International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Aug-2023)

http://dx.doi.org/10.12785/ijcds/140149

Bengali Query Processing System for Disease Detection using LSTM and GRU

Kailash Pati Mandal[1], Prasenjit Mukherjee[2], Souvik Ganguly[3] and Baisakhi Chakraborty[4]

[1]*Computer Science and Engineering, National Institute of Technology, Durgapur, India*
[2]*Computer Science and Engineering, National Institute of Technology, Durgapur, India*
[3]*Bachelor of Computer Application, Dr. B.C. Roy Engineering College, Durgapur, India*
[4]*Computer Science and Engineering, National Institute of Technology, Durgapur, India*

Received 08 May 2022, Revised 22 Jul. 2023, Accepted 25 Jul. 2023, Published 01 Aug. 2023

Abstract: This paper proposes a disease detection system where it receives the query in form of symptoms of the disease in the Bengali language. This system is able to handle natural language queries in Bengali. The proposed system assists a layman to detect a probable disorder or disease in their body using disease symptoms. The proposed research work is challenging due to insufficient resources in vernacular languages like Bengali. This system receives a description of the patient's symptoms in the Bengali language and after processing the natural language text, it detects any potential disorders or diseases that may have occurred. This research work has been implemented separately by using the two most popular sequential prediction models. One is Bi-directional Long Short-Term Memory (Bi-LSTM) and the other is Bi-directional Gated Recurrent Unit (Bi-GRU). Both Bi-GRU and Bi-LSTM have provided significant results on a dataset of 3714 samples. The raw clinical text categorization data has been gathered from the Kaggle to build the detection model. The performances of disease detectability of both models have been measured using precision, recall and f1-score. The accuracy of the proposed system using the Bi-LSTM and the Bi-GRU models are 97.85% and 99.73%, respectively.

Keywords: Bi-LSTM network, Bi-GRU network, Bengali language, Disease prediction, Natural Language Processing (NLP)

1. INTRODUCTION

People are attacked by different types of diseases at several stages of life. Good health is not an option; it is a requirement of life. When people face some health problems, it takes a long time to detect the disease and mostly in remote areas, there are few medical practitioners to detect diseases. With the help of the proposed system, quick detection of diseases is possible which may enable patients to access quick precautions. The symptom-based machine-learning system proposed in this work detects the appropriate disease depending upon the description of the symptoms. This system enables immediate first help to the patient for the first aid in their nearby area as soon as a disease is detected. Various computerized automated assessment, prediction and detection systems have already been developed. We shall discuss a few of them. By incorporating relevant data from Arabic feature patterns, a Bi-directional LSTM Network (Bi-LSTM) has been studied to improve the Arabic Emotion Assessment. The results of these experiments on six standard emotion classification datasets show that this model outperforms both state-of-the-art deep neural networks and standard traditional machine-learning approaches as mentioned in [1]. The unsupervised Continuous Bag of Words (CBOW) and Skip-gram models have been used to predict the political ideology of Bangladesh's people, as well as comparative analysis, has also been done as described in [2]. The Bengali news articles are divided into different categories by using the Multi-Label K-Nearest Neighbor (ML-KNN) algorithm, Neural Network and Count Vectorizer as a feature vector as described in [3]. Written text attitude characterization is essential in various Human Computer Interaction (HCI) applications where the wording is used as the medium of interaction, such as responses, comments, forums and other Web 2.0 platforms. To complete the linguistic expression categorization job, an emotion corpus including 8047 Bengali texts was created as discussed in [4]. The GloVe embedding and Very Deep Convolution Neural Network (VDCNN) classifiers have been utilized for Bengali text document classification. The best embedding parameter has been chosen using Embedding Parameters Identification (EPI) Algorithm for low resource languages like Bengali. The GloVe

E-mail: biltu.cse@gmail.com, prasen.mscit09@gmail.com, souvikganguly027@gmail.com, 2baisakhichakraborty@gmail.com

embedding model performs better than another embedding model as explained in [5]. The multi-class classifier has been implemented for Bengali news classification using a Multi-layer Neural network, Naive Bayes and Support Vector Machine as described in [6]. The human abnormality is detected from the Bengali text stated by the person. The count vector and TF-IDF have been utilized to capture patterns from the dataset. A comparative study has also been performed on Naive Bayes and Support Vector Machine (SVM) as described in [7]. Convolutional Neural Network (CNN), LSTM, SVM and Stochastic Gradient Descent (SGD) have been utilized to develop the authorship classification model. The authorship classification model has been tested on Bengali authorship classification corpus (BACC-18), Bangla Authorship Attribution Dataset (BAAD16) and Linked Data (LD) Bengali datasets as explained in [8]. A medical recommendation system has been developed by using Altibbi disease classification databases. For next word prediction, various machine-learning models are evaluated and tuned. The algorithms are analyzed with regards to training accuracy and loss, validation accuracy and loss, and testing accuracy utilizing 3-gram and 4-gram interpretations of the datasets that have been described in [9]. The text mining approach has been applied to predict the highly crime-prone area in the West Bengal from the newspaper as implemented in [10]. Various machine-learning based prediction systems, as well as the classification system, have been implemented in various domains. But very few implementations have been done in the medical domain. The proposed system is a supervised machine-learning model that accepts the patient's symptoms description in the Bengali language. This system allows a patient personally to predict a possible illness in their body. The proposed system predicts diseases like Diabetes, Chickenpox, etc, that are depending upon symptoms description. Clinical text classification data has been collected from the Kaggle [11] to train the proposed system.

This research article consists of various sections. Section 2 narrates related works. Section 3 explains the proposed system. Section 4 narrates the methodology. Section 5 analyses experimental results of the proposed system. Section 6 presents the limitations and future works. The proposed system is concluded in Section 7.

2. RELATED WORKS

The multi-label news classification has been developed for the Arabic language. By wrapping each classifier in an OneVsRest classifier, the news classifier is able to leverage multi-labeling classifiers like Logistic Regression and XGBoost as explained in [12]. Finding the reasons behind an emotion that is expressed in a document is known as emotion cause extraction. To combine the aforementioned data, a Bi-directional Long Short-Term Memory Condition Random Fields (Bi-LSTM-CRF) network based on attention is implemented as described in [13]. The random forest classifier has been used to identify whether the Bengali news is fake or real which is implemented in [14]. Transformer-based structures produce multimodal representations of an utterance. The segment-level audio features and word-level lexical features are used as inputs for emotion recognition as described in [15]. The BanglaLekhaImageCaptions dataset has been employed for Bengali text generation from the image as explained in [16]. Logistic Regression (LR), Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) have been utilized as baseline classifiers for the detection of Extreme Guilt as well as Grave Fault in the Bengali language as discussed in [17]. A rule-based sentimental analysis has been implemented for Bengali text. The Support Vector Machine (SVM) and term frequency-inverse document frequency (tf-idf) have been employed to develop the system as described in [18]. A hate speech detection system for the Bengali language has been implemented using an attention-based decoder that is explained in [19]. Today, the majority of web interfaces for databases only accept search queries that are input as text strings or keywords. An enhanced neural model for Natural Language (NL) database querying has been applied on the IRNet architecture. A public dataset with experimental results demonstrates intriguing possibilities as described in [20]. A Hindi text-based hostility detection system has been developed using Constraint@AAAI 2021 Hindi animosity identification data source. In this hostile data source, there are overlapping classes such as fraudulent, objectionable, hatred and defamation. CNN, LSTM and Bidirectional Encoder Representations from Transformers (BERT) are examples of multi-label categorization approaches which is used to identify hostility as implemented in [21]. Long Short-Term Memory (LSTM) and gated recurrent units (GRUs) are utilized to train the hotel customer feedback for recognizing customer opinion with a significant accuracy rate. Naive Bayes, Decision Tree, Random Forest, and SVM are also used to test the dataset which is discussed in [22]. Artificial Intelligence (AI) and Natural Language Processing (NLP) have been used to construct a healthcare and rehabilitation aid system. This system uses Automated machine-learning techniques and NLP, where this system can aid in the diagnosis and treatment of long-term illnesses as implemented in [23]. A convolution layer and gated recurrent units are used to leverage serial context for assessing the usefulness of a review. This method provides better results than five classical learning methods and two deep learning models as explained in [24]. Chinese location components are separated from the address statement using a bidirectional gated recurrent unit (Bi-GRU) neural network. The Viterbi algorithm has been applied to perform the separation of Chinese address elements as described in [25]. Although neural networks have achieved astoundingly high levels of accuracy in applications like image recognition and natural language processing, they are sometimes viewed as opaque models that are challenging to understand. The equation learner (EQL) network, a neural network-based symbolic regression architecture has been utilized for extraction the digit as explained in [26]. A Twitter dataset is used to evaluate multiple current Language Identification (LID) technologies in order to create a group of interrelated standards and publish it publicly so that someone else can contribute to this group of benchmarks. An enhanced N-Gram regional convolution neural network with an attentiveness approach produces a significant result as described in [27]. The performances of various current negative), which were then classified into bigger topics using thematic analysis. This study finds 34 adverse themes, 17 of which are economic, socio-political, educational and political concerns as discussed in [29]. A three-stage method based on Deep Learning and Natural Language Processing techniques to examine, understand and forecast the future stock prices as explained in [30]. The comparative study has been given in Table I.

TABLE I. A COMPARISON OF SYSTEMS THAT ARE QUITE SIMILAR TO THE PROPOSED SYSTEM

AUTHOR(S)	METHODOLOGY USED IN IDENTICAL TYPE SYSTEM	METHODOLOGY APPLIED IN THE PROPOSED SYSTEM
Zerin Tasnim et al. [2]	1) Political position has been predicted from Bengali textual data using the CBOW and Skip-gram word embedding models. 2) This is an unsupervised model. 3) This prediction system provides 76.22%.	1) The Bi-directional LSTM as well as Bi-directional GRU have been used to implement the proposed system. 2) The proposed system is a supervised model. 3) The Bidirectional LSTM and Bi-directional GRU achieve accuracy of 97.85%, 99.73% respectively.
Wahiduzzaman Akanda et al. [3]	1) The multi-label Bengali text classification has been done using the ML-KNN algorithm and Neural Network. 2) This system provides accuracy 75%. 3) This system uses Bengali text to classify an article.	1) The proposed system is a Bengali multi-label text classification based prediction system that uses Bidirectional LSTM as well as GRU. 2)The Bidirectional LSTM and Bi-directional GRU exhibit accuracy of 97.85%, 99.73% respectively. 3) The proposed approach uses symptoms written in Bengali to forecast sickness.
Tanzia Parvin et al. [4]	1) The system uses eight different Machine-Learning approaches to explore emotion based on Bengali textual input. 2) The maximum weighted f1-score of 62.39% is provided by this system. 3) This system predicts emotions.	1) The proposed system uses Bi-directional LSTM and Bi-directional GRU to explore Bengali text. 2) The Bidirectional LSTM and Bi-directional GRU provides accuracy of 97.85%, 99.73%. 3) The proposed system predicts the disease.
Md. Rajib Hossain et al. [5]	1) A deep convolutional neural network have been used to classify Bengali text content. 2) This system provides accuracy of 96.96%. 3) This system determines the Bengali text document's class.	1) The proposed system uses Bi-directional Recurrent Neural Network. 2) This system exhibits accuracy of 97.85%, 99.73% using Bidirectional LSTM and Bi-directional GRU respectively. 3) The proposed system determines the disease based on Bengali text.
Sharmin Yeasmin et al. [6]	1) The multi-layer dense neural network has been used to classify Bengali news. 2) The accuracy of SVM, Logistic regression, and Multi-layer dense neural network, are 94.99%, 94.60%, and 95.50% respectively. 3) This is a news classification system.	1) The proposed system uses Bidirectional LSTM and GRU to classify disease based Bengali text. 2) The accuracy of Bidirectional LSTM and Bi-directional GRU are 97.85%, 99.73% respectively. 3) The proposed system is a disease prediction system.
M. Firoz Mridha et al. [7]	1) This system is developed using Nave Bayes and Support Vector Machine. 2) The accuracy of this system is 89%. 3) This system predicts human Irregularity.	1) The proposed system is developed using Bi-directional LSTM and Bi-directional GRU. 2) The proposed system shows 97.85%, 99.73% using Bidirectional LSTM and Bi-directional GRU respectively. 3) The proposed system predicts disease.
Md. Rajib Hossain et al. [8]	1) Convolutional Neural Networks has been utilized to developed Authorship Classification system. 2) For the datasets BACC-18, BAAD16, and LD, this approach returns 93.45%, 95.02%, and 98.67%. 3) The authorship of a Bengali text is recognized by this system.	1) The proposed system has been developed using Bidirectional LSTM and Bi-directional GRU. 2) The proposed system exhibits 97.85%, 99.73% using Bidirectional LSTM and Bi-directional GRU respectively. 3) The proposed system recognizes disease using symptoms.

Deep Learning based models for multi-label text categorization algorithms have been assessed using assessment measures. Glove, word2vec and FastText have been utilized to prepare the embedding corpus for the provided models as implemented in [28]. Over 1 million chosen random responses yielded pertinent opinionated keywords and their related sentiment polarity (positive or

3. THE PROPOSED SYSTEM

The workflow diagram for the proposed system has been given in figure 1.

A. *Execution steps of the proposed system*

The proposed system's execution steps have been given below.

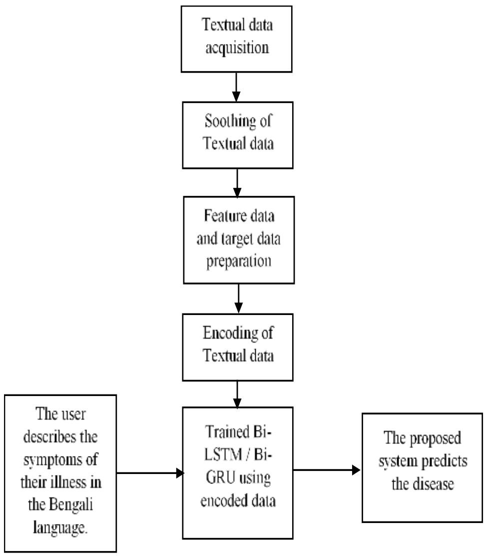

Figure 1. Workflow diagram of the proposed system

Step 1: Textual data acquisition

Health-related clinical tabular data has been collected from the Kaggle [19] in the English language. The tabular data consists of diseases and their allied symptoms. The English raw data has been transformed into the Bengali language except for the disease column. The transformed dataset has been given in Table II. The IPA notation has been used to represent the actual pronunciation of the Bengali data. In this article, Table II is the transformed form of English data whereas it has shown the sub-part of the original data as an example.

Step 2: Soothing of textual data

For machine-learning purposes, the proposed system uses data from columns Disease, Symptom_1, Symptom_2, Symptom_3, Symptom_4 and Symptom_5 of the dataset. The transformed Bengali data also contains non-Bengali symbols and punctuation marks. These irrelevant symbols are removed from the dataset. The proposed system discards the row if any of its column(s) contains null values. The filtered dataset is shown in Table III as the sub-part of the original data as an example.

Step 3: Feature data and target data preparation

In order to prepare the feature data, Symptom_1, Symptom_2, Symptom_3, Symptom_4 and Symptom_5 columns were combined. The data from the Disease column has been used as the target data.

Step 4: Encoding textual data

Deep learning models are all mathematical models that require numbers to perform tasks. Table II depicts categorical data that has not been fitted with a deep learning model and hence requires encoding. The proposed system assigns a number to each unique Bengali word in the Symptom_1, Symptom_2, Symptom_3, Symptom_4 and Symptom_5 columns. All of the English words in the disease columns have been converted into numerical values. The number sequence represents all of the textual values in the Symptom_1, Symptom_2, Symptom_3, Symptom_4 and Symptom_5 columns. The encoded

TABLE II. CONVERTED RAW BENGALI DATA

Disease	**Symptom_1**	**Symptom_2**	**Symptom_3**	**Symptom_4**	**Symptom_5**
Hypertension	বুকব্যাথা। (bukbæt̪ʰa)	মাথাঘোরা, (mat̪ʰagʱora)	ভারসাম্য হারানো(bʱarʃammo harano)	মনোযোগের অভাব (monoɟoger ɔbʱab)	
Jaundice	চুলকানি (culkani)	বমি (bomi)	ক্লান্তি (klant̪i)	“ওজন কমানো” (oɟon kɔmano)	মাত্রাতিরিক্ত জ্বর,(mat̪rat̪irikt̪o ɟɔr)
Osteoarthristis	সংযোগে ব্যথা (ʃɔŋɟoge bæt̪ʰa)	ঘাড়ব্যথা (gʱaɽbæt̪ʰa)	RE:হাঁটুরব্যাথা (hãturbæt̪ʰa)	জয়েন্টগুলোতে ফোলা (ɟɔʲentgulot̪e pʰola)	বেদনাদায়ক হাঁটা (bed̪onad̪aʲok hãta)

TABLE III. FILTERED DATA

Disease	**Symptom_1**	**Symptom_2**	**Symptom_3**	**Symptom_4**	**Symptom_5**
Jaundice	চুলকানি (culkani)	বমি (bomi)	ক্লান্তি (klant̪i)	ওজন কমানো (oɟon kɔmano)	মাত্রাতিরিক্ত জ্বর (mat̪rat̪irikt̪o ɟɔr)
Osteoarthristis	সংযোগে ব্যথা (ʃɔŋɟoge bæt̪ʰa)	ঘাড়ব্যথা (gʱaɽbæt̪ʰa)	হাঁটুরব্যাথা (hãturbæt̪ʰa)	জয়েন্টগুলোতে ফোলা (ɟɔʲentgulot̪e pʰola)	বেদনাদায়ক হাঁটা (bed̪onad̪aʲok hãta)

numbers of feature data, as well as target data has been given below. These encoded data are used to train the Bi-LSTM and the Bi-GRU models.

Encoded feature data

[বমি, বদহজম, ক্ষুধামান্দ্য, পেটে, ব্যথা, গ্যাসের, উত্তরণ]

[3, 36, 37, 13, 20, 31, 122]

Encoded target data

[Drug Reaction, Diabetes, Chicken pox]

[11, 9, 5]

Step 5: Trained the Bi-LSTM and the Bi-GRU Model using encoded data

Bi-LSTM

Different types of multiclass text classification models are available and the Long Short-Term Memory (LSTM) is one of them. The LSTM is made up of forget gate, input gate and output gate. The current study uses Bi-LSTM because it produces promising results for multiclass sequential data where accuracy is concerned. The Bi-LSTM model is a revised version of the LSTM model. The Bi-LSTM enhances the model's performance of categorization tasks for sequential data. This model uses two LSTMs instead of one throughout in the learning phase. The Bi-LSTM solves the limitations of traditional Recurrent Neural Networks (RNNs). The Bi-LSTM is used to access all prior data as well as expected future data at the same time. Users can provide the actual information to the learning process twice via Bi-LSTMs. One from start to finish and the other from finish to start [31]. The block diagram of an LSTM is shown in figure 2.

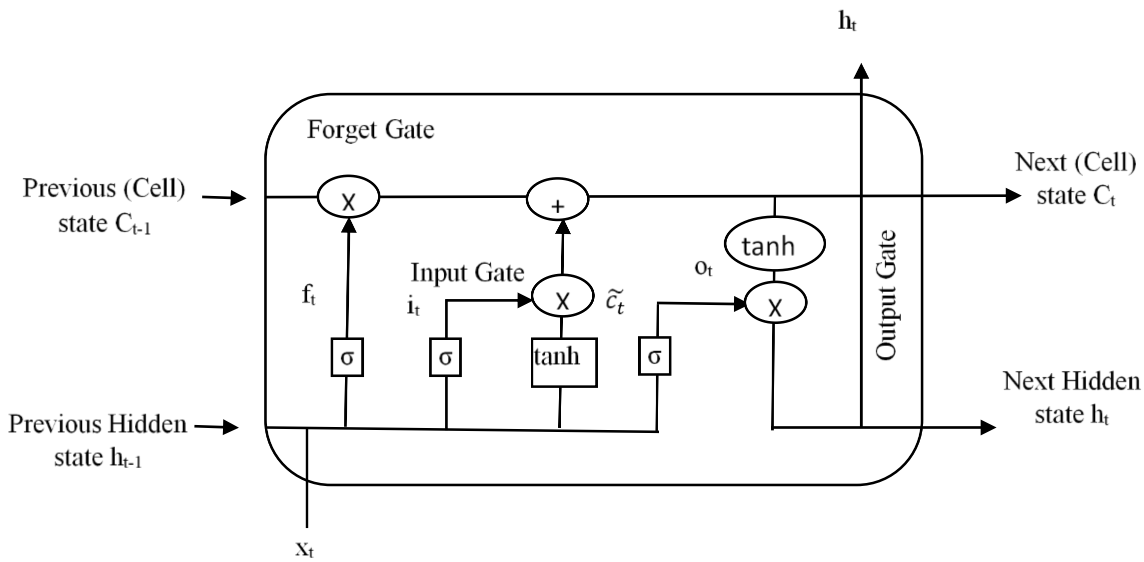

Figure 2. Long Short Term Memory (LSTM)

The forget gate determines which information from the cell state will be discarded. Mathematical equation (1) expresses the forget gate.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (1)$$

Here,

f_t: Forget gate.

W_f: Associated weight matrix of the input data.

h_{t-1}: Hidden state at the earlier timestamp. The hidden state is basically just an encoding of the information provided to check for time dependencies.

x_t: Input at the current timestamp t.

b_f: Bias parameter value which is learned from input training data.

σ: Sigmoid function will output values between 0 and 1.

The input gate calculates what new data is to be stored in the cell state. Mathematical expressions (2) and (3) represent input gate.

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

Here,

i_t: Input gate.

W_t: Associated weight matrix of the input data.

h_{t-1}: A hidden state at the previous timestamp.

x_t: Input at the current timestamp t.

b_i: Bias parameter value.

σ: Sigmoid function will output values between 0 and 1.

$\widetilde{C_t}$: Candidate cell state.

W_c: Associated weight matrix of the input data.

b_c: Bias parameter value.

tanh: tanh function will output values between -1 and 1.

The mathematical expression (4) describes the updated cell state.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

Here,

C_t: Update cell state.

f_t: forget gate.

C_{t-1}: Previous cell state.

$\widetilde{C_t}$: Candidate cell state.

Output gate and hidden state are represented by the mathematical equations (5) and (6).

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t * \tanh(C_t) \quad (6)$$

Here,

O_t: Output gate.

W_o: Weight associated with the input.

h_{t-1}: Hidden state at the earlier timestamp.

x_t: Input at the current timestamp t.

b_o: Bias parameter value.

h_t: A hidden state of the current timestamp.

C_t: Current cell state.

σ: Sigmoid function will output values between 0 and 1.

tanh: tanh function will output values between -1 and 1.

Bi-GRU

A Gated Recurrent Unit (GRU) in machine-learning algorithms (RNN) is a gated mechanism equivalent to a Long Short-Term Memory (LSTM) unit having reset and update gates. When the execution time is concerned then GRU performed better because of their relatively simple design. The GRU consists of an update gate, reset gate and current memory gate. The update gate specifies how much previous information must be transferred for the future. The reset gate controls what kind of one's previous knowledge is erased [32]. The proposed system uses the Bi-GRU model that consists of two-GRU models for sequence processing. One receives input in a forward direction, whereas the other receives it in backwards. The block diagram of a GRU is depicted in figure 3.

The mathematical equation of the update gate is given by the equation (7).

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (7)$$

Here,

z_t: Update gate.

W_z: Associated weight matrix of the input data.

h_{t-1}: Hidden state at the earlier timestamp.

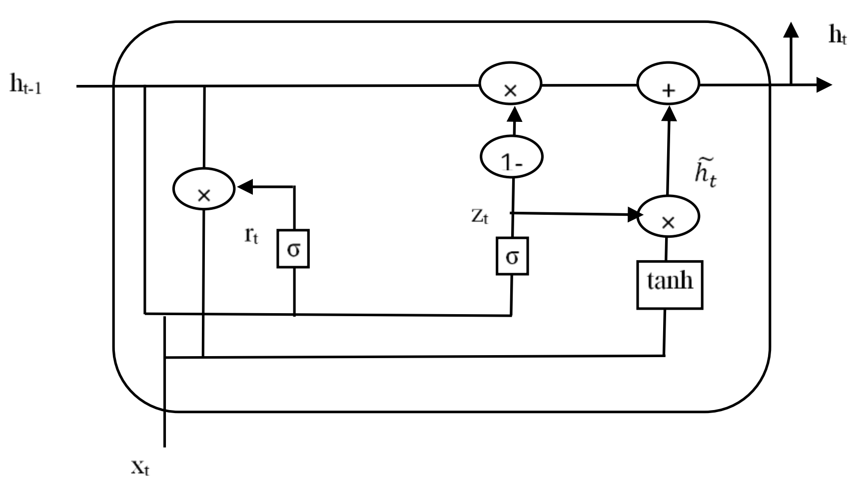

Figure 3. Gated Recurrent Unit (GRU)

x_t: Input at the current timestamp t.

σ: Sigmoid function will output values between 0 and 1.

The mathematical equation of the reset gate is given by the equation (8).

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (8)$$

Here,

r_t: Reset gate.

W_r: Associated weight matrix of the input data.

h_{t-1}: Hidden state at the earlier timestamp.

x_t: Input at the current timestamp t.

σ: Sigmoid function will output values between 0 and 1.

The mathematical equation of the current memory gate is given by equation (9).

$$\widetilde{h_t} = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (9)$$

Here,

$\widetilde{h_t}$: Current memory gate.

W: Associated weight matrix of the input data.

h_{t-1}: Hidden state at the earlier timestamp.

x_t: Input at the current timestamp t.

r_t: Reset Gate.

tanh: tanh function will output values between -1 and 1.

The mathematical equation of the current hidden state is given by equation (10).

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \quad (10)$$

Here,

h_t: Current hidden state.

z_t: Update gate.

$\widetilde{h_t}$: Current memory gate.

h_{t-1}: A hidden state at the previous timestamp. The hidden state is essentially just an encoding of the information you gave it keeping the time-dependencies in check.

Bi-directional Recurrent Neural Networks (BRNN)

RNN processing is repeated in a Bi-directional RNN (BRNN). The inputs are handled in both forward and reverse sequences in a BRNN. This enables a BRNN to think about future context. Bi-LSTM and Bi-GRU models are frequent choices in NLP for this reason. Figure 4 depicts the BRNN.

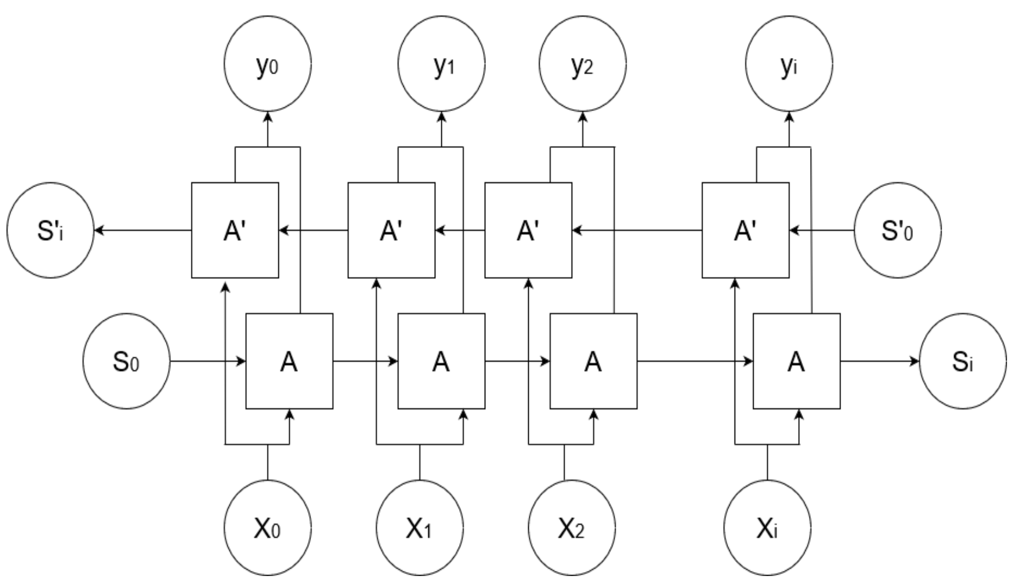

Figure 4. Bi-directional Recurrent Neural Networks (BRNN)

Here,

x_i: Input.

y_i: Output.

A: Number of hidden units.

S_i: Forward hidden state.

A': Number of hidden units.

S'_i: Backward hidden state.

B. Input validation steps of the proposed system

The input validation steps of the proposed system have been shown in figure 5. The validation steps have been illustrated below.

Step 1: Submission of user's symptoms

In Bengali, the patient describes his or her illness as input to the proposed system. The following is an example of a query.

Input =“আমার ত্বকে চুলকানি হয়, আগের থেকে ওজন কম হয়েছে, বমি হয়, ক্লান্তি লাগে এবং ত্বক হলুদ হয়েছে।” (amar t̪ke culkani hɔʲ, ager t̪ʰeke oɟon kɔm hoʲecʰe, bomi hɔʲ, klant̪i lage eboŋ t̪k holud̪ hoʲecʰe)

Step 2: Tokenization of user’s symptoms

The symptoms description entered by the user is tokenized and kept in a list. The list of tokens has been given below.

Token_list = [“আমার”, “ত্বকে”, “চুলকানি”, “হয়”, “,” ,“আগের”, “থেকে”, “ওজন”, “কম”, “হয়েছে”, “,”,“বমি”, “হয়”, “,”,“ক্লান্তি”, “লাগে”, “,”, “এবং”, “ত্বক”, “হলুদ” ,“হয়েছে”, “।”]. ([“amaro”, “t̪t̪oke”, “culkani”, “hɔʲ”, “,” ,“ager”, “t̪ʰeke”, “oɟɔn”, “km”, “hoʲecʰe”, “,”, “bomi”, “hoʲ”, “,”, “klant̪i”, “lage”, “,”, “eboŋ”, “t̪t̪ok”, “holud̪”, “hoʲecʰe”, “।”])

Step 3: Removal of punctuation marks

The punctuation marks are dropped from the token list. The token list has been given below after the removal of punctuation marks.

Token without punctuation marks = [“আমার”, “ত্বকে”, “চুলকানি”, “হয়”, “আগের”, “থেকে”, “ওজন”, “কম”, “হয়েছে”, “বমি”, “হয়”, “ক্লান্তি”, “লাগে”, “এবং”, “ত্বক”, “হলুদ”, “হয়েছে”] ([“amaro”, “t̪t̪oke”, “culkani”, “hɔʲo”, “ager”, “t̪ʰeke”, “oɟɔn”, “km”, “hoʲecʰe”, “bmi”, “hɔʲ”, “klant̪i”, “lage”, “eboŋ”, “t̪t̪ok”, “holud̪”, “hoʲecʰe”])

Step 4: Removal of stops words

The Bengali stop(s) are removed from the token list. Below is a list of filtered tokens.

Filtered token = [“ত্বকে”, “চুলকানি”, “আগের”, “ওজন”, “কম”, “বমি”, “ক্লান্তি”, “লাগে”, “ত্বক”, “হলুদ”] ([“ t̪t̪oke”, “culkani”, “ager”, “oɟon”, “kom”, “bomi”, “klant̪I”, “lage”, “t̪k”, “holud̪”])

Step 5: Passing of the padded sequence of filtered tokens through both Bi-LSTM and Bi-GRU

The proposed approach allows users to describe their medical symptoms in up to 150 words. After removing the stop word(s), the padded sequence is passed into both Bi-LSTM and Bi-GRU.

[[1 16 1 13 1 2 3 1 9 7 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0]]

Step 6: Output generation

The previously trained Bi-LSTM and the Bi-GRU models are used to determine the probability vector for a certain input sequence with 33 different diseases symptoms. The Bi-LSTM and the Bi-GRU probabilities are listed below.

Bi-LSTM

[[0.0195459 0.02801642 0.00228327 0.01037019
0.02141957 0.04064025 0.07692058 0.01964119
0.03173987 0.00591955 0.06556281 0.00894377
0.04046099 0.03802254 0.02681678 0.04879355
0.01290938 0.03576077 0.01732328 0.03042845
0.00636274 0.08043915 0.21874340 0.01351665
0.00324032 0.00435869 0.00949593 0.01769513
0.00616443 0.02744493 0.00453764 0.00267629
0.0238056]]

Bi-GRU

[[9.7272851e-02 2.0754285e-02 4.1573490e-03
1.1142943e-02 2.1368492e-02 9.8857597e-02
7.4467383e-02 5.3682018e-02 2.4345320e-02
5.8964103e-02 2.3417292e-02 7.2171111e-03
1.1351527e-02 7.3789254e-02 2.8960998e-03
1.6664708e-02 1.6821600e-03 3.9005004e-02
4.5338366e-03 1.6695756e-03 4.3332012e-04
2.0512661e-03 1.9103217e-01 2.3759296e-03
3.9633900e-02 6.4627871e-05 5.5297604e-03
2.3034506e-03 6.2391777e-03 6.0639709e-02
1.0780096e-02 2.8194321e-02 3.4832866e-03
]]

Finally, previously trained models select the vector with maximum probabilities that help to predict the probable disease.

The Bi-LSTM model in its output, predicts jaundice.

The Bi-GRU model in its output, predicts jaundice.

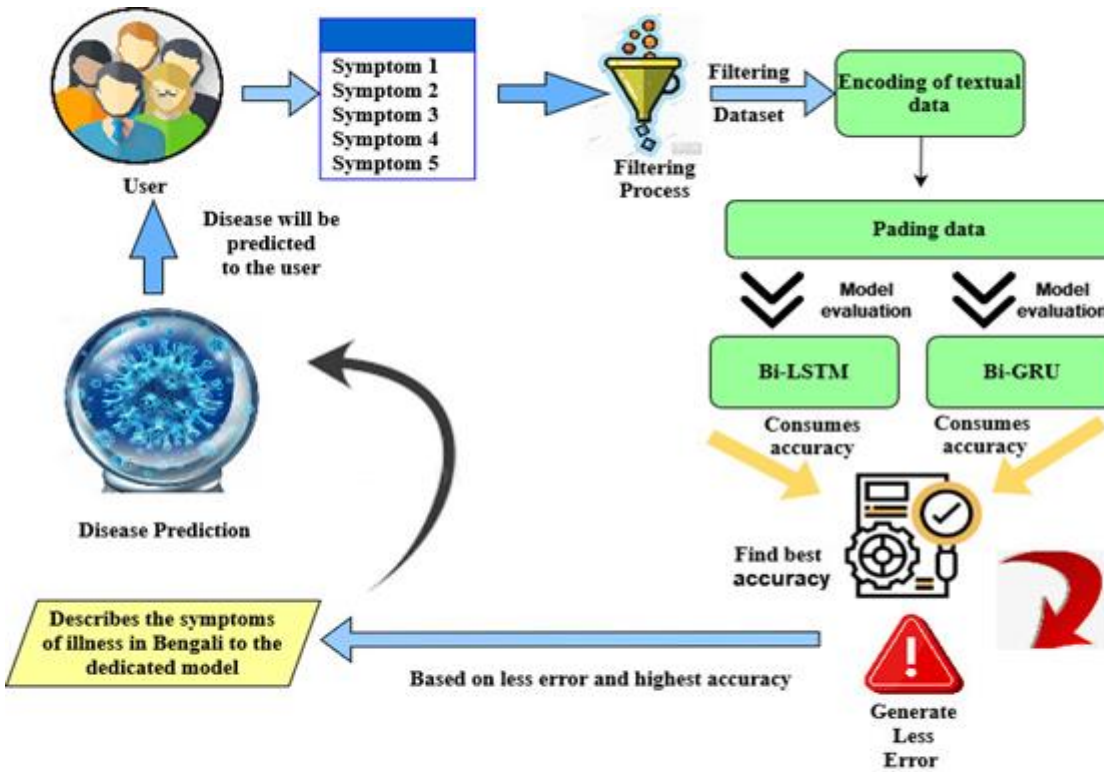

Figure 5. Visualization of input validation steps

4. EXPERIMENTAL RESULT ANALYSIS OF THE PROPOSED SYSTEM

The proposed study conducts numerous observations on the Bi-LSTM and the Bi-GRU such as Root Mean Square Error (RMSE), Mean Squared Error (MSE), precision, recall, F1-score and model accuracy respectively.

A. Classification reports of Bi-directional LSTM and Bi-directional GRU

The Bi-LSTM as well as the Bi-GRU models classification reports of each disease have been depicted in Table IV and Table V respectively.

$$\text{Precesion} = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \qquad (11)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \qquad (12)$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precesion\ \times\ Recall}{Precesion\ +\ Recall} \qquad (13)$$

B. Accuracy of the proposed system using Bi-directional LSTM and Bi-directional GRU

TABLE VI shows the training testing and validation accuracy of the Bi-LSTM and the Bi-GRU respectively. TABLE VII demonstrates the overall accuracy of the above mention models.

C. Error analysis for proposed system

The error analysis for the models has been given in TABLE VIII. TABLE VIII indicates that the Bi-GRU model outperforms the existing Bi-LSTM model. Low RMSE and MSE imply that the model accurately predicts the superior future value. TABLE VI indicates that the training, validation and testing accuracy of the Bi-LSTM model are 95.55%, 95.67% and 96.24% respectively whereas the training, validation and testing accuracy of the

TABLE IV. CLASSIFICATION REPORTS OF THE BI-DIRECTIONAL LSTM

Diseases	Precision	Recall	F1-Score
(vertigo) Paroymsal Positional Vertigo	100.00	100.00	100.00
Alcoholic hepatitis	100.00	100.00	100.00
Arthritis	100.00	100.00	100.00
Bronchial Asthama	100.00	100.00	100.00
Cervical spondylosis	100.00	100.00	100.00
Chicken pox	100.00	100.00	100.00
Chronic cholestasis	100.00	100.00	100.00
Common cold	100.00	100.00	100.00
Dengue	100.00	100.00	100.00
Diabetes	100.00	100.00	100.00
Dimorphic hemmorhoids(piles)	100.00	100.00	100.00
Drug Reaction	100.00	100.00	100.00
GRED	100.00	100.00	100.00
Hepatitis B	100.00	100.00	100.00
Hepatitis C	100.00	100.00	100.00
Hepatitis D	85.71	60.00	70.59
Hepatitis E	100.00	75.00	85.71
Hypertension	100.00	100.00	100.00
Hyperthyroidism	100.00	100.00	100.00
Hypoglycemia	100.00	100.00	100.00
Hypothyroidism	100.00	100.00	100.00
Impetigo	100.00	100.00	100.00
Jaundice	92.31	100.00	96.00
Malaria	100.00	100.00	100.00
Migraine	100.00	100.00	100.00
Osteoarthristis	100.00	100.00	100.00
Peptic ulcer disease	100.00	100.00	100.00
Pneumonia	100.00	100.00	100.00
Psoriasis	100.00	100.00	100.00
Tuberculosis	100.00	100.00	100.00
Typhoid	94.12	100.00	96.97
Varicose veins	100.00	100.00	100.00
Hepatitis A	64.29	100.00	78.26

TABLE. V. CLASSIFICATION REPORTS OF THE BI-DIRECTIONAL GRU

Diseases	Precision	Recall	F1-Score
(vertigo) Paroymsal Positional Vertigo	100.00	100.00	100.00
Alcoholic hepatitis	100.00	100.00	100.00
Arthritis	100.00	100.00	100.00
Bronchial Asthama	100.00	100.00	100.00
Cervical spondylosis	100.00	100.00	100.00
Chicken pox	100.00	100.00	100.00
Chronic cholestasis	100.00	100.00	100.00
Common cold	100.00	100.00	100.00
Dengue	100.00	100.00	100.00
Diabetes	100.00	100.00	100.00
Dimorphic hemmorhoids(piles)	100.00	100.00	100.00
Drug Reaction	100.00	100.00	100.00
GRED	100.00	100.00	100.00
Hepatitis B	100.00	100.00	100.00
Hepatitis C	100.00	100.00	100.00
Hepatitis D	90.91	100.00	95.24
Hepatitis E	100.00	87.50	93.33
Hypertension	100.00	100.00	100.00
Hyperthyroidism	100.00	100.00	100.00
Hypoglycemia	100.00	100.00	100.00
Hypothyroidism	100.00	100.00	100.00
Impetigo	100.00	100.00	100.00
Jaundice	100.00	100.00	100.00
Malaria	100.00	100.00	100.00
Migraine	100.00	100.00	100.00
Osteoarthristis	100.00	100.00	100.00
Peptic ulcer disease	100.00	100.00	100.00
Pneumonia	100.00	100.00	100.00
Psoriasis	100.00	100.00	100.00
Tuberculosis	85.71	100.00	92.31
Typhoid	100.00	100.00	100.00
Varicose veins	100.00	100.00	100.00
Hepatitis A	100.00	100.00	100.00

TABLE VI. ACCURACY OF THE PROPOSED MODELS

Models	Training Accuracy	Validation Accuracy	Testing Accuracy
Bi-LSTM	95.55%	95.67%	96.24%
Bi-GRU	99.48%	99.55%	99.73%

TABLE VII. OVERALL ACCUARCY OF THE PROPOSED SYSTEM

Models	Accuracy
Bi-LSTM	97.85%
Bi-GRU	99.73%

Bi-GRU model are 99.48%, 99.55% and 99.73% respectively. The overall accuracy of the Bi-LSTM model is 97.85%, while the Bi-GRU model is 99.73%, which is greater than the value of the Bi-LSTM model, as shown in Table VII.

TABLE VIII. ERROR ANALYSIS FOR THE PROPOSED SYSTEM

Model	Root Mean Squared Error (RMSE)	Mean Squared Error (MSE)
Bi-LSTM	1.9831277565816912	3.932795698924731
Bi-GRU	0.05184758473652126	0.002688172043010753

D. Graphical representation of training-validation accuracy range and loss range

Figure 6 and figure 7 reflect the training validation accuracy ranges of the Bi-LSTM and Bi-GRU respectively. Figure 8 and figure 9 indicate the training validation loss ranges of the Bi-LSTM and Bi-GRU respectively.

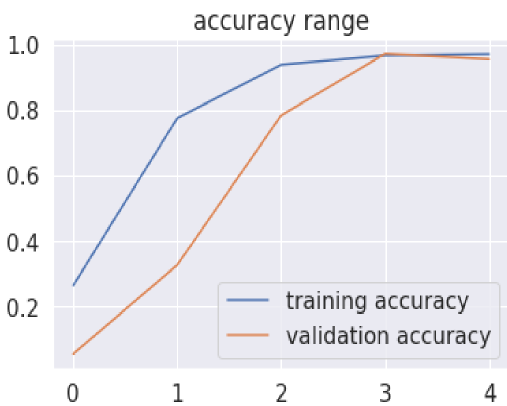

Figure 6. Accuracy range of the Bi-directional LSTM

Figure 7. Accuracy range of the Bi-directional GRU

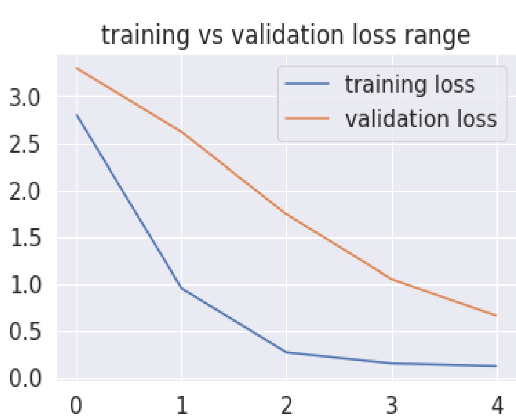

Figure 8. Loss range of the Bi-directional LSTM

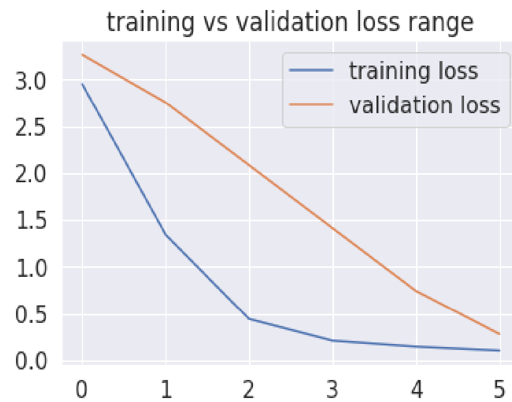

Figure 9. Loss range of the Bi-directional GRU

5. LIMITATIONS AND FUTURE WORKS

The proposed system diagnoses only a small number of diseases. This system is only allowed to forecast illnesses using different disease symptoms in Bengali. It is not yet prepared to foretell diseases in a variety of languages and it will not be able to suggest treatments after diseases have been foretold. The proposed system may be enhanced to identify more diseases in future. This system might be enhanced to anticipate disease using an audio input question and to prescribe appropriate precautions and medications based on the disease.

6. CONCLUSION

The proposed system is used to diagnose a disease based on a description of symptoms in the Bengali language. Disease prediction based on symptoms is essential because most people are unaware of numerous sickness signs and symptoms. The proposed system exhibits significant accuracy using both sequential models. The Bi-LSTM and Bi-GRU networks will be used to learn each symptom and the activations of the intermediate layers will be normalized to increase the network's stability during training. Additionally, each step of the training process will be carried out separately, resulting in a learning rate that is faster than that of any model. The patients describe their sickness symptoms in the Bengali language. The proposed system removes irrelevant text from symptoms description. The filtered text is sent to the pre-trained models that predict the disease that closely matches the given symptoms. The proposed work has utilized Bi-LSTM and Bi-GRU separately. Both Bi-LSTM and Bi-directional GRU have produced promising results. The accuracy Bi-LSTM exhibits 97.85% whereas Bi-directional GRU shows 99.73% accuracy.

ACKNOWLEDGMENT

The authors extend deepest regards and heartfelt gratitude to the Computer Science and Engineering department, NIT Durgapur, India for providing us with their valuable support and necessary help in publication of this research paper.

References

[1] H.J.J.o.I.S. Elfaik, "Deep bidirectional lstm network learning-based sentiment analysis for arabic text," De Gruyter: Journal of Intelligent Systems, vol. 30, pp. 395-412, 2021.

[2] Z. Tasnim, S. Ahmed, A. Rahman, J.F. Sorna, M. Rahman, "Political Ideology Prediction from Bengali Text Using Word Embedding Models," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, 2021, pp. 724-727.

[3] W. Akanda, A. Uddin, "Multi-Label Bengali article classification using ML-KNN algorithm and Neural Network," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE, 2021, pp. 466-471.

[4] T. Parvin, M.M.J.P.C.S. Hoque, "An Ensemble Technique to Classify Multi-Class Textual Emotion," vol. 193, 2021, pp. 72-81.

[5] M.R. Hossain, M.M. Hoque, N. Siddique, I.H.J.E.S.w.A. Sarker, "Bengali text document categorization based on very deep convolution neural network," vol. 184, 2021.

[6] S. Yeasmin, R. Kuri, A.M.H. Rana, A. Uddin, A.S.U. Pathan, H. Riaz, "Multi-category Bangla News Classification using Machine Learning Classifiers and Multi-layer Dense Neural Network," International Journal of Advanced Computer Science and Applications, vol. 12, 2021, pp. 757-767.

[7] M.F. Mridha, M.S. Rahman, A.Q. Ohi, "Human Abnormality Detection Based on Bengali Text," 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 1102-1105.

[8] M.R. Hossain, M.M. Hoque, M.A.A. Dewan, N. Siddique, M.N. Islam, I.H.J.I.A. Sarker, "Authorship Classification in a Resource Constraint Language Using Convolutional Neural Networks," vol. 9, 2021, pp. 100319-100338.

[9] M. Habib, M. Faris, R. Qaddoura, A. Alomari, H.J.I.A. Faris, "A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context," vol. 9, 2021, pp. 85690-85708.

[10] S. Mukherjee, K. Sarkar, "Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas,", 2020 IEEE Calcutta Conference (CALCON), IEEE, 2020, pp. 444-450.

[11] https://www.kaggle.com/ritheshsreenivasan/clinical-text classification/data

[12] H.E. Rifai, L.A. Qadi, A. Elnagar, "Arabic Multi-label Text Classification of News Articles," International Conference on Advanced Machine Learning Technologies and Applications, Springer, 2021, pp. 431-444.

[13] L. Liang, X. Ji, F. Ren, "Attention-based bi-lstm-crf network for emotion cause extraction in texts," 2020 IEEE International Conference on Mechatronics and Automation (ICMA), IEEE, 2020, pp. 1670-1675.

[14] F. Islam, M.M. Alam, S.S. Hossain, A. Motaleb, S. Yeasmin, M. Hasan, R.M. Rahman, "Bengali fake news detection," 2020 IEEE 10th International Conference on Intelligent Systems (IS), IEEE, 2020, pp. 281-287.

[15] Z. Lian, B. Liu, J.J.I.A.T.o.A. Tao, Speech,, L. Processing, "CTNet: Conversational transformer network for emotion recognition," 29 (2021) 985-1000.

[16] A.H. Kamal, M.A. Jishan, N. Mansoor, "Textmage: The automated bangla caption generator based on deep learning," 2020 International Conference on Decision Aid Sciences and Application (DASA), IEEE, 2020, pp. 822-826.

[17] R. Ul Haque, M.F. Mridha, A.K. Saha, M.A. Hamid, K. Nur, "Identification of Extreme Guilt and Grave Fault in Bengali Language," Proceedings of the International Conference on Computing Advancements, 2020, pp. 1-5.

[18] N.R. Bhowmik, M. Arifuzzaman, M.R.H. Mondal, M.J.N.L.P.R. Islam, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary," 1 (2021) 34-45.

[19] A.K. Das, A. Al Asif, A. Paul, M.N.J.J.o.I.S. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," 30 (2021) 578-591.

[20] T. Bai, Y. Ge, S. Guo, Z. Zhang, L.J.I.A. Gong, "Enhanced natural language interface for web-based information retrieval," 9 (2020) 4233-4241.

[21] R. Joshi, R. Karnavat, K. Jirapure, R. Joshi, "Evaluation of deep learning models for hostility detection in hindi text," 2021 6th International Conference for Convergence in Technology (I2CT), IEEE, 2021, pp. 1-5.

[22] M.S. Hossen, A.H. Jony, T. Tabassum, M.T. Islam, M.M. Rahman, T. Khatun, "Hotel review analysis for the prediction of business using deep learning approach," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021, pp. 1489-1494.

[23] J. Carriere, H. Shafi, K. Brehon, K. Pohar Manhas, K. Churchill, C. Ho, M.J.F.i.a.i. Tavakoli, "Case Report: Utilizing AI and NLP to assist with healthcare and rehabilitation during the COVID-19 pandemic," (2021) 2.

[24] M.E. Basiri, S. Habibi, "Review helpfulness prediction using convolutional neural networks and gated recurrent units," 2020 6th International Conference on Web Research (ICWR), IEEE, 2020, pp. 191-196.

[25] P. Li, A. Luo, J. Liu, Y. Wang, J. Zhu, Y. Deng, J.J.I.I.J.o.G.-I. Zhang, "Bidirectional gated recurrent unit neural network for Chinese address element segmentation," 9 (2020) 635.

[26] S. Kim, P.Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, V. Čeperić, M.J.I.T.o.N.N. Soljačić, L. Systems, "Integration of neural network-based symbolic regression in deep learning for scientific discovery," 32 (2020) 4166-4177.

[27] D.-T. Vo, R. Khoury, "Language identification on massive datasets of short messages using an attention mechanism CNN," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2020, pp. 16-23.

[28] H.H. Mohammed, E. Dogdu, A.K. Görür, R. Choupani, "Multi-Label Classification of Text Documents Using Deep Learning," 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 4681-4689.

[29] O. Oyebode, C. Ndulue, D. Mulchandani, B. Suruliraj, A. Adib, F.A. Orji, E. Milios, S. Matwin, R.J.J.o.H.I.R. Orji, "COVID-19 pandemic: Identifying key issues using social media and natural language processing," (2022) 1-34.

[30] I. Chatterjee, J. Gwan, Y.J. Kim, M.S. Lee, M. Cho, "An NLP and LSTM Based Stock Prediction and Recommender System for KOSDAQ and KOSPI," International Conference on Intelligent Human Computer Interaction, Springer, 2020, pp. 403-413.

[31] M. Schuster, K.K.J.I.t.o.S.P. Paliwal, "Bidirectional recurrent neural networks," 45 (1997) 2673-2681.

[32] J. Chung, C. Gulcehre, K. Cho, Y.J.a.p.a. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," (2014).

Appendix

The steps of the proposed system has been given below.
Step 1: Submission of user's symptoms

In Bengali, the patient describes his or her illness as input to the proposed system. The following is an example of a query.

Input = "আমার ঠান্ডা লাগছে, বমি লাগছে, ক্লান্তি, মাত্রাতিরিক্ত জ্বর এবং মাথাব্যথা আছে।". (amar tʰanda lagcʰe, bomi lagcʰe, klant̪i, mat̪rat̪irikt̪o ɟɔr eboŋ mat̪ʰabæt̪ʰa acʰe)

Step 2: tokenization of user's symptoms

The symptoms description entered by the user is tokenized and kept in a list.

Token_list = ["আমার", "ঠান্ডা", "লাগছে", ",", "বমি", "লাগছে", ",", "ক্লান্তি", ",", "মাত্রাতিরিক্ত", "জ্বর", "এবং", "মাথাব্যথা", "আছে", "।"] (["amar", "tʰanda", "lagcʰe", ",", "bomi", "lagcʰe", ",", "klant̪i", ",", "mat̪rat̪irikt̪o", "ɟɔr", "eboŋ", "mat̪ʰabæt̪ʰa", "acʰe", "।"])

Step 3: Removal of punctuation marks

The punctuation marks are removed from the token list.

Token without punctuation marks = ["আমার", "ঠান্ডা", "লাগছে", "বমি", "লাগছে", "ক্লান্তি", "মাত্রাতিরিক্ত", "জ্বর", "এবং", "মাথাব্যথা", "আছে"] (["amar", "tʰanda", "lagcʰe", "bomi", "lagcʰe", "klant̪i", "mat̪rat̪irikt̪o", "ɟɔr", "eboŋ", "mat̪ʰabæt̪ʰa", "acʰe"])

Step 4: Removal of stops words

Token list after removal of stops words

Filtered token = ["ঠান্ডা", "বমি", "ক্লান্তি", "মাত্রাতিরিক্ত", "জ্বর", "মাথাব্যথা"] (["tʰanda", "bomi", "klant̪i", "mat̪rat̪irikt̪o", "ɟɔr", "mat̪ʰabæt̪ʰa"])

Step 5: Passing of the padded sequence of filtered tokens through both Bi-LSTM and Bi-GRU

The suggested approach allows users to describe their medical symptoms in up to 150 words. After removing the stop word(s), the padded sequence is passed into both Bi-LSTM and Bi-GRU.

[[8 2 3 5 6 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0]]

Step 6: Output generation

A previously trained Bi-LSTM and Bi-GRU models are used to determine the probability vector for a certain input sequence with 33 different diseases symptoms. Bi-LSTM and Bi-GRU probabilities are listed below.

Calculated Probabilities of diseases using the Bi-LSTM

[[3.0718576e-03	6.3315179e-04	4.6452018e-03
7.1735648e-03	1.7948746e-04	5.2585334e-02
3.0989149e-03	2.5011949e-02	7.2264314e-02
6.6530653e-03	2.7806337e-03	3.6711033e-04
1.3682268e-03	2.8016903e-03	7.5606978e-04
5.6596319e-03	6.6981227e-03	2.6593607e-02
3.2681905e-02	4.1824870e-02	4.8604775e-03
1.7215291e-03	9.4527371e-02	2.6555119e-02
1.2937210e-04	1.3634887e-03	4.8098706e-02
2.2045560e-02	7.4929543e-05	1.1046443e-01
3.7646595e-01	1.6630737e-02	2.1362548e-04
]]		

Calculated Probabilities of diseases using the Bi-GRU

[[1.61466189e-03	3.15410271e-03	4.84571327e-03
1.52669859e-03	5.07328041e-05	2.85809822e-02
9.75651294e-03	4.82729748e-02	1.53859716e-03
1.96136795e-02	1.31886508e-02	1.07880784e-02
6.26955996e-04	8.17974359e-02	2.24439502e-02
1.17179414e-04	2.04833783e-03	3.71934322e-04
1.00059006e-02	1.69328600e-03	9.37276147e-03
5.16746441e-05	5.46980083e-01	6.44648774e-03
5.90274744e-02	2.83650472e-03	3.62986024e-03
2.67376262e-03	2.97757238e-03	8.77324045e-02
9.24125873e-03	7.44428253e-04	6.24938402e-03
]]		

Finally, previously trained models select the vector with maximum probabilities that help to predict the probable disease.

The Bi-LSTM model in its ouput, predicts typhoid.

The Bi-GRU model in its ouput, predicts typhoid.

Kailash Pati Mandal his obtained Bachelor of Technology in Computer Science and Engineering from Bengal College of Engineering and Technology, Durgapur, India in 2006. In 2011, he received Master's degree in Computer Science and Engineering from the Jadavpur University, Kolkata, India. He is currently a part-time Researcher in Computer Science and Engineering in the field of Natural Language Processing at the National Institute of Technology (NIT), Durgapur, India.

Prasenjit Mukherjee obtained his PhD in Computer Science and Engineering from National Institute of Technology (NIT), Durgapur, India under Visvesvaraya PhD scheme Program, Ministry of Electronics and Information Technology (MeitY), Govt. of India. His research interests include machine learning, deep learning, natural language processing, knowledge

engineering, and database management systems. He has more than 15 international publications.

Souvik Ganguly obtained his Bachelor of Computer Application from Dr. B.C. Roy Engineering College, Durgapur, India in 2021. He is currently pursuing Master of Computer Application from Dr. B.C. Roy Engineering College, Durgapur, India. He is currently working as a junior technical instructor at Ardent Computech Pvt. Ltd., India.

Baisakhi Chakraborty obtained her PhD in Computer Science and Engineering from National Institute of Technology, Durgapur, India in 2011. Her research interests include knowledge systems, knowledge engineering and management, database systems, data mining, natural language processing, and software engineering. She has several researchers under her direction. She has more than 30 international publications. She has a decade of industry experience and 17 years of academic experience.