Received: 23 May 2023 | Revised: 4 December 2023 | Accepted: 3 January 2024

DOI: 10.1002/ima.23025

RESEARCH ARTICLE

WILEY

CCheXR-Attention: Clinical concept extraction and chest x-ray reports classification using modified Mogrifier and bidirectional LSTM with multihead attention

Somiya Rani[1] | Amita Jain[2] | Akshi Kumar[3] | Guang Yang[4,5,6,7]

[1]Department of Computer Science and Engineering, Netaji Subhas University of Technology East Campus (erstwhile A.I.A.C.T.R.), Guru Gobind Singh Indraprastha University, Delhi, India

[2]Department of Computer Science and Engineering, Netaji Subhas University of Technology, Delhi, India

[3]Department of Computing, Goldsmiths, University of London, London, UK

[4]Bioengineering Department and Imperial-X, Imperial College London, London, UK

[5]National Heart and Lung Institute, Imperial College London, London, UK

[6]Cardiovascular Research Centre, Royal Brompton Hospital, London, UK

[7]School of Biomedical Engineering & Imaging Sciences, London, UK

Correspondence

Guang Yang, Bioengineering Department and Imperial-X, Imperial College London, London W12 7SL, UK.
Email: g.yang@imperial.ac.uk

Funding information

Innovative Medicines Initiative, Grant/Award Number: 101005122; Medical Research Council, Grant/Award Number: MC/PC/21013; Horizon 2020 Framework Programme, Grant/Award Number: 952172; Nvidia; Royal Society, Grant/Award Number: IEC\NSFC \211235; UK Research and Innovation, Grant/Award Number: MR/V023799/1; Boehringer Ingelheim; Wellcome Leap

Abstract

Radiology reports cover different aspects from radiological observation to the diagnosis of an imaging examination, such as x-rays, magnetic resonance imaging, and computed tomography scans. Abundant patient information presented in radiology reports poses a few major challenges. First, radiology reports follow a free-text reporting format, which causes the loss of a large amount of information in unstructured text. Second, the extraction of important features from these reports is a huge bottleneck for machine learning models. These challenges are important, particularly the extraction of key features such as symptoms, comparison/priors, technique, finding, and impression because they facilitate the decision-making on patients' health. To alleviate this issue, a novel architecture CCheXR-Attention is proposed to extract the clinical features from the radiological reports and classify each report into normal and abnormal categories based on the extracted information. We have proposed a modified Mogrifier long short-term memory model and integrated a multihead attention method to extract the more relevant features. Experimental outcomes on two benchmark datasets demonstrated that the proposed model surpassed state-of-the-art models.

KEYWORDS

clinical concept extraction, clinical name entity recognition, deep learning, Mogrifier LSTM, multihead attention, natural language processing

1 | INTRODUCTION

The emergence of electronic health records (EHR) has created new prospects in the healthcare industry, and the growing use of digital content has provided numerous advantages.[1] A large amount of EHR information exists in the form of unstructured free text.[2] "The free-text reporting format tends to offer a more natural and expressive approach to documenting clinical events and facilitate communication among the care team in the healthcare environment".[3]

Medical imaging techniques such as x-rays, magnetic resonance imaging (MRI), and computed tomography (CT) scans are among the few clinical examinations used by radiologists to diagnose pulmonary diseases.[4] Radiology reports offer the primary means of documenting imaging diagnostics and communicating findings to physicians.[5] Radiology reports are specifically important for conveying valuable information; however, harnessing quality information from them is difficult because of their free-text format.[6–9]

Deep learning (DL) techniques have recently demonstrated outstanding performance in natural language processing (NLP). DL methods have been adopted for various tasks in the medical domain. Unlike machine learning (ML) and rule-based methods which require handcrafted features and manual design of rules for training purposes, DL methods automatically learn features and have stronger generalization ability. Long short-term memory (LSTM) is one of the most popular DL models predominantly used by researchers owing to its potential to capture long dependencies. Bidirectional LSTM (BiLSTM), a variant of LSTM, has forward and backward hidden layers to address sequential modeling issues. BiLSTM models have achieved impressive results for clinical-named entity recognition (CNER),[10–12] coreference resolution,[13,14] relation extraction,[15,16] and classification[17,18] for mainstream text processing tasks. Despite having a multitude of advantages over LSTM, there are some key problems with the BiLSTM model: (1) the model becomes complex due to the presence of high-dimensional input distance; (2) the model sometimes fails to capture contextual features; and (3) reduced performance due to the absence of medical words in pre-trained word embeddings.

We propose a model to address the aforementioned issues for the clinical concept extraction (CCE) and classification of chest radiographs with a modified Mogrifier and bidirectional LSTM with multihead attention (CCheXR-Att). CCheXR-Att utilizes pre-trained embeddings to generate the contextual vectors of the input words, and because the CNER can be improved by extracting the character-level information,[19] we propose generating character embeddings by adopting the self-attention method. The word representations are fed into the bidirectional Mogrifier LSTM (BiMogrifier LSTM) layer, where backward representations are computed to capture contextual information. Global and local character embeddings are fed into the conventional BiLSTM model. The outputs from both the BiMogrifier LSTM and traditional BiLSTM are concatenated, and the concatenated result is provided as input to the multihead attention (MHA) layer to capture important features. Finally, SoftMax is used to predict the final label.

The main contributions in this paper are summarized as follows:

- Propose a novel architecture, CCheXR-Att, for radiological concept extraction and classification of chest x-ray reports.
- The model integrates pre-trained embeddings and adopts a multihead self-attention method to generate character-level information.
- Propose a modification in the Mogrifier LSTM to compute backward representations for capturing contextual information.
- Employed MHA mechanism to improve the extraction of important features.
- Experimental results show that the proposed model outperforms state-of-the-art models.

The remainder of the paper is structured as follows: Section 2 discusses related work. In Section 3, we discuss the proposed methodology. The datasets, evaluation metrics, and training details are discussed in Section 4. We provide a detailed analysis of the results obtained in Section 5. Finally, the conclusion and future work are discussed in Section 6.

2 | RELATED WORK

In the field of healthcare and clinical practice, a substantial amount of text is generated, encompassing various aspects such as symptoms, test results, diagnoses, treatments, prevention, and patient outcomes. These textual data hold valuable information, and accurately identifying all the details within a clinical report can greatly assist healthcare professionals in understanding a patient's overall context during their diagnosis or treatment phase, thereby enhancing healthcare support. The potential application of clinical-concept detection and extraction within the healthcare domain is noteworthy. This involves creating systems that can extract relevant clinical information from medical narratives or data found on social media platforms.

The techniques employed in constructing CCE applications have largely been adapted from the broader realm of NLP.[20] These methodologies generally fall into two categories: rule-based approaches and statistical approaches which are further categorized into ML,[21–24] DL,[25–34] and hybrid methodologies.[35,36]

Recently, several researchers have explored this domain and achieved remarkable performance outcomes. For instance, Li et al.[37] proposed a model combining character-level CNN-BiLSTM-CRF and trained it using the Nadam algorithm, achieving an F1-score of 84.61% on the 2019 i2b2/VA concept extraction task. Gerevini et al.[38] used NLP tools for the annotation of chest CT reports and ML methods to classify them. As DL methods have shown competitive performance in different domains in recent years, many authors have used DL methods in the clinical domain. For example, Venkataraman and Pineda[39] used an LSTM RNN-based model on textual human and veterinary records and compared them with decision trees and random forests. The model scored higher than the baselines achieving macro-*F*1 scores of 74% and 68% for veterinary and human text narratives, respectively.

In addition to ML and DL methods, neural-based contextual embeddings have recently gained considerable popularity due to their superior performance compared to traditional word embeddings. Si et al.[40] explored various neural-based embeddings for extracting clinical concepts from textual narratives. Similarly, López-Ubeda et al.,[41] used transfer learning methods for the classification of Spanish radiological reports and achieved a score of up to 70% *F*1 score using a pre-trained multilingual model. For the classification of medical texts, Prabhakar and Won[42] developed a hybrid DL model with MHA, achieving an accuracy of 96.72% for the QC-LSTM model and 95.76% for the hybrid BiGRU. Olthof et al.[43] evaluated ML techniques based on NLP for categorizing Dutch radiology records of fractured extremities, chest radiography, and pneumothorax. A summary of recent studies related to clinical concept extraction is shown in Table 1.

3 | PROPOSED METHODOLOGY

This section discusses the proposed CCheXR-Att model for extracting clinical concepts from chest x-ray reports

TABLE 1 A summary of recent studies on the extraction of clinical concepts.

Reference	Objectives	Model	Dataset	Performance
Li et al.[37]	To improve and optimize a named entity recognition method based on the LSTM-CRF model.	CNN-BiLSTM-CRF	2010 i2b2/VA	*F*1 score: 84.61% Recall: 85.41%
Gerevini et al.[38]	To construct a system using annotated radiology reports to train machine learning classifiers based on the radiologist-developed schema and identified textual evidence.	BoW-based Model	Chest computed tomography reports	Accuracy: Up to 98.5% *F*-measure: Up to 98.3%
Venkataraman and Pineda[39]	To assign ICD-9 codes to clinical and veterinary records.	LSTM RNN	CSU and MIMIC-III	Macro-*F*1 score: 91% and 70%, respectively
Si et al.[40]	To enhance the contextual embeddings for clinical concept extraction.	Contextual embeddings	i2b2 (2010 and 2012) and SemEval (2014 and 2015)	*F*1-measures: 80.74%–93.18%
López-Ubeda et al.[41]	To apply transfer-learning models for the text classification of Spanish radiological reports.	Transfer-learning models	Radiological reports	*F*1-score: 70%
Prabhakar and Won[42]	To propose hybrid DL models with MHA for medical text classification.	QC-LSTM and hybrid BiGRU	Hallmark dataset and AIM dataset	Accuracy (QC-LSTM): 96.72% Accuracy (Hybrid BiGRU): 95.76%
Olthof et al.[43]	To compare different ML NLP methods for classifying radiology reports.	Rule-based, ML, BERT	Dutch Radiology reports on fractures, chest radiographs, and pneumothorax	Accuracy: Up to 96% *F*1-score: Up to 95%

Abbreviations: DL, deep learning; LSTM, long short-term memory; MHA, multihead attention; ML, machine learning; NLP, natural language processing.

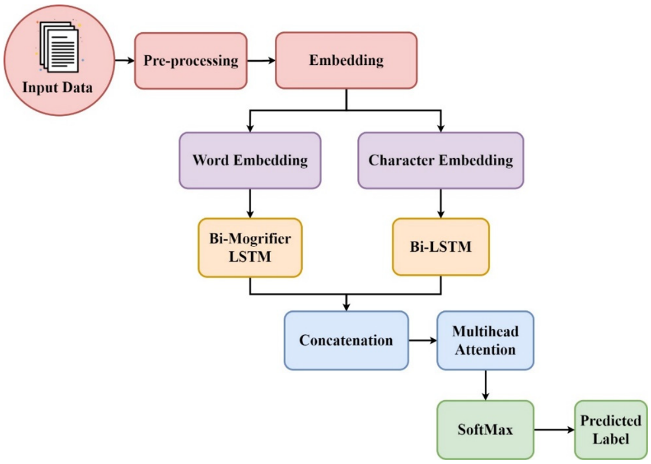


FIGURE 1 Flowchart of the proposed model.

and classifying each report into normal and abnormal categories. Figure 1 shows a flowchart of the proposed model. The input sentences are first pre-processed and fed into the embedding layers. We used two embedding layers: a word-embedding layer and a character-embedding layer. The word-embedding layer builds word embeddings for each word in a sentence using global vector embedding (GloVe) embedding.[44] In addition, the model adopts a multi-attention neural network to generate local and global character-level features. The proposed BiMogrifier LSTM layer processes word embeddings as input, whereas the standard BiLSTM layer processes character-level embeddings. The output from both layers is concatenated, and the result is fed into the MHA layer to find important features. The representations from the attention layer are then inserted into SoftMax to determine the ideal label (normal or abnormal).

3.1 | Embedding

The sentences are pre-processed and converted into a vector. Given the limitations of conventional embeddings in capturing contextual information, we opted for the utilization of pre-trained embeddings in this paper. We have used GloVe as our first embedding generation method, and the self-attention mechanism as the second method for generating local and global-level character embedding to capture more character-level features.

Prior research predominantly concentrated on word embeddings; however, it has recently been observed that character embeddings (CE) based on the self-attention method capture more information than word embeddings. As a result, we employ the self-attention mechanism to concurrently generate character embeddings at both local and global levels. The attention score is computed as

$$\alpha_i = \frac{\exp(s(x_i,q))}{\sum_{j=1}^{n} \exp(s(x_j,q))}, \tag{1}$$

where $s(x_i,q) = x_i^T \times q$, x_i represents the input status of a word or character, and the input state represented by q corresponds to x_i.

All the character representations found in a single sentence are combined into a global feature matrix for global character embedding. The BiLSTM model processes the feature matrix to incorporate more contextual information and then employs a self-attention technique to generate a new representation matrix. This results in the creation of new character-level features. The next step in obtaining the global character embeddings is to compute the average value of each character feature, followed by max pooling, which chooses the largest value of all the characters contained in a word.

Similarly, local character embedding is generated by employing a self-attention mechanism within one word. Self-attention often results in a large output dimension, hence, we construct a layer using the back-to-back pool method. Selecting a feature as word embedding with a single max-pool is not sufficient, and character information is usually lost in the first pooling layer if two max-pool layers are applied. Therefore, we employ average pooling before max pooling, which chooses the highest value as word embedding to generate character representation based on attention. “Back-to-back pool layers allow for the unification of the dimensions of each output in the character-level feature extraction layer”.[45] The proposed architecture of the CCheXR-Att model is illustrated in Figure 2.

3.2 | Bidirectional Mogrifier LSTM

In this paper, we present a modified version of the Mogrifier LSTM.[46] The motivation behind the model is to capture contextual information that traditional LSTM cannot extract efficiently. The BiLSTM model, which also processes the backward information in addition to forward information, shows a significant improvement over the LSTM model. The functions followed in LSTM are:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i), \tag{2}$$

$$F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right), \tag{3}$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o), \tag{4}$$

$$\tilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c), \tag{5}$$

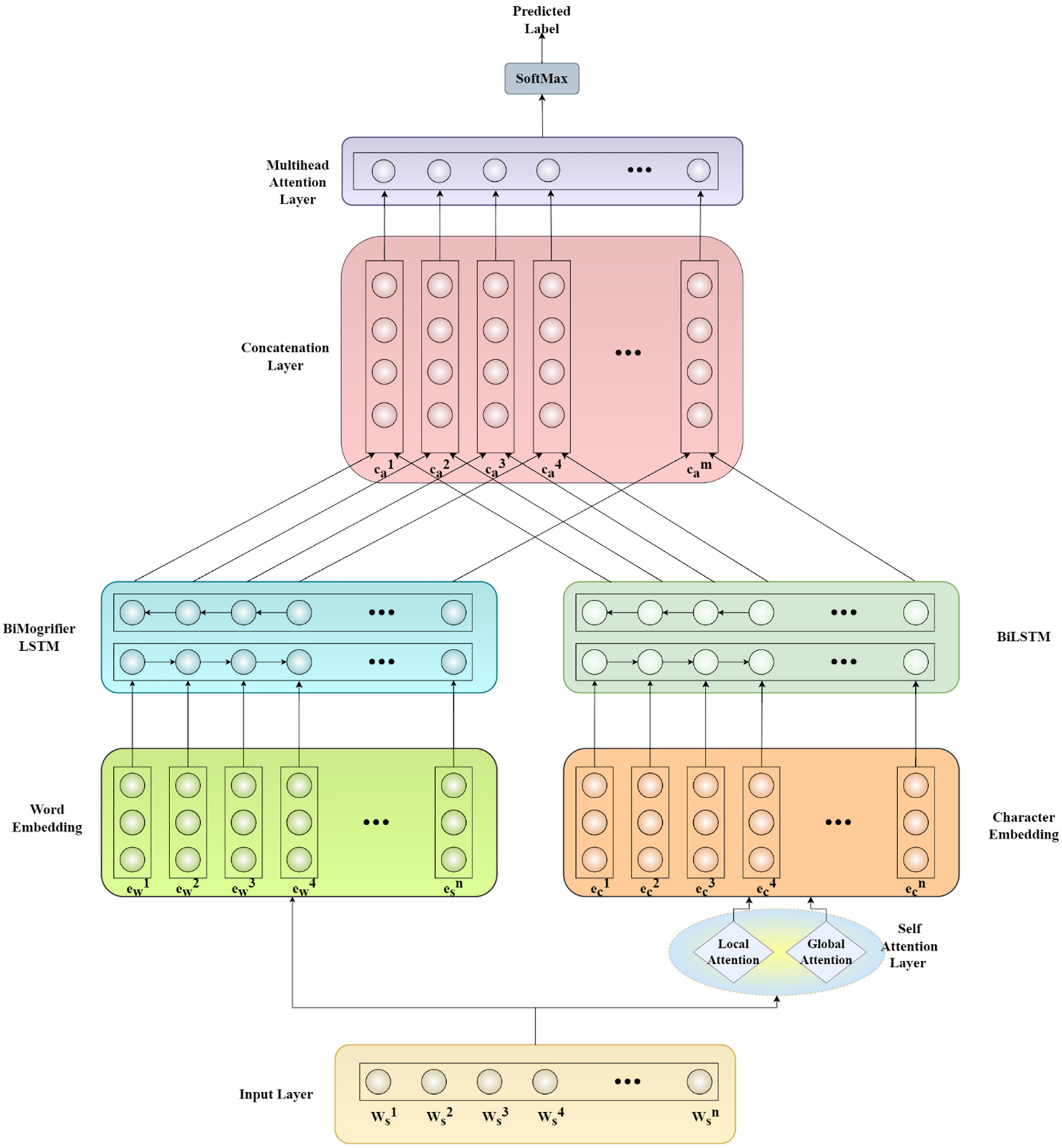


FIGURE 2 Proposed architecture of CCheXR-Attention model. CCheXR-Att, classification of chest radiographs with a modified Mogrifier and bidirectional LSTM with multihead attention.

$$P_t = P_t \bigodot P_{t-1} + I_t \bigodot \tilde{C}_t, \tag{6}$$

$$H_t = O_t \bigodot \tanh(P_t), \tag{7}$$

where $I_t \in R^{n\times h}$ denotes input gate, $F_t \in R^{n\times h}$ denotes forget gate, and $O_t \in R^{n\times h}$ denotes the output gate. X_t represents the input information. W_{xi}, W_{xf}, W_{xo}, W_{xc}, W_{hi}, W_{hf}, W_{ho}, and W_{hc} are the weight parameters, and b_i, b_f, b_o, and b_c are the bias parameters. $\tilde{C}_t$ and P_t represent the candidate memory and present memory cells respectively, and H_t represents the present hidden state. The input, forget, and output states are then computed by the LSTM cell. From the X_t and H_{t-1}, relevant information such as the input and forget states, can be calculated. The input and forget gates produce the cell state. The next LSTM cell is then loaded with hidden and cell states.

Since there is no connection between the previous state h_{prev} and the present input x in the LSTM, and with no opportunity for interaction before the gate, this absence of interaction could cause contextual information to be lost. Inspired by the Mogrifier LSTM,[46] which improves the contextual modeling by providing the interaction before the gate, we propose a bidirectional Mogrifier LSTM that not only improves the contextual modeling ability but also the concept extraction. Figure 3 shows the proposed BiMogrifier LSTM cell structure.

BiMogrifier LSTM updates the input and prior hidden states through mutual gating. The input is crossed with the gate in each cycle. The BiMogrifier LSTM model derives bidirectional hidden information $\overrightarrow{h}$ and $\overleftarrow{h}$ from the forward and backward directions, respectively. x^i representing the input embedding and h^i_{prev} representing the previous hidden states are processed by the Mogrifier operation.

$$\overrightarrow{h} = \text{Mogrifier}\left(\overrightarrow{x}, \overrightarrow{h}_{\text{prev}}\right), \tag{8}$$

$$\overleftarrow{h} = \text{Mogrifier}\left(\overleftarrow{x}, \overleftarrow{h}_{\text{prev}}\right), \tag{9}$$

$$h_{\text{BiMogrifier-LSTM}} = \left[\overrightarrow{h}, \overleftarrow{h}\right], \tag{10}$$

$$x^i = 2\sigma\left(Q^i h^{i-1}_{\text{prev}}\right) \odot x^{i-2}, \text{for odd } i \in [1 \dots r], \tag{11}$$

$$h^i_{\text{prev}} = 2\sigma\left(R^i x^{i-1}\right), \text{for even } i \in [1 \dots r], \tag{12}$$

where Q^i and R^i are matrices with randomly initialized values, and r represents the number of rounds. σ represents a logistic sigmoid function and $\odot$ represents an element-wise product.

3.3 | Bidirectional LSTM

The basic BiLSTM network is used as the second layer to process the character-level embeddings. Local and global character-level embeddings are fed into the BiLSTM network. The forward and backward hidden representations $\overrightarrow{h}$ and $\overleftarrow{h}$ for BiLSTM are computed using Equations (13) and (14), respectively. The final output hidden representation h_{BiLSTM} is obtained by concatenating $\overrightarrow{h}$ and $\overleftarrow{h}$ as given in Equation (15).

$$\overrightarrow{h} = \text{LSTM}\left[\overrightarrow{h_1}, \dots \overrightarrow{h_n}\right]. \tag{13}$$

$$\overleftarrow{h} = \text{LSTM}\left[\overleftarrow{h_1}, \dots \overleftarrow{h_n}\right]. \tag{14}$$

$$h_{\text{BiLSTM}} = [h_1, \dots, h_n]. \tag{15}$$

The contextual representations obtained from the BiMogrifier LSTM and the traditional BiLSTM are then fed into the concatenation layer.

3.4 | Concatenation layer

This layer performs element-wise summations of the representations from the previous layers.[47] The concatenation of two independent networks is represented by h_{Final} as shown in Equation (16).

$$h_{\text{Final}} = \left[h_{\text{BiMogrifier-LSTM}} + h_{\text{BiLSTM}}\right]. \tag{16}$$

3.5 | Multihead attention

Clinical-named entities are not present in isolation in clinical reports and hold dependencies among them,

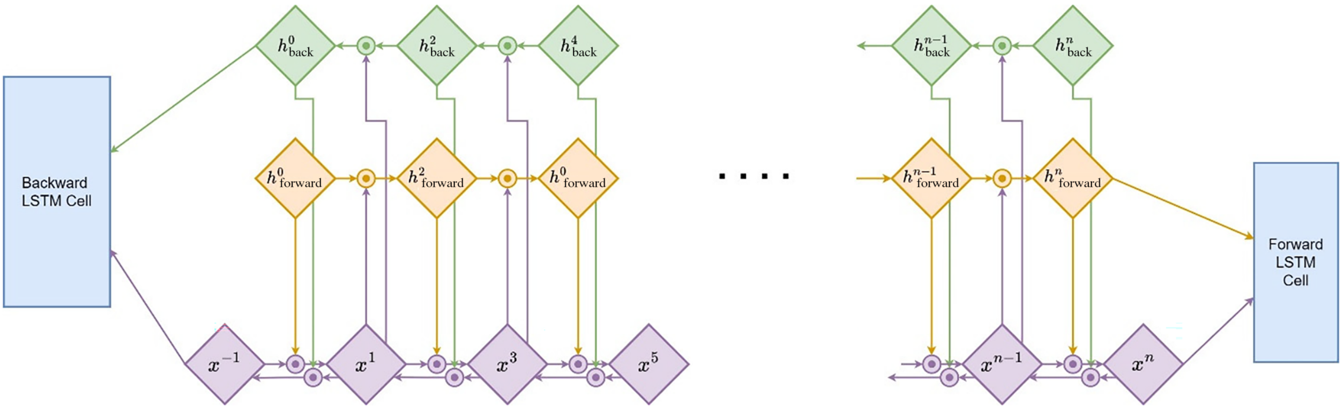


FIGURE 3 Proposed Bidirectional Mogrifier long short-term memory cell structure.

accompanied by a long interval between entity characters. Given the significance of this dependency, it must be captured by assigning dependent characters more attention by assigning them higher weights for significant characters and lower weights for less important characters. To capture this dependency, the model needs to give extra importance to the characters that are dependent on the current character.

We used an MHA method to locate important features. The MHA structure is shown in Figure 4. Attention scores are computed using Equation (17).

$$\text{attention}(Q,K,V)=\text{SoftMax}\left(\frac{QK^t}{\sqrt{d}}\right)V. \tag{17}$$

The MHA employs parallel h heads to concentrate on various components of the value vector channels. The $Q,K,\text{and}\,V$ parameters represent the characters in the sentence and are set to be equal while calculating self-attention. The learning parameters are defined as $W_i{}^Q \in \mathbb{R}^{n\times\frac{d}{h}}, W_i{}^K \in \mathbb{R}^{n\times\frac{d}{h}}$, and $W_i{}^V \in \mathbb{R}^{n\times\frac{d}{h}}$. The ith head attention is calculated using Equation (18).

$$M_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right). \tag{18}$$

The computation output from these parameters is concatenated h times, a linear transformation is performed, and the output of the phrase's tth character is obtained using Equation (19), where concat() denotes the splicing function and $W^0 \in \mathbb{R}^{n\times\frac{d}{h}}$ is the weight parameter.

$$M_t=\text{concat}(h_1,h_2,h_3,\ldots,h_h)W^0. \tag{19}$$

3.6 | SoftMax

The next layer that we implemented after the MHA layer is the SoftMax layer to decode the predicted labels. The evaluation score is computed using Equation (20).

$$S(X,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n}W_{y_i y_{i+1}}, \tag{20}$$

where X represents the input sequence, y represents the corresponding label sequence, $P_{i,j}$ represents the score of the ith character labeled as label j, W represents the transition matrix, and $W_{i,j}$ is the state transition score from label i to j.

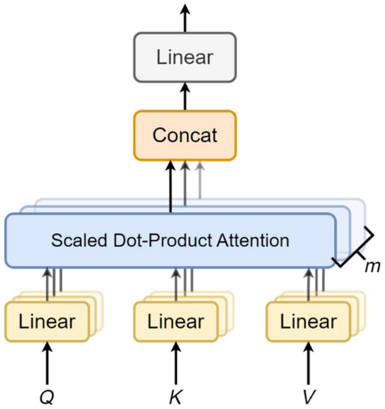


FIGURE 4 Multihead attention structure.

SoftMax is used to compute the conditional probability of the sequence label y given X, using Equation (21).

$$P(y|X)=\frac{e^{S(X,y)}}{\sum_{\tilde{y}\in Y_x} e^{S(X,\tilde{y})}}. \tag{21}$$

4 | EXPERIMENT

This section gives an overview of datasets, evaluation metrics, and the training process. We also evaluated the effectiveness of several baseline models and extensively analyzed both the proposed model and its different variants. In addition, we thoroughly investigated the model's performance through an ablation study to gain deeper insights into the proposed model.

4.1 | Dataset

We employed two standard benchmark datasets in our study. The first dataset, known as the Indiana University Chest X-ray Reports (IU-CXR) dataset, comprises 3955 radiology reports and is sourced from the National Library of Medicine.[48] The second dataset, called MIMIC-CXR,[49] encompasses 377 110 images linked to 227 827 textual reports. For our research, we focused solely on the textual reports within this dataset.

4.2 | Evaluation metrics

We assess the performance of the proposed approach through various evaluation metrics: accuracy, $F1$-score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). These metrics are computed using the following equations:

$$\text{Accuracy} = \frac{T_{\text{Positive}} + T_{\text{Negative}}}{T_{\text{Positive}} + F_{\text{Negative}} + T_{\text{Negative}} + F_{\text{Positive}}}. \quad (22)$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (23)$$

$$\text{Sensitivity} = \frac{T_{\text{Positive}}}{T_{\text{Positive}} + F_{\text{Negative}}}. \quad (24)$$

$$\text{Specificity} = \frac{T_{\text{Negative}}}{T_{\text{Negative}} + F_{\text{Positive}}}. \quad (25)$$

$$\text{PPV} = \frac{T_{\text{Positive}}}{T_{\text{Positive}} + F_{\text{Positive}}}. \quad (26)$$

$$\text{NPV} = \frac{T_{\text{Negative}}}{T_{\text{Negative}} + F_{\text{Negative}}}. \quad (27)$$

We have also used the Matthews Correlation Coefficient (MCC), a statistical tool for evaluating model performance that considers T_{Positive}, T_{Negative}, F_{Positive}, and F_{Negative}, thus making it a balanced measure of classification performance. This is particularly important in CCE, where the goal is to identify all relevant concepts accurately without missing any or generating F_{Positive}. The MCC is computed using Equation (28).

$$\text{MCC} = \frac{(T_{\text{Positive}} * T_{\text{Negative}}) - (F_{\text{Positive}} * F_{\text{Negative}})}{\sqrt{(T_{\text{Positive}} + F_{\text{Positive}})(T_{\text{Positive}} + F_{\text{Negative}})(T_{\text{Negative}} + F_{\text{Positive}})(T_{\text{Negative}} + F_{\text{Negative}})}}. \quad (28)$$

4.3 | Training details

The assessment of all models is done on a Windows machine with an NVIDIA RTX3070 GPU. The word ‘embedding’ originated as a 100-dimensional GloVe pre-trained representation sourced from Wikipedia domain text. Character embedding was configured at a dimension of 40. ADAM optimizer with a learning rate of 0.0001 is employed. Each iteration was based on a batch size of 16, and the training spanned 40 epochs. Dropout regularization with a rate of 0.5 was employed to enhance the model's stability.

5 | RESULTS AND DISCUSSION

This section provides a detailed analysis of the results obtained from the proposed model. A comparative analysis is conducted between the proposed model and traditional as well as state-of-the-art models to assess its efficacy.

5.1 | Baselines

We evaluated the performance of the CCheXR-Att model in comparison with several baseline and state-of-the-art models, including LSTM, BiLSTM, CNN-BiLSTM-CRF,[37] BoW-based model,[38] FasTag,[39] $\text{BERT}_{\text{LARGE}}$,[40] Hybrid BiGRU,[42] Fine-Tuned BERT,[43] Mogrifier LSTM,[46] and MSAM.[50]

LSTM relies on a single LSTM network to represent the sentence. The final sentence representation is derived by averaging all hidden states. BiLSTM creates representations in both the forward and backward directions, enabling it to capture the semantic meanings of a sentence from both directions. CNN-BiLSTM-CRF[37] receives pre-processed text and offers two choices for character-level representation generation: CNN and BiLSTM. The BiLSTM method captures neighboring information, with forward LSTM capturing the left context and backward LSTM capturing the right context. The final labeling, influenced by local label dependencies, is optimized using the CRF layer to assign the most suitable tag to each word. BoW-based model[38] relies on the bag-of-words framework wherein the complete textual content of the report is employed for classification purposes. This method omits the use of manually curated annotated datasets and automated annotation tools. FasTag[39] involves sequential embeddings of terms abstracted from medical narratives. By utilizing the GloVe technique, these terms undergo compact encoding, enabling the portrayal of a vector space in which semantically similar terms exhibit close associations. $\text{BERT}_{\text{LARGE}}$[40] is improved by adding extra BiLSTM layers atop its architecture, making it deeper and more complex. $\text{BERT}_{\text{LARGE}}$ is fine-tuned by replacing the CRF layer with a BiLSTM architecture due to BERT's effective sequence labeling. Hybrid BiGRU[42] consists of a CNN for extracting local features and a BiGRU along with an MHA mechanism to model the semantic features to enhance the overall effectiveness of the model. Fine-Tuned BERT[43] understands the relationships within single words and complete sentences. BERT is initialized with pre-trained parameters and then optimizes all parameters using labeled data for the CCE task. Mogrifier LSTM[46] introduces a “Mogrifier” update, a gating mechanism that enhances LSTM networks by integrating information from different time steps, improving their ability to capture complex dependencies in sequential data. MSAM[50] employs self-attention mechanisms with MHA heads to capture temporal dependencies and patterns effectively.

TABLE 2 Performance comparison of traditional, state-of-the-art, and proposed models on the IU-CXR dataset.

Model	Accuracy	*F*1-score	Sensitivity	Specificity	PPV	NPV	MCC
LSTM	82.19	81.78	80.61	77.54	79.07	81.56	72.38
BiLSTM	85.36	84.12	79.53	76.03	81.91	83.48	75.41
CNN-BiLSTM-CRF[37]	86.44	82.62	81.31	80.42	83.27	87.17	73.55
BoW-based model[38]	83.43	80.74	79.25	78.47	81.59	85.38	72.42
FasTag[39]	84.62	83.41	81.53	79.82	81.72	86.75	73.86
$\text{BERT}_{\text{LARGE}}$[40]	86.91	84.38	80.42	77.43	83.64	87.29	76.34
Hybrid BiGRU[42]	85.35	82.63	80.26	78.29	84.64	87.85	74.64
Fine-Tuned BERT[43]	86.71	83.24	81.36	79.36	85.62	87.59	75.35
Mogrifier LSTM[46]	85.59	84.93	82.93	78.36	84.60	85.74	74.76
MSAM[50]	84.6	82.4	79.91	75.42	80.7	84.13	75.23
LSTM-Att	85.12	85.9	81.81	78.49	84.3	87.17	74.68
BiLSTM-Att	88.93	87.69	83.01	78.83	85.16	88.53	80.35
Mogrifier LSTM-Att	86.55	85.4	80.32	79.45	82.87	86.41	78.59
BiMogrifier LSTM-Att	90.73	91.99	84.81	80.73	89.57	90.38	81.27
CCheXR-Att	92.89	89.31	88.23	81.51	87.75	90.45	88.62

Abbreviations: LSTM, long short-term memory; NPV, negative predictive value; MCC, Matthews Correlation Coefficient; PPV, positive predictive value.

This allows the model to focus on the most relevant features, enabling the accurate classification of radiological concepts.

5.2 | Performance comparison with traditional and state-of-the-art models

We compared CCheXR-Att with traditional and state-of-the-art models to prove its effectiveness. Performance achieved by various models on the IU-CXR and MIMIC-CXR datasets is presented in Tables 2 and 3, respectively.

It is observed that LSTM achieved the lowest accuracy of 82.19% on IU-CXR, and an accuracy of 80.65% on the MIMIC-CXR dataset. One reason for the poor performance of LSTM is that it is unidirectional and processes each word in a sentence equally. BiLSTM exhibited a slightly improved performance compared with LSTM on both datasets. In comparison with conventional LSTM and BiLSTM models, CNN-BiLSTM-CRF exhibited superior performance. The inclusion of both CNN and BiLSTM architectures combined with CRF contributed to an enhanced performance level on both datasets. The BoW-based model achieved slightly reduced performance as it neglects word order and syntactic information. The existence of misspellings, abbreviations, and medical terminologies influenced the performance of FasTag. $\text{BERT}_{\text{LARGE}}$ showcased great performance on both datasets. This is due to its extensive pre-trained knowledge, providing an inherent benefit in comprehending specialized terminology and patterns within clinical reports.

The hybrid BiGRU model also achieved comparable performance with fine-tuned BERT. The performance comparison on the IU-CXR dataset indicates that the models consistently achieved relatively similar accuracies with the Hybrid BiGRU model. By contrast, the fine-tuned BERT model demonstrated the highest accuracy as it involves training BERT on labeled data, allowing it to learn task-specific features, relationships, and intricacies present in the clinical reports. Notably, the Mogrifier LSTM demonstrated a substantial accuracy of 85.59% on the IU-CXR dataset and an accuracy of 81.43% on the MIMIC-CXR dataset. MSAM which incorporates self-attention mechanisms, exhibited an accuracy of 84.6% on the IU-CXR dataset and 85.91% on the MIMIC-CXR dataset. MSAM showed a better performance than the baseline owing to the self-attention method adopted by the model to give extra weight to important entities.

Remarkably, the CCheXR-Att model emerged as a standout performer across both datasets. On the IU-CXR dataset, it attained the highest accuracy of 92.89%, an *F*1-score of 89.31%, sensitivity of 88.23%, a specificity of 81.51%, PPV of 87.75%, NPV of 90.45%, and MCC of 88.62%. Similarly, on the MIMIC-CXR dataset, CCheXR-Att achieved an accuracy of 93.58%, an *F*1-score of 92.03%, sensitivity of 88.32%, specificity of 81.49%, PPV of 90.23%, NPV of 92.64%, and MCC of 82.72%. This impressive performance can be attributed to the incorporation of word and character embeddings in the construction of a bidirectional model of the basic Mogrifier LSTM. BiMogrifier LSTM-Att also displayed favorable outcomes, achieving an accuracy of 90.73% on the IU-

TABLE 3 Performance comparison of traditional, state-of-the-art, and proposed models on the MIMIC-CXR dataset.

Model	Accuracy	*F*1-score	Sensitivity	Specificity	PPV	NPV	MCC
LSTM	80.65	79.19	79.38	76.12	78.92	81.72	71.61
BiLSTM	84.89	83.92	80.82	78.87	82.67	85.58	72.48
CNN-BiLSTM-CRF[37]	87.22	83.18	82.53	80.31	84.61	88.32	75.37
BoW-based Model[38]	84.64	81.36	80.83	79.46	83.52	85.23	74.15
FasTag[39]	86.42	84.23	84.05	83.75	84.13	87.76	75.44
BERT$_{\text{LARGE}}$[40]	87.84	85.61	82.32	78.92	85.75	89.43	77.82
Hybrid BiGRU[42]	87.83	83.52	82.36	81.27	83.41	88.28	76.26
Fined-Tuned BERT[43]	88.57	84.64	82.23	80.41	84.34	87.73	76.35
Mogrifier LSTM[46]	81.43	79.01	80.50	80.30	80.34	85.12	70.62
MSAM[50]	85.91	85.28	82.91	80.73	84.78	88.39	73.57
LSTM-Att	83.42	82.39	80.5	79.59	81.61	85.18	74.21
BiLSTM-Att	90.06	89.73	86.69	85.25	87.21	88.45	76.43
Mogrifier LSTM-Att	84.11	82.1	82.33	78.72	80.46	89.44	75.63
BiMogrifier LSTM-Att	89.6	88.97	87.12	81.68	85.93	89.92	80.58
CCheXR-Att	93.58	92.03	88.32	81.49	90.23	92.64	82.72

Abbreviations: CCheXR-Att, classification of chest radiographs with a modified Mogrifier and bidirectional LSTM with multihead attention; LSTM, long short-term memory; NPV, negative predictive value; MCC, Matthews Correlation Coefficient; PPV, positive predictive value.

CXR dataset and 89.6% on the MIMIC-CXR dataset. The Boxplots of performance metrics for the CCheXR-Att and its variants on both datasets are shown in Figure 5. The training and validation accuracy comparisons per epoch for different variants of the CCheXR-Att on both datasets are shown in Figure 6.

5.3 | Proposed model analysis

We examined different variations of CCheXR-Att to see how well they work. The variants include LSTM-Att, BiLSTM-Att, Mogrifier LSTM-Att, and BiMogrifier LSTM-Att.

The LSTM-Att model utilizes separate LSTM layers for word and character-level representation learning. The representations obtained from both LSTM layers are concatenated at the concatenation layer followed by MHA to learn relevant position-specific information. SoftMax classifier is then used to predict the final output. By contrast, the BiLSTM-Att variant replaces both LSTM layers with a bidirectional LSTM while keeping other layers consistent. Conversely, the Mogrifier LSTM-Att introduces a Mogrifier LSTM layer in place of BiLSTM in the previous variant, to learn word and character-level representation. The rest of the architecture remains the same. Similarly, the BiMogrifier LSTM-Att employs two bidirectional Mogrifier LSTM layers among which one layer learns the word and the other layer learns character-level embeddings, followed by concatenation, MHA, and finally a SoftMax layer to predict the class. Alternatively, the CCheXR-Att model considers bidirectional Mogrifier LSTM and bidirectional LSTM layers for word and character-level representations, respectively. The output from both layers is then concatenated at the concatenation layer and is followed by the MHA layer. A SoftMax classifier is then applied for class prediction, as elaborated in Section 3 of the paper.

We have also analyzed different parameter values, including learning rate and dropout rates, which affected model performance. Among the tested learning rates, 0.0001 stood out as optimal, delivering the highest accuracy and *F*1-score for IU-CXR and MIMIC-CXR datasets. Various dropout rates were tested, and optimal outcomes emerged as dropout rates increased incrementally from 0.2 to 0.5. The best accuracy and *F*1-score occurred at a 0.5 dropout rate for both IU-CXR and MIMIC-CXR datasets. Specifically, IU-CXR achieved 90.17% accuracy and 84.97% *F*1-score, while MIMIC-CXR attained 91.62% accuracy and 90.65% *F*1-score.

We explored local and global self-attention methods using various head values (20, 40, 60, and 80) for the CCheXR-Att. Results from the experiments revealed that, for the local character-level self-attention approach, the highest accuracy occurred at 80 attention heads, and the highest F1-score emerged at 60 attention heads for the IU-CXR dataset. Similarly, on the MIMIC-CXR dataset, the accuracy occurred at 60 attention heads and the *F*1-score at 80 attention heads. On analyzing various head values with the global self-attention method, we

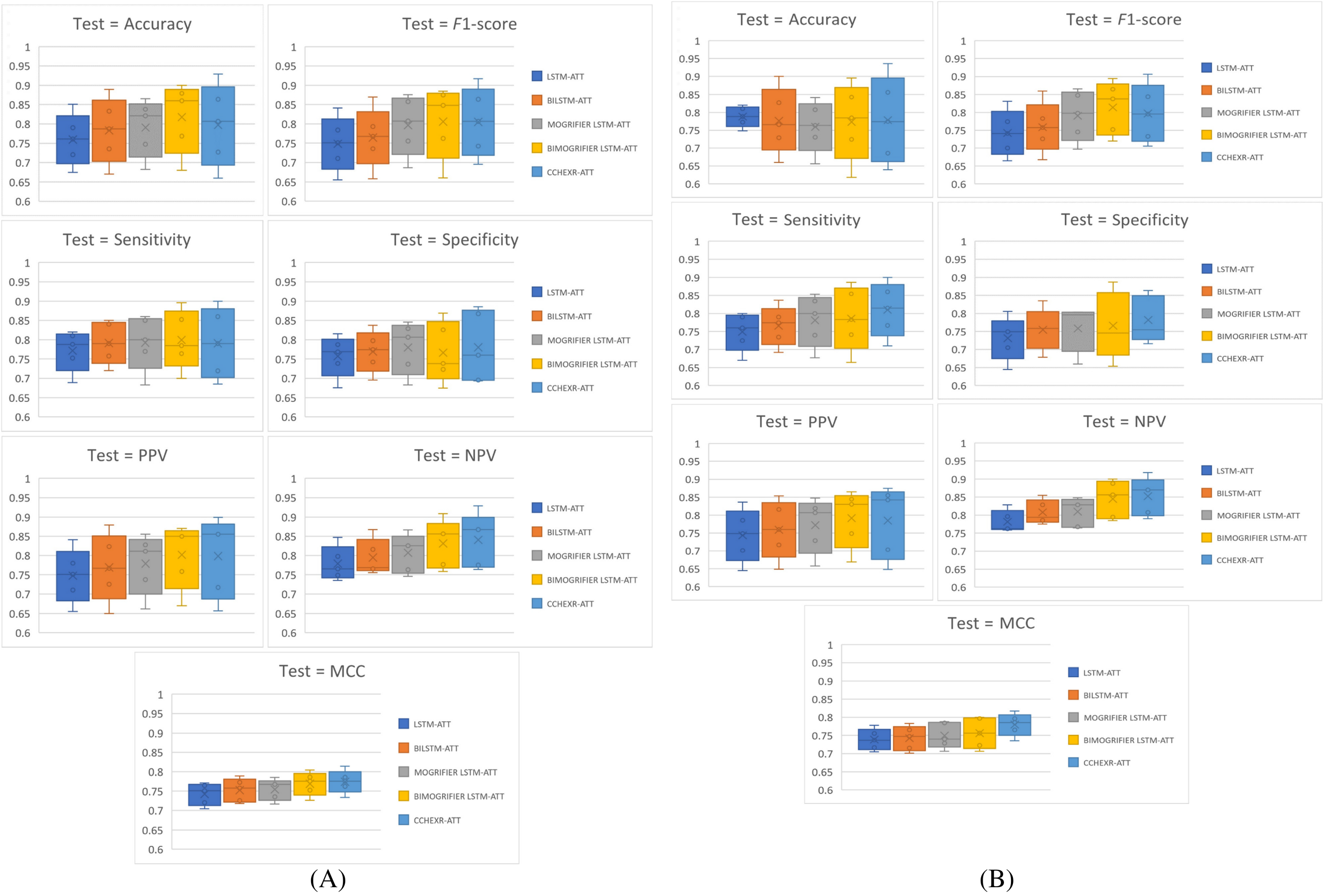


FIGURE 5 Boxplot of performance metrics for the proposed model and its variants on the datasets: (A) IU-CXR, and (B) MIMIC-CXR. IU-CXR, Indiana University Chest X-ray Reports.

observed that CCheXR-Att achieved its highest accuracy at a head value of 40, while the *F*1-score reached its maximum at 60 for the IU-CXR dataset. By contrast, the highest accuracy and *F*1-score for the MIMIC-CXR dataset were observed with an 80 head value.

5.4 | Ablation study

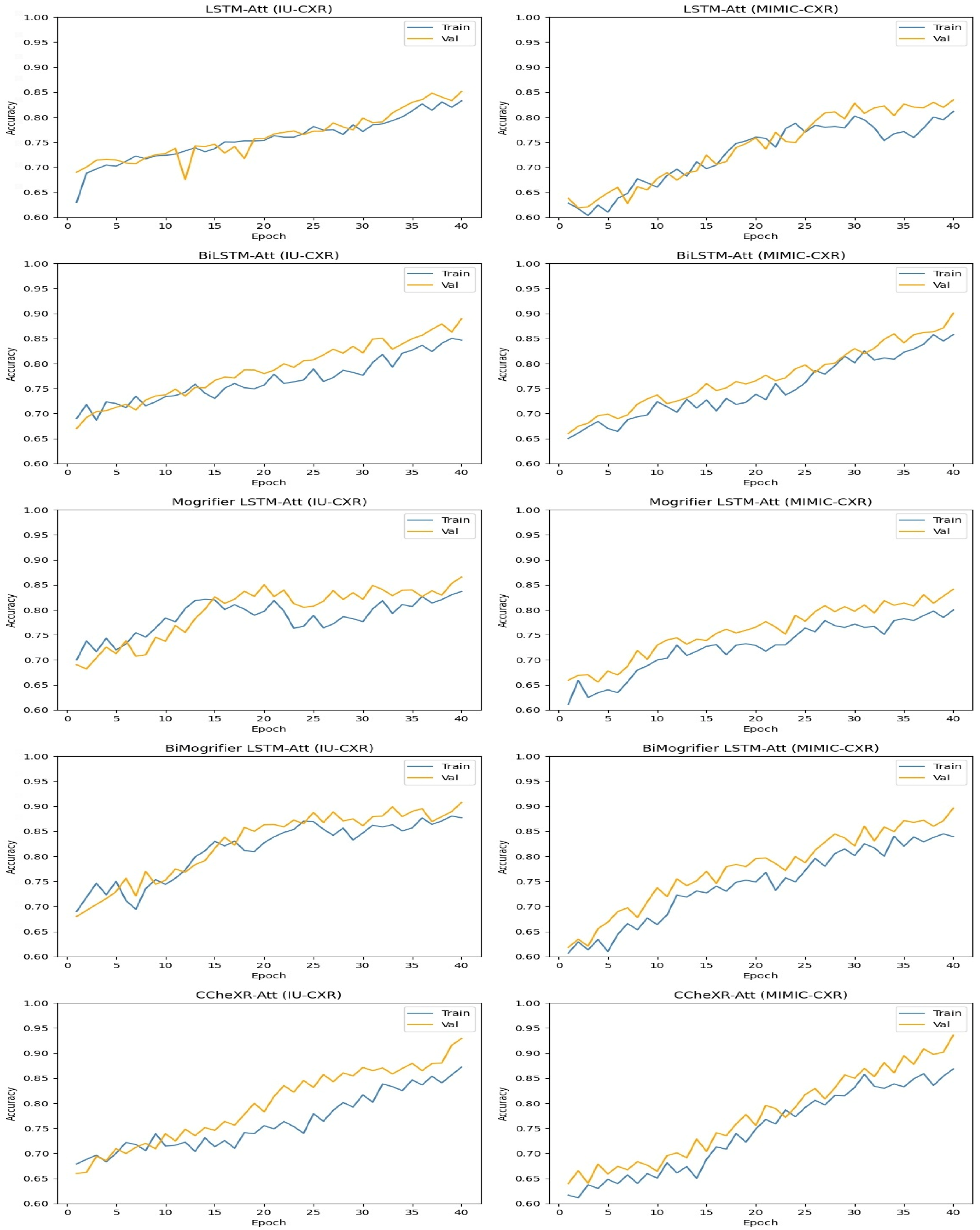
To comprehensively assess the individual contributions of the components comprising CCheXR-Att, we conducted an ablation study using the IU-CXR and MIMIC-XR datasets. The results achieved by our model in the absence of various proposed modules are presented in Table 4.

5.4.1 | Removing self-attention-based character-embedding module

Utilizing CE through the self-attention technique proves to be more informative than word embeddings. The findings in Table 4 show the relatively reduced performance of the model in the absence of CE_s, thus affirming their efficacy. This highlights the valuable role of character embeddings at both local and global levels in helping the model understand detailed character-level features. The model's performance decline without CE_s indicates that the omission of such embeddings introduces contextual information gaps, potentially introducing bias and consequent performance deterioration.

5.4.2 | Removing BiMogrifier LSTM module

The incorporation of the BiMogrifier LSTM introduces a dynamic updating mechanism that involves mutual gating, enabling contextual modeling by facilitating interaction preceding the gating process. Our results show a significant performance drop on both IU-CXR and MIMIC-CXR datasets when the BiMogrifier LSTM module is removed. This confirms the essential nature of the dynamic updating mechanism and strongly suggests that the including of the BiMogrifier LSTM module greatly improves concept extraction and classification.



(A) (B)

FIGURE 6 Training and validation accuracy throughout 40 epochs for the proposed model and its variants on datasets: (A) IU-CXR, and (B) MIMIC-CXR. IU-CXR, Indiana University Chest X-ray Reports.

TABLE 4 The performance comparison of our model without various proposed modules on the IU-Xray and MIMIC-CXR datasets.

	IU-CXR		MIMIC-CXR	
Model	Accuracy	*F*1-score	Accuracy	*F*1-score
Without CE	80.07	72.94	82.37	79.21
Without BiMogrifier LSTM	82.53	77.23	85.63	80.76
Without MHA	84.42	80.77	84.11	81.54
CCheXR-Att	92.89	89.31	93.58	92.03

Abbreviations: CCheXR-Att, classification of chest radiographs with a modified Mogrifier and bidirectional LSTM with multihead attention; CE, character embeddings; IU-CXR, Indiana University Chest X-ray Reports; LSTM, long short-term memory; MHA, multihead attention.

5.4.3 | Removing MHA module

The integration of an MHA mechanism that extracts important insights from different parts of a sentence and gives priority to relevant elements greatly benefits classification results. Our experimental findings confirm that incorporating this MHA mechanism significantly improves the effectiveness of the proposed model. Conversely, the lack of such a mechanism suggests that the model struggles to capture essential semantic information that spans various representation dimensions and positions, resulting in a noticeable decline in performance.

5.5 | Discussion and findings

Applying the proposed model to chest x-ray reports offers comparatively simple and low-effort means to overcome the limitations discussed in this paper. Traditional ML and DL models are not able to effectively capture contextual features, and the presence of a high-dimensional input distance causes complexity in the models. Another reason for the poor performance of traditional models is the absence of medical words in pre-trained word embeddings.

To this end, the proposed model is useful given its ability to utilize word and character embedding by integrating the local and global-level attention methods, and further enhancing the Mogrifier LSTM to include information from the backward direction through the gating mechanism. Additionally, the model is improved through an MHA method to capture important information obtained from the concatenation layer.

In addressing complex CCE tasks, we find the combination of word and character embeddings to be highly effective. Word embeddings capture comprehensive word information, while character embeddings excel in handling out-of-vocabulary (OOV) words, collectively enhancing our model's CCE performance. CCheXR-Att utilizes both BiMogrifier LSTM and BiLSTM to successfully extract meaningful features. CCheXR-Att exhibited higher accuracy compared with BiMogrifier LSTM-Att. This variation in accuracy between CCheXR-Att and BiMogrifier-Att is attributed to distinctions in clinical narratives, the presence of rare words, and dataset size. By leveraging both word and character embeddings, our model adeptly captures word relationships, leading to improved concept extraction. It excels in annotating entities of varying lengths, demonstrating superior information extraction capabilities. While BiLSTMs are particularly effective at learning complex features and patterns, the introduction of Bi-Mogrifier LSTM enhances contextual information identification, thereby boosting the overall performance of CCheXR-Att. The combination of architectural elements and embedding techniques positions CCheXR-Att as a robust solution for advancing CCE tasks, showcasing its proficiency in capturing contextual information and achieving superior performance across diverse datasets.

6 | CONCLUSION AND FUTURE WORK

This paper presents CCheXR-Att, a novel approach for the clinical concept extraction and classification of chest x-ray reports. Specifically, the model integrates pre-trained word embeddings and character-level embeddings based on the self-attention method that is further processed with proposed Bi-Mogrifier LSTM and BiLSTM, respectively.

The proposed model aims to fetch useful entities from clinical narratives and provide support to healthcare professionals, radiologists, and researchers to make better decisions through the detected information, and to increase patients' quality of life. In addition, the experiment demonstrated the effectiveness of CCheXR-Att which suggests that the framework with different components introduced in the model can capture accurate information and classify the reports correctly. In the two benchmark datasets, the proposed model performs better than the state-of-the-art models.

In the future, an investigation into the efficacy of incorporating state-of-the-art language models will be undertaken. Furthermore, a comprehensive assessment and comparison of alternative neural architectures for clinical concept extraction and classification will be conducted. It is also planned to include external domain-specific knowledge in future implementation. The extensibility of the model to encompass multiple languages is also a prospective avenue of exploration.

FUNDING INFORMATION

Guang Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd., and the UKRI Future Leaders Fellowship (MR/V023799/1).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Somiya Rani https://orcid.org/0000-0002-6345-4013
Akshi Kumar https://orcid.org/0000-0003-4263-7168
Guang Yang https://orcid.org/0000-0001-7344-7733

REFERENCES

1. Pons E, Braun LM, Hunink MM, Kors JA. Natural language processing in radiology: a systematic review. *Radiology*. 2016;279(2):329-343. doi:10.1148/radiol.16142770
2. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep*. 2017;7(1):1-2. doi:10.1038/srep46226
3. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;1(77):34-49. doi:10.1016/j.jbi.2017.11.011
4. Agrawal T, Choudhary P. Segmentation and classification on chest radiography: a systematic survey. *Visual Computer*. 2023;39(3):875-913. doi:10.1007/s00371-021-02352-7
5. Ignácio FD, Souza LR, D'Ippolito G, Garcia MM. Radiology report: what is the opinion of the referring physician? *Radiol Bras*. 2018;51:308-312. doi:10.1590/0100-3984.2017.0115
6. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging*. 2011;2(2):93-96. doi:10.1007/s13244-011-0066-7
7. The Royal College of Radiologists. *Standards for the Reporting and Interpretation of Imaging Investigations*. The Royal College of Radiologists; 2006. Accessed September 2010. www.rcr.ac.uk
8. Groupe de travail SFR_CRR. General recommendations for the development of a radiological report. *J Radiol*. 2007;88:304-306.
9. Brady AP. Radiology reporting—from Hemingway to HAL? *Insights Imaging*. 2018;9(2):237-246. doi:10.1007/s13244-018-0596-3
10. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J Biomed Inform*. 2019;92:103133. doi:10.1016/j.jbi.2019.103133
11. An Y, Xia X, Chen X, Wu FX, Wang J. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artif Intell Med*. 2022;127:102282. doi:10.1016/j.artmed.2022.102282
12. Zhang R, Lu W, Wang S, Peng X, Yu R, Gao Y. Chinese clinical named entity recognition based on stacked neural network. *Concur Comput Pract Exp*. 2021;33(22):e5775. doi:10.1002/cpe.5775
13. Lu P, Poesio M. Coreference resolution for the biomedical domain: a survey. *arXiv*. Preprint arXiv:2109.12424. 2021. doi:10.48550/arXiv.2109.12424
14. Hourali S, Zahedi M, Fateh M. A new model for coreference resolution based on knowledge representation and multi-criteria ranking. *J Intell Fuzzy Syst*. 2021;40(1):877-892. doi:10.3233/JIFS-201050
15. Zhang X, Zhang Y, Zhang Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int J Med Inform*. 2019;132:103985. doi:10.1016/j.ijmedinf.2019.103985
16. Li Z, Yang J, Gou X, Qi X. Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts. *Artif Intell Med*. 2019;97:9-18. doi:10.1016/j.artmed.2019.04.003
17. Dahl FA, Rama T, Hurlen P, et al. Neural classification of Norwegian radiology reports: using NLP to detect findings in CT-scans of children. *BMC Med Inform Decis Mak*. 2021;21(1):1-8. doi:10.1186/s12911-021-01451-8
18. Yuan J, Zhu H, Tahmasebi A. Classification of pulmonary nodular findings based on characterization of change using radiology reports. *AMIA Summ Transl Sci Proc*. 2019;2019:285. https://pubmed.ncbi.nlm.nih.gov/31258981
19. Li J, Zhao S, Yang J, et al. WCP-RNN: a novel RNN-based approach for Bio-NER in Chinese EMRs. *J Supercomput*. 2020;76:1450-1467. doi:10.1007/s11227-017-2229-x
20. Cowie J, Wilks Y. Information extraction. *Handb Nat Lang Process*. 2000;56:57.
21. Onan A. Consensus clustering-based undersampling approach to imbalanced learning. *Sci Program*. 2019;2019:5901087. doi:10.1155/2019/5901087
22. Onan A, Korukoğlu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst Appl*. 2016;15(57):232-247. doi:10.1016/j.eswa.2016.03.045
23. Onan A. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*. 2019;7:145614-145633. doi:10.1109/ACCESS.2019.2945911
24. Onan A, Korukoğlu S. A feature selection model based on genetic rank aggregation for text sentiment classification. *J Inf Sci*. 2017;43(1):25-38. doi:10.1177/0165551515613226
25. Onan A. Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion. *J King Saud Univ Computer Inf Sci*. 2023;13:101610. doi:10.1016/j.jksuci.2023.101610

26. Onan A. GTR-GA: harnessing the power of graph-based neural networks and genetic algorithms for text augmentation. *Expert Syst Appl*. 2023;24:120908. doi:10.1016/j.eswa.2023.120908
27. Onan A. SRL-ACO: a text augmentation framework based on semantic role labeling and ant colony optimization. *J King Saud Univ Computer Inf Sci*. 2023;9:101611. doi:10.1016/j.jksuci.2023.101611
28. Onan A. Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *J King Saud Univ Computer Inf Sci*. 2022;34(5):2098-2117. doi:10.1080/08839514.2022.2145641
29. Onan A. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurr Comput Pract Exp*. 2021;33(23):e5909. doi:10.1002/cpe.5909
30. Onan A. Mining opinions from instructor evaluation reviews: a deep learning approach. *Computer Appl Eng Educ*. 2020;28(1): 117-138. doi:10.1002/cae.22179
31. Onan A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Computer Appl Eng Educ*. 2021;29(3):572-589. doi:10.1002/cae.22253
32. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification. *J Inf Sci*. 2018;44(1):28-47. doi:10.1177/0165551516677911
33. Onan A, Toçoğlu MA. A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access*. 2021;9:7701-7722. doi:10.1109/ACCESS.2021.3049734
34. Onan A. Topic-enriched word embeddings for sarcasm identification. InSoftware Engineering Methods in Intelligent Algorithms: Proceedings of 8th Computer Science On-line Conference 2019, Vol. 1 8 2019 (pp. 293-304). Springer International Publishing. doi:10.1007/978-3-030-19807-7_29
35. Onan A, Korukoğlu S, Bulut H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Inf Process Manag*. 2017;53(4):814-833. doi:10.1016/j.ipm.2017.02.008
36. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. *J Biomed Inform*. 2020;109:103526. doi:10.1016/j.jbi.2020.103526
37. Li L, Xu W, Yu H. Character-level neural network model based on Nadam optimization and its application in clinical concept extraction. *Neurocomputing*. 2020;414:182-190. doi:10.1016/j.neucom.2020.07.027
38. Gerevini AE, Lavelli A, Maffi A, et al. Automatic classification of radiological reports for clinical care. *Artif Intell Med*. 2018; 91:72-81. doi:10.1016/j.artmed.2018.05.006
39. Venkataraman GR, Pineda AL, Bear Don't Walk OJ IV, et al. FasTag: automatic text classification of unstructured medical narratives. *PloS One*. 2020;15(6):e0234647. doi:10.1371/journal.pone.0234647
40. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc*. 2019;26(11):1297-1304. doi:10.1093/jamia/ocz096
41. López-Ubeda P, Díaz-Galiano MC, López LA, Martín-Valdivia MT, Martín-Noguerol T, Luna A. Transfer learning applied to text classification in Spanish radiological reports. In: *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*; 2020:29–32.
42. Prabhakar SK, Won DO. Medical text classification using hybrid deep learning models with multihead attention. *Comput Intell Neurosci*. 2021;2021:1-16. doi:10.1155/2021/9425655
43. Olthof AW, Shouche P, Fennema EM, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Comput Methods Programs Biomed*. 2021; 208:106304. doi:10.1016/j.cmpb.2021.106304
44. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014:1532–1543.
45. Yang Z, Chen H, Zhang J, Ma J, Chang Y. Attention-based multi-level feature fusion for named entity recognition. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI2020)*; 2021:3594–3600.
46. Melis G, Kočiský T, Blunsom P. Mogrifier lstm. *arXiv*. Preprint arXiv:1909.01792. 2019. doi:10.48550/arXiv.1909.01792
47. Rani S, Jain A. Aspect-based sentiment analysis of drug reviews using multi-task learning based dual BiLSTM model. *Multimed Tools Appl*. 2023;7:1-29. doi:10.1007/s11042-023-16360-3
48. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc*. 2016;23(2):304-310. doi:10.1093/jamia/ocv080
49. Johnson A, Lungren M, Peng Y, et al. MIMIC-CXR-JPG-chest radiographs with structured labels (version 2.0. 0). *PhysioNet*. 2019;10:8360-t248. doi:10.13026/8360-t248
50. Gu P, Wu T, Zou M, et al. Multi-head self-attention model for classification of temporal lobe epilepsy subtypes. *Front Physiol*. 2020;11:604764. doi:10.3389/fphys.2020.604764

How to cite this article: Rani S, Jain A, Kumar A, Yang G. CCheXR-Attention: Clinical concept extraction and chest x-ray reports classification using modified Mogrifier and bidirectional LSTM with multihead attention. *Int J Imaging Syst Technol*. 2024;34(1):e23025. doi:10.1002/ima.23025