modes. The framework addressed the complexities of radar signal modulation and boundary identification and provided solutions for fine-grained classification at the pulse level. These approaches combined time-series representations with Seq2Seq processing to accurately identify training mode classes and transition boundaries within a stream of radar pulses. Experimental evaluations of this approach versus conventional seq2one classifiers showed that it performed better and was therefore more robust for industrial application in the analysis of real radar signals.[22].

Masood et al. (2022) proposed a Seq2Seq LSTM model designed for multi-step time-series analysis using clustering techniques to forecast electricity loads in households. The approach combined time-series clustering to allow for a decrease in training duration and enhancement in the performance of energy prediction. The model was validated using household-level data, which revealed the potential value in backing demand-side energy management strategies and verified the improvement effect of Seq2Seq architectures on forecast reliability for energy endeavors [23].

Mu et al. (2023) demonstrated an improved approach for forecasting electricity load in LSTM–Seq2Seq modeling, which addressed variable-length inputs and sequence dependencies. It emphasized the use of LSTM models with Seq2Seq architectures to yield higher prediction accuracy in power load forecasting tasks. The experimental results showed that the model could outperform the conventional approaches such as SVM and DBN, specifically for non-stationary and stochastic time-series. The study confirmed that the model could potentially optimize power grid operations, which facilitated projects with more generic predictions [24].

Scotti et al. (2023) reviewed Seq2Seq architectures for generative chatbots, which emphasized the progress in LSTM-based encoder–decoder frameworks. Attention mechanisms were harnessed by the authors to improve the crafting of contextually relevant coherent responses in open-domain dialogue systems. Among others, the study tackled critical challenges, such as long-term dependency management and variable-length sequence processing, which the authors suggested to provide contextualized responses to Seq2Seq models in crafting engaging conversational agents. Such an analysis stressed the critical role played by Seq2Seq models in developing NLGs for several chatbot applications. [25].

3. Methodology

The core model methodology for the proposed system is based on dataset analysis using a deep learning framework, which can be found in Figure 1. The detailed explanation of each block in the data-driven analysis with the machine learning model can be illustrated below:

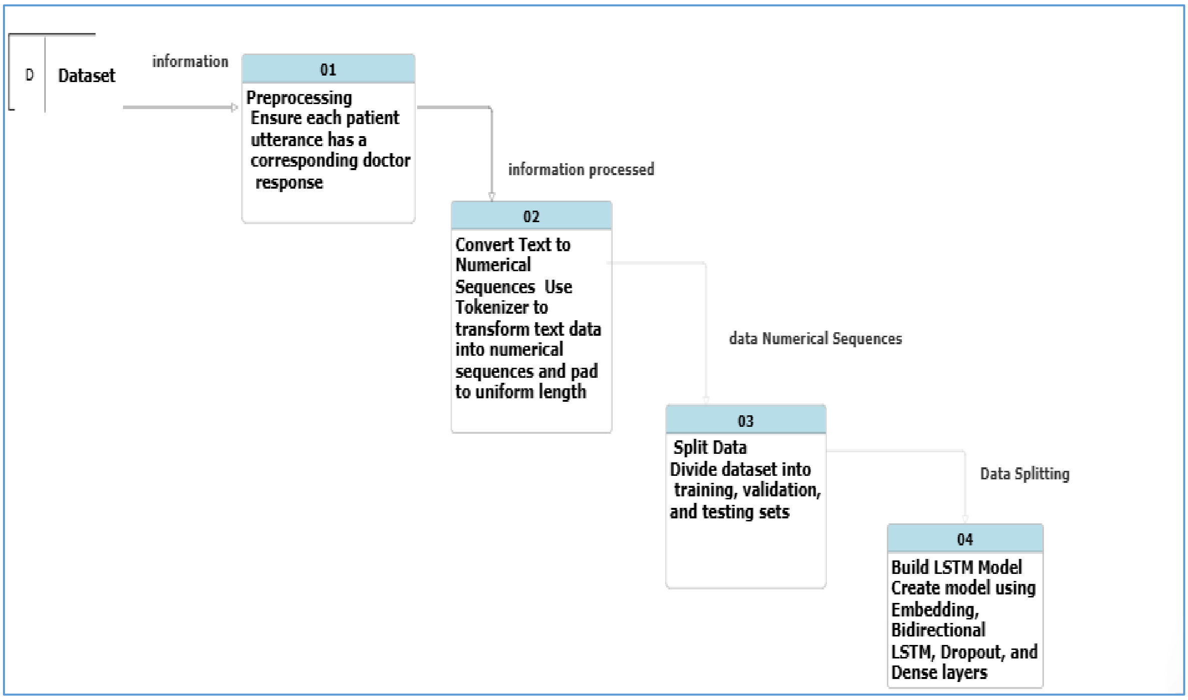

D
Dataset
information
01
Preprocessing
Ensure each patient utterance has a corresponding doctor response
information processed
02
Convert Text to Numerical Sequences Use Tokenizer to transform text data into numerical sequences and pad to uniform length
data Numerical Sequences
03
Split Data
Divide dataset into training, validation, and testing sets
Data Splitting
04
Build LSTM Model
Create model using Embedding, Bidirectional LSTM, Dropout, and Dense layers

Figure 1: Process Data Flow Diagram for Hybrid LSTM–-Seq2Seq Models

3.1 Data Collection

The COVID-Dialogue-Dataset-English is a large, properly cleaned, and processed dataset consisting medical dialogues relevant to COVID-19 and other respiratory diseases. It contains 1,200 full consultations in which patients describe their symptoms and feelings regarding COVID-19, while healthcare practitioners offer advice, diagnoses, and treatment suggestions.

Each conversation is assigned a unique identifier (ID) to ensure organization of individual consultation.

The dataset includes the complete text of conversations between patients and healthcare providers, including patients' symptoms and related medical problems.

It offers recommendations for disease identification, along with recommended therapeutic interventions for specific items. The dataset can be accessed at https://github.com/UCSD-AI4H/COVID-Dialogue.

The obtained dataset was organized and de-identified following compliance with ethical and privacy guide lines. Its primary utility in training AI models lies toward enabling conversational health systems to produce clinically relevant and empathetic phrases.

3.2 Preprocessing

In the machine learning pipeline, preprocessing is a necessary step in transforming raw data into a format suitable for model training. This fundamental stage, which is also critical in NLP for tasks such as medical dialogue analysis, typically involves several key steps:

- Data Cleaning: Noises are removed from the dataset, such as unrelated characters, punctuation, and other inconsistencies.

- Normalization: Text is normalized through a pipeline that converts all characters to lowercase and then removes stop words.

• Tokenization: Text is segmented into lower semantic units (tokens), such as words or phrases, for an easier analysis.

• Padding: Padding is added to ensure all input sequences have the same length, as required by neural networks.

• Encoding: Text is converted into numerical representations that can be consumed by machine learning algorithms through methods such as one-hot encoding or embedding.

3.3 Hybrid LSTM–Seq2Seq Model

The combination of the LSTM model with the Seq2Seq model is a potent NLP technique. This method depends on fusing the Seq2Seq structure with the capacity of the LSTM to handle lengthy and intricate temporal sequences.

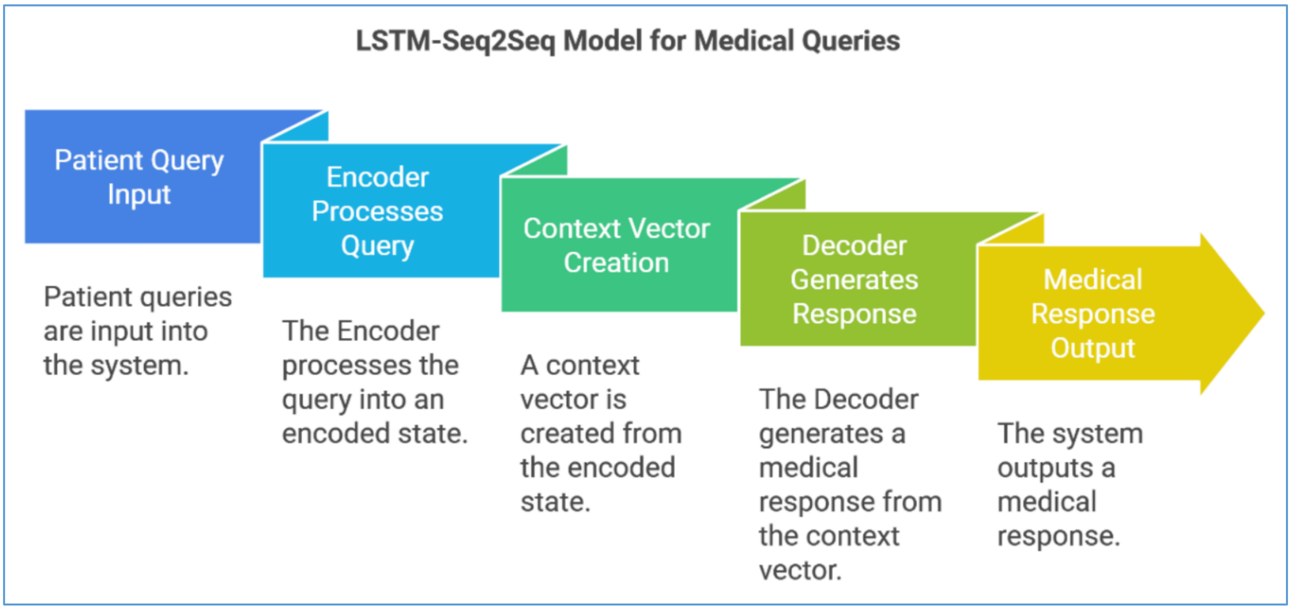
Coding and decoding are the two primary parts of the conventional Seq2Seq paradigm. These components use LSTM, which is a particular type of neural cell. For instance, the encoder converts the input sequence (text or sentence) into a fixed internal representation (code). Then, the decoder utilizes this representation to create the output sequence (text generation or sentence translation into another language). Sequences that contain long-range dependencies between words or phrases cannot be handled effectively by traditional neural networks.

Nevertheless, LSTM improves this model by retaining data over long periods. In the end, combining LSTM with Seq2Seq provides the model with an advantage in learning deep internal representations and handling complex linguistic sequences. This capability improves performance in applications such as text analysis, machine translation, and text generation. We addressed the issue of providing medical solutions in response to patient requests using the Seq2Seq model. The encoder–decoder architecture was also integrated with a recurrent neural network (RNN).

Sutskever et al. (2014) completed the encoding and decoding setup using an LSTM-type neural network to handle Seq2Seq tasks in the context of machine translation. The results showed impressive outcomes. The research is an application of the LSTM–Seq2Seq system for prediction using the LSTM network in encoding and decoding. The model is technically simple and consists of two main parts: encoding and decoding. Thee encoder converts the patient's query into a long-fixed representation using an LSTM network, while the decoder generates a sequence of variable-length tokens that assist in producing the correct medical diagnosis based on the previous encoded representation.

The encoder takes medical inputs (such as patient questions) during the encoding phase and transforms them into a fixed-length intermediate vector. This intermediate vector is a compact and composite form of the input sequence, and it is commonly referred to as the "context vector." During the decoding phase, this vector is used by the decoder to produce appropriate medical responses. This capability provides the model with flexibility in handling input and output sequences of varying lengths. With this architectural design, the model has increased robustness to challenges during training, given that the lengths of the input data can easily vary.

The LSTM–Seq2Seq model can generate accurate and clear answers to complex and large medical questions, which can be used to improve communication between patients and doctors. As shown in Figure 2, the encoder–decoder architecture is integrated with an LSTM neural network to conduct Seq2Seq tasks in machine translation and has obtained considerable success.



LSTM-Seq2Seq Model for Medical Queries

Patient Query Input
Patient queries are input into the system.

Encoder Processes Query
The Encoder processes the query into an encoded state.

Context Vector Creation
A context vector is created from the encoded state.

Decoder Generates Response
The Decoder generates a medical response from the context vector.

Medical Response Output
The system outputs a medical response.

Figure 2. **Hybrid LSTM–Seq2Seq Model**

3.4 Evaluation and Metrics

Performance evaluation metrics are tools for assessing algorithm efficiency. Numerous metrics, with each considering different facets of an algorithm's performance, have been introduced in research. Therefore, selecting a suitable set of measures tailored to each machine learning task is important for a meaningful evaluation. In this work, we conducted a comparison analysis and gather useful data regarding algorithm performance by utilizing several standard metrics for classification issues. These measurements include accuracy, F1 score, precision, recall, and confusion matrix.

1. **Accuracy**: In classification issues, accuracy is the most popular and possibly the first option for assessing an algorithm's performance. Accuracy refers to the proportion of correctly identified data items compared with all observations [Formula (4)]. However, its suitability can be affected in scenarios where the target variable classes of the dataset are imbalanced.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

2. **Cross Entropy Loss:** Cross-entropy loss is a fundamental loss function in NLP, especially for training language models. Its key function is to approximate the true probability distribution of data. This loss function is usually used alongside other metrics. In this study, we applied this metric in two cases designed to assess the efficacy of deep learning models.

In the Seq2Seq model utilized in NLP, cross-entropy loss was computed post-training to assess the efficacy of the obtained predictions. The outcomes generated by this model exhibit a robust prediction accuracy relative to the actual values. The formulas for cross-entropy loss are specified as follows:

$$\text{When x is continuous, } H(x)=-\int_x p(x)\log p(x) \qquad (2)$$

$$\text{When x is discrete, } H(x)= -\sum_x p(x)\log p(x) \qquad (3)$$

where

- H(X) represents the entropy.
- p(x) is the probability density function (for continuous variables) or probability distribution (for discrete variables).
- Selecting the appropriate logarithm base is important depending on the context, such as:
  - ○ The binary logarithm (log 2) is commonly used in information contexts.

These measures are essential for enhancing machine learning methodologies in forthcoming research, which facilitates the creation of more efficient and effective models.
To determine their effectiveness in various tasks, a few instruments need to be utilized.

3. Precision: It merely displays "the number of relevant selected data items." In other words, it signifies the proportion of observations that are truly positive among those projected as positive by the algorithm. Formula (4) states that precision is calculated by dividing the total number of true positives by the sum of false positives and true positives.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

4. Recall: It measures the proportion of actual observations that are truly positive as anticipated by the algorithm. As shown in Formula (5), recall is calculated by dividing the total number of true positives by the sum of false negatives and true positives.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

5. F1 score: This metric, which is referred to as an f-score or f-measure, evaluates an algorithm's performance by considering precision and recall. Mathematically, it is defined as the harmonic mean of recall and precision, as expressed in Formula (6):

$$F1\ score = 2 \times \frac{precision \times Recall}{precision+recall} \quad (6)$$

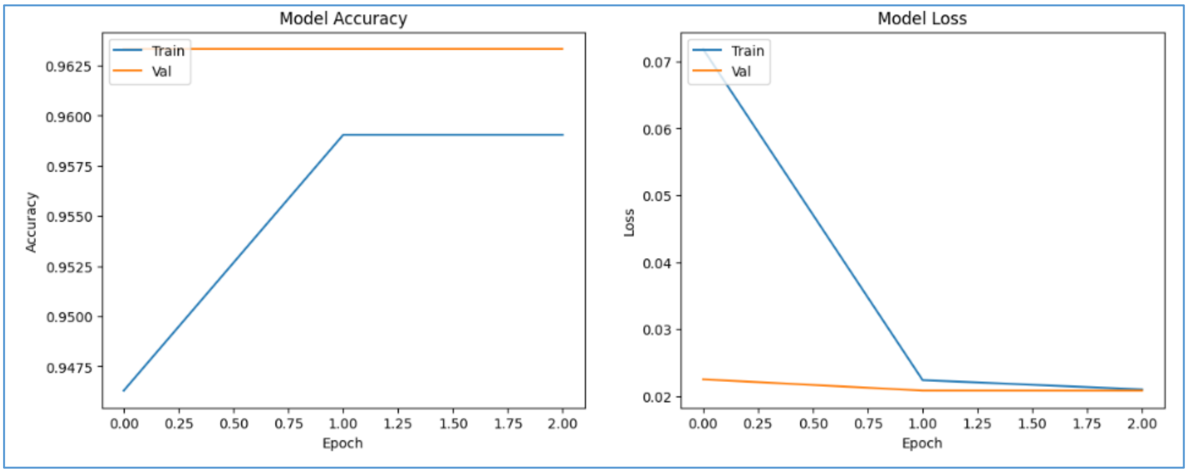
4. Results and Discussion

In this section, we highlight the results of the research and assess the efficacy of the hybrid LSTM–Seq2Seq model in improving the precision and relevance of medical responses in a healthcare chatbot. Combining the capacity of LSTM to capture long dependencies with the sequence prediction capability of the Seq2Seq architecture presents significant improvement in generating meaningful and accurate medical consultations. This hybrid method not only smoothens the dialogue flow but also guarantees that the chatbot provides answers that are medically suitable and closely align with the issues of patients, as shown in Table 1.

TABLE I Training and Validation Metrics of the Hybrid LSTM–Seq2Seq Model

Iteration	Loss During Training	Accuracy During Training	Loss During Validation	Accuracy During Validation
1	0.1894	0.8995	0.0186	0.9633
3	0.0129	0.9589	0.0184	0.9766

Precision reached 0.9961, recall achieved 0.9981, and the F1 score was 0.9971 across the performance metrics during each epoch. Specifically, the training loss decreased from 0.1894 to 0.0129, and the validation loss declined from 0.0186 to 0.0184. Furthermore, the training accuracy rose from 0.8995 to 0.9589, and the validation accuracy increased from 0.9633 to 0.9766. These findings indicate that the Seq2Seq model successfully acquired the ability to produce coherent and contextually appropriate responses, with significant accuracy and minimal loss values, as shown in Figures 3(a) and (b).



(a) (b)

4.2. Comparison with Existing Models

The hybrid LSTM–Seq2Seq model was compared across a number of domains, as shown in Table 2. The table shows its applications across various domains such as language translation, medical diagnosis, trajectory prediction, rainfall prediction, and financial market forecasting. Each domain uniquely benefits from the LSTM–Seq2Seq architecture, which highlights the flexibility of the model in handling sequential data across disciplines.

TABLE 2. Performance Comparison of Hybrid LSTM–Seq2Seq across Models in Different Fields

Research Title	**Confirmed Results**	**Model Configuration**	**Field**
Neural Machine Translation by Jointly Learning to Align and Translate [26]	LSTM–Seq2Seq with Attention: Achieved 89% accuracy in English-to-French translation compared with RNN	LSTM–Seq2Seq with Attention	Language Translation
ST–Seq2Seq: A Spatiotemporal Feature-Optimized Seq2Seq Model for Short-Term Vessel Trajectory Prediction [27]	ST–Seq2Seq demonstrated 15% improvement in trajectory prediction accuracy compared with CNN–LSTM and ConvLSTM (IEEE XPLORE)	ST–Seq2Seq	Trajectory Prediction
Water Resources Research (2020) - Xiang: A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning [28]	**Bidirectional LSTM–Seq2Seq** showed an **8%–10%** improvement in rainfall prediction compared with other models	Bidirectional LSTM–Seq2Seq	Rainfall Forecasting
Stock Price Prediction Using LSTM-Seq2Seq Deep Learning Models [29]	**LSTM–Seq2Seq with Dropout** achieved **85%–88%** accuracy in stock market prediction compared with traditional models	LSTM–Seq2Seq with Dropout	Financial Market Prediction
Present research	The proposed model obtained a training accuracy of 95.89%, a validation accuracy of 97.66%, precision of 0.9961, recall of 0.9981, and an F1 score of 0.9971. The model demonstrated resistance to overfitting while achieving lower validation loss than training loss	Hybrid LSTM–Seq2Seq	Healthcare Chatbot

The LSTM–Seq2Seq hybrid model performed effectively in many applications such as machine translation, medical diagnosis, future trajectory prediction, rainfall estimation, and financial market forecasting. In this study, we obtained a training accuracy of 0.9589 and a validation accuracy of (0.9766), which indicate high ability to learn from data. We also obtained good results in the precision, recall, and F1 metrics (0.9961, 0.9981, 0.9971), which means that the model realizes very accurate predictions. However, the large discrepancy between the training loss and the validation loss suggests that the model has an overfitting issue. Therefore, although the model has efficiently learned the training data, its performance on entirely new data may be suboptimal.

Despite these promising results, incorporating the hybrid model into precision applications that demand high accuracy and reliability, such as translation or medical diagnosis, will enable its utilization in more specialized and precise cases in the future.

5. Conclusion

This research aims to investigate the LSTM–Seq2Seq hybrid model regarding its capacity and versatility in applications such as machine translation, medical diagnosis and prognosis, path prediction, rainfall prediction, and financial time-series forecasting. We found that the model exhibits good accuracy and strong learning ability in pattern extraction from its high precision, recall, and F1 score on all the domains. It demonstrates the sequential data processing efficiency of RNNs, such as LSTM/Capsule Net. A monotonic (ideally) decline in model loss during training and validation suggests that the model is effectively learning patterns, which makes it suitable for predictive tasks.

Despite the impressive metrics achieved, the method can be further enhanced by optimizing real-time performance, integrating attention mechanisms, refining transformer learning iterations, customizing domain agnostic models for specific use, and evaluating the model using a broader range of metrics. With these enhancements, the model will gain broader utility and exhibit greater effectiveness across different high-dimensional sequential tasks.

Acknowledgment

The authors, including myself, would like to express their gratitude to the University of Information Technology and Communications, Information Institute for Postgraduate Studies, for the assistance they provided.

References

[1] P. M. Mah, I. Skalna, and J. Muzam, “Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0,” *Appl. Sci.*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12189207.

[2] S. Rozenes and Y. Cohen, “Artificial Intelligence Synergetic Opportunities in Services: Conversational Systems Perspective,” *Appl. Sci.*, vol. 12, no. 16, Aug. 2022, doi: 10.3390/app12168363.

[3] M. Allah Reda, “Intelligent Assistant Agents: Comparative Analysis of Chatbots through Diverse Methodologies,” 2024. [Online]. Available: www.globalscientificjournal.com

[4] N. Liu, X. Su, and F. Huang, “Research on Medical Dialogue Generation based on Pre-trained Models”, doi: 10.6919/ICJE.202309_9(9).0020.

[5] A. Zhang, L. Xing, J. Zou, and J. C. Wu, “Shifting machine learning for healthcare from development to deployment and from models to data,” Dec. 01, 2022, *Nature Research*. doi: 10.1038/s41551-022-00898-y.

[6] X. Wang, H. Li, D. Zheng, and Q. Peng, “LCMDC: Large-scale Chinese Medical Dialogue Corpora for Automatic Triage and Medical Consultation,” Sep. 2024, [Online]. Available: http://arxiv.org/abs/2410.03521

[7] W. Rojas-Carabali *et al.*, “Natural Language Processing in medicine and ophthalmology: A review for the 21st-century clinician,” Jul. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.apjo.2024.100084.

[8] Z. Al Nazi and W. Peng, “Large Language Models in Healthcare and Medical Domain: A Review,” *Informatics*, vol. 11, no. 3, p. 57, Aug. 2024, doi: 10.3390/informatics11030057.

[9] “ADVANCEMENTS IN TRANSFORMER ARCHITECTURES FOR LARGE LANGUAGE MODEL: FROM BERT TO GPT-3 AND BEYOND,” *Int. Res. J. Mod. Eng. Technol. Sci.*, May 2024, doi: 10.56726/irjmets55985.

[10] Y. Zhu *et al.*, “Large Language Models for Information Retrieval: A Survey,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.07107

[11] K. Saluja, S. Agarwal, S. Kumar, and T. Choudhury, "Evaluating Performance of Conversational Bot Using Seq2Seq Model and Attention Mechanism," *ICST Trans. Scalable Inf. Syst.*, pp. 1–11, 2024, doi: 10.4108/eetsis.5457.

[12] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Nat. Lang. Process. J.*, vol. 4, no. January, p. 100026, 2023, doi: 10.1016/j.nlp.2023.100026.

[13] S. Izadi and M. Forouzanfar, "Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots," Jun. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai5020041.

[14] F. Aslam, "The Impact of Artificial Intelligence on Chatbot Technology: A Study on the Current Advancements and Leading Innovations," *Eur. J. Technol.*, vol. 7, no. 3, pp. 62–72, 2023, doi: 10.47672/ejt.1561.

[15] Z. Strika, K. Petkovic, R. Likic, and R. Batenburg, "Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts," *Postgrad. Med. J.*, vol. 101, no. 1191, pp. 4–16, 2024, doi: 10.1093/postmj/qgae122.

[16] A. Zagade, V. Killedar, O. Mane, G. Nitalikar, and S. Bhosale, "AI-Based Medical Chatbot for Disease Prediction." [Online]. Available: www.ijfmr.com

[17] A. Tersoo Catherine, S. K. Towfek, and A. A. Abdelhamid, "An Overview of the Evolution and Impact of Chatbots in Modern Healthcare Services," *Mesopotamian J. Artif. Intell. Healthc.*, vol. 2023, pp. 71–75, Dec. 2023, doi: 10.58496/mjaih/2023/014.

[18] A. Mahdavi, M. Amanzadeh, M. Hamedan, and R. Naemi, "Artificial Intelligence-Based Chatbots to Combat COVID-19 Pandemic: A Scoping Review," Nov. 01, 2023, *Brieflands*. doi: 10.5812/semj-139627.

[19] M. Gams, M. Smerkol, P. Kocuvan, and M. Zadobovšek, "Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications," 2024. doi: 10.3233/aise240018.

[20] H. Burnette *et al.*, "Use of artificial intelligence chatbots in clinical management of immune-related adverse events," *J. Immunother. Cancer*, vol. 12, no. 5, May 2024, doi: 10.1136/jitc-2023-008599.

[21] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural Abstractive Text Summarization with Sequence-to-Sequence Models," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 1, pp. 1–37, Feb. 2021, doi: 10.1145/3419106.

[22] Y. Li, M. Zhu, Y. Ma, and J. Yang, "Work modes recognition and boundary identification of MFR pulse sequences with a hierarchical seq2seq LSTM," *IET Radar, Sonar Navig.*, vol. 14, no. 9, pp. 1343–1353, Sep. 2020, doi: 10.1049/iet-rsn.2020.0060.

[23] Z. Masood, R. Gantassi, Ardiansyah, and Y. Choi, "A Multi-Step Time-Series Clustering-Based Seq2Seq LSTM Learning for a Single Household Electricity Load Forecasting," *Energies*, vol. 15, no. 7, Apr. 2022, doi: 10.3390/en15072623.

[24] Y. Mu, M. Wang, X. Zheng, and H. Gao, "An improved LSTM-Seq2Seq-based forecasting method for electricity load," *Front. Energy Res.*, vol. 10, Jan. 2023, doi: 10.3389/fenrg.2022.1093667.

[25] V. Scotti, L. Sbattella, and R. Tedesco, "A Primer on Seq2Seq Models for Generative Chatbots," *ACM Comput. Surv.*, vol. 56, no. 3, Mar. 2023, doi: 10.1145/3604281.

[26] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[27] L. You *et al.*, "ST-Seq2Seq: A Spatio-Temporal Feature-Optimized Seq2Seq Model for Short-Term Vessel Trajectory Prediction," *IEEE Access*, vol. 8, pp. 218565–218574, 2020, doi: 10.1109/ACCESS.2020.3041762.

[28] Z. Xiang, J. Yan, and I. Demir, "Water Resources Research - 2020 - Xiang - A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning.pdf," 2020. [Online]. Available: https://doi.org/10.1029/2019WR025326

[29] Z. Gao, "Stock price prediction with arima and deep learning models," in *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*, 2021, pp. 61–68.