Soft Computing (2022) 26:8253–8272
https://doi.org/10.1007/s00500-022-07261-y

FOUNDATIONS

A deep learning approaches and fastai text classification to predict 25 medical diseases from medical speech utterances, transcription and intent

Yogesh Kumar[1] · Apeksha Koul[2] · Seema Mahajan[3]

Accepted: 17 May 2022 / Published online: 25 July 2022

Abstract
The article examined the deep learning models and Fastai text classification technique to predict the medical speech utterances, transcriptions, and intent to extract the 25 medicals problems. The experimental work was conducted using a large amount of data which contains 6661.wav files and one.csv file, including 13 distinct categorization fields of medical speech utterances. Each illness's exploratory data analysis demonstrated the phrase length classes and disease categorization based on the recorded speech sound of patients for each disease. The preprocessing of the task included the wordcloud consisting of all the vocabulary words having different sizes based on the number of speech utterances in each category, eliminating Nan values, verifying for duplicates, and computing the corpus and their term index. Further, features are extracted to determine the number of words in each category, the length of phrases, and the number of words in each phrase, followed by lemmatization and tokenization. Deep learning models such as GRU (Gated Recurrent Unit), LSTM (Long Short Term Memory), bidirectional gated recurrent unit, bidirectional long short-term memory, and Fastai classifier have been used to exact category of disease from the medical speech utterances and their textual phrases. After the assessment, it was discovered that Fastai earned the most incredible precision, recall, accuracy, and lowest loss rate by 96.89%, 95.8%, 93.32%, and 0.169, respectively. In comparison, bidirectional LSTM had achieved the highest F1 score by 95.69% to predict the medical speech utterances for each category.

Keywords Gated recurrent unit · Long short-term memory · Fastai · Lemmatization · Tokenization · Medical speech · Phrases

1 Introduction

Physicians have been using SRS (Speech Recognition System) to help with clinical documentation for a long time. They can dictate clinical notes using SRS, which converts voice into electronic text edited in real-time. When using a voice recognition solution to create a medical record, the risk of omitting vital information during a visit is reduced (Noort et al. 2021). This is particularly relevant when a variety of medical services are reported in a single visit. SRS, among other things, makes it easier to work with insurance companies. Errors in healthcare records, papers, and notes developed with speech recognition technology can be caused by medical language precision, misspelled or missing words, user accents, and various language patterns. To avoid this, the product's versatility should be expanded by incorporating

✉ Yogesh Kumar
yogesh.arora10744@gmail.com;
Yogesh.Kumar@sot.pdpu.ac.in

Apeksha Koul
apekshakoulo9@gmail.com

Seema Mahajan
mm_seema@yahoo.com

[1] Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

[2] Department of Computer Science and Engineering, Punjabi University, Patiala, India

[3] Department of Computer Engineering, Indus Institute of Technology & Engineering, Indus University, Rancharda, Shilaj, Ahmedabad 382115, Gujarat, India

dictionaries for medical specializations. Users can quickly and easily repair spelling errors using semi-autocorrecting, as well as physicians can be reminded to speak slower or louder when a voice recognition technology detects a crucial amount of incompetencies while speaking to reduce the risk of misinterpretation (Poder et al. 2018).

Speech recognition technology (Fig. 1) is beneficial in the healthcare industry because medical practitioners spend a considerable portion of their day filling out paperwork. Speech recognition technology can help in this situation. It takes time to write or type notes, but speaking them out loud is faster. They have to talk into a recording system, and speech recognition technology can convert what they say into written words (Santosh 2019). This can be very useful for many physicians who struggle to find time to complete their paperwork.

Using the speech recognition solution, clinicians can easily record their voices and have diagnoses, medical history, and other relevant notes recorded without taking time after hours or minimizing face-to-face time with patients. Simultaneously, saving time on paperwork helps doctors to spend more time with their patients, resulting in better overall patient care. Speech technology may enable clinicians to be present with their patients instead of wasting time by sorting through a medical chart or reviewing details on a computer (Zhang et al. 2019). When a clinician listens intently and communicate with their patient on a one-on-one basis, they do a higher, more personalized level of treatment. As a result, patients' well-being increases, medical personnel feel satisfied with the job, and hospitals and clinics prosper. But, for several reasons, dictating in high-traffic areas can be challenging. HIPAA (Health Insurance Portability and Accountability Act of 1996) regulations necessitate the protection of confidential medical details from an unauthorized access (Kumah-Crystal et al. 2018). Furthermore, while most speech recognition systems will filter out background noise, there is still the risk of background dialogue or the sounds of medical equipment interfering with recognition. According to a study on speech recognition in healthcare, 62% of healthcare providers currently use speech recognition technology for their records, and 4% implement medical speech recognition in EHRs (Electronic Health Records) (Johnson et al. 2014).

The present study focuses on classifying 25 diseases based on medical voices and their associated textual phrases. The purpose of this study is to determine whether deep learning models can be used to validate medical speech utterances, transcriptions, and intent to extract medical issues. Long short-term memory, bidirectional long short-term memory, gated recurrent unit, and bidirectional gated recurrent unit have all been created to detect illnesses using medical discourse. For the study in this article, one.csv (comma separated value) file and 6661 wav (waveform audio) files are used, which are then preprocessed to extract characteristics through lemmatization and tokenization. The acquired features are displayed in terms of the number of words in each category, the length of phrases, and the number of words in each medical speech phrase. Following the evaluation, it has been determined that Fastai has the highest precision, recall, accuracy, and lowest loss rate, with 96.89%, 95.8%, 93.32%, and 0.169, respectively, with 98% (validation accuracy) and 0.99 (validation loss), while bidirectional LSTM has the highest F1 score, with 95.69%.

The remaining paper is structured as follows: Sect. 2 reviews the literature and available datasets on medical speech, transcription, and purpose prediction. Section 3 summarizes the contributions made in this study. Section 4 describes the materials utilized in this study in-depth, including the platform and libraries, dataset collection, preprocessing, feature extraction, and applied deep neural network algorithms, as well as their evaluation parameters. Section 5 includes the proposed work's experimental findings and analyses. Section 6 contains the study's discussion, and Sect. 7 closes the paper with an explanation of

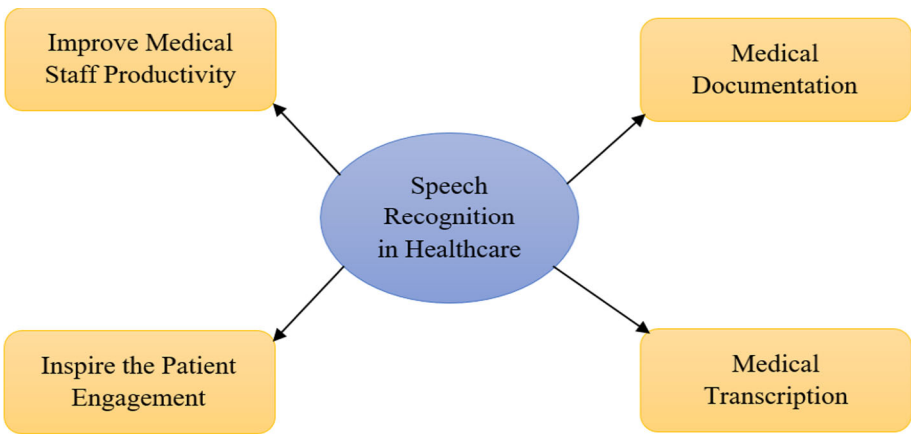


Fig. 1 Speech recognition in healthcare

the acquired results and their comparison to baseline techniques with room for improvement.

2 Related work

This section discusses how voice recognition technology is used in the healthcare industry. Based on the current work, we know a lot more work can be done using deep learning models for medical speech, transcription, and intent prediction. In (Ismail et al. 2020), a novel approach based on speech recognition has been proposed for providing an easy-to-use control device to the aged, ill, and incapacitated. The goal was to create a low-cost speech recognition system that allows users to easily access Internet of things (IoT) devices that have been installed in smart homes and hospitals without the need for a centralized supervisory system. A Raspberry Pi board was used in the proposed method to enable wireless charging of domestic items via mobile phones. The main objective was to leverage IoT connections driven by speech commands to speed up interactions between customers and household equipment. The recommended platform contribution enhances speech recognition by integrating a SVM (Support Vector Machine) with a DTM (Dynamic Time Warping) algorithm. The proposed solution is a machine learning-based system that can monitor smart devices with 97% when spoken instructions are utilized. In (Latif et al. 2021), the authors addressed the open issues and proposed some research paths to completely exploit the benefits of other technologies for the efficiency of speech-based healthcare solutions. The authors also gave an overview of various barriers to the growth of voice-based health care services. Authors had also conducted a study of current research in Blackley et al. (2019) on the use of speech recognition technology for clinical recording in order to gain a better understanding of the impact of speech recognition on document quality, provider efficiency, and institutional expenditure, among other factors. While speech recognition has been used extensively for clinical testing, this research remains largely heterogeneous, sometimes using various assessment criteria with mixed findings. In addition, further study of the use and efficacy of SR-assisted documentation in clinical settings is warranted outside radiology. The opportunities of speech technology in Netherlands hospitals and the use of speech for medical documentation have been examined (Luchies et al. 2018). In addition, the authors explained why the Dutch hospital workers still marginally used speech technology. The authors then conducted interviews, in which participants included users of speech technology, hospital managers, and software suppliers. The authors transcribed the interviews and summarized the advantages and disadvantages of language technology and key obstacles to adoption. Then, the findings showing different factors clarify that only 1% of medical personnel in the Netherlands use speech technology. In (Sonal and Dodiya 2016), the authors discussed and investigated the effects of these systems for the medical domain. They examined the features and implementations of automatic speech recognition systems. Since medical issues are very long and complicated, the findings showed that this domain differs from the open domain and needs more work to be adjusted according to the field while automatically speaking**.** The authors in Vij and Pruthi (2018) proposed data science applications in healthcare focused on examining different feelings and methodologies of dynamic analysis. The authors presented the complexity of the feeling and emotional examination of patients' medical record visualization and medical history. Therefore, there has been a collective vision of data mining, data analytics, and visualization in healthcare. In (Alhussein and Muhammad 2018), the authors used deep learning to detect speech pathology in mobile healthcare. In the speech pathology detection system, voices were gathered using smart mobile devices. Voice inputs were analyzed until a co-evolutionary neural network was used. The authors utilized the methods VGG-16 and CaffeNet to apply current stable CNN models. In Saarbrucken, the database was used in voice problems research. In (Nassif et al. 2019)**,** the authors used the assessment method that had been extracted from a comprehensive statistical study of the use of profound education in speech applications. Most researchers found that their device efficiency was determined by the use of WER (Word Error Rate). Preliminary research aimed at clarifying the use of CNN in voice pathology detection was carried out (Mohammed et al. 2020). They worked on a deep CNN (Convolutional Neural Network) learning system for detecting speech pathologies that had a high prediction accuracy of up to 94.54% in training and 95.41% in research. The automatic Punjabi-language spontaneous language recognition has been proposed in Kumar et al. (2021). The vast Punjabi text corpus vocabulary was collected through interviews, presentations, and other sources with sizes that match natural voice recognition. According to research, the 2,073,456 unique combinations of Punjabi tri-phonemes in the lexicon make up 231 phonemes. The suggested automated models for spontaneous speech included 13,218 Punjabi words and more than 200 min of registered speech. Punjabi qualifying sentences and word accuracy increased to 94.19% for 13,218 Punjabi phrases in 2381, bringing the suggested model's performance to 87.10%. In (Takao et al. 2018), the authors had established a new reporting system based on standardized data entry, which selectively only extracts endoscopic findings from oral statements of gastroenterologists and automatically

enters them into appropriate columns on an endoscopic basis in real-time. The proposed device showed that gastroenterologists could complete the report efficiently in real-time by endoscopic procedures. The role of three Netherlands systemic intermediaries in agriculture, energy development, and health care were analyzed in Lente et al. (2020). The authors found that structural intermediaries work with the field pledge and have to reposition themselves as a result. They both profit and experience the dilemma between revolutionary structural changes and continuing them in various phases. Table 1 represents the comparative analysis of the work done by the researchers.

3 Contribution

Working on Fastai and deep learning models to validate the medical speech utterances, transcriptions, and intent to extract the 25 medicals problems. The study concerns the large amount dataset having the speech sounds and description file for each sound unit. This allows us to have various objectives: The deep learning models include GRU, bidirectional GRU, LSTM, bidirectional LSTM and Fastai for prediction of exact category of disease from the medical speech utterances and their textual phrases.

(a) As no state-of-the-art works provide a large amount of data for validation of different types of medical speech utterances, the dataset contains the 6661.wav files and one.csv file, which includes the 13 other columns for classification. This brings us a solution to quantify/test deep learning models
(b) Further applied deep learning models have been compared based on training accuracy, training loss, validation accuracy, validation loss, precision, recall, and F1 score.

The proposed approach, as depicted in Fig. 2, works by collecting medical speech utterances dataset, their exploratory data analysis, preprocessing tasks, feature extraction, which includes the lemmatization and

Table 1 Comparative analysis

References	Dataset	Technique	Outcome	Limitations
Suominen et al. 2016)	NICTA synthetic nursing handover data	Multiclass classification, statistical techniques	F1 score = 81%	The clinical features of many nurses traveling between patient locations hampered the system, resulting in a loud, minimally customized multi-speaker
Akinloye et al. 2020)	Data collected from ASD patient	Wearable emotional-basede-Healthcare controller, Speech Recognizable Sensor	Accuracy = 92.07%	The model could not be generalized to all patients without taking into account their preferences and dislikes
Lam et al. 2020)	Elderly data Resource data Staff data	Face Recognition Technology, Intelligent mHealthcare System	Confident scores = 77% 87% 92%	Real-time data monitoring and automated care should be integrated into the ImHS to enable personalized healthcare services
Paulett and Langlotz 2009)	2 million de identified radiology reports	Trigram Model, n-gram model	Concordant = 0.251% ± 2.23 Discordant = 36.5% ± 5.60	The system was unable to validate itself by comparing several language models in a real-world voice recognition system
Lu et al. 2021)	Data collected from 24 young participants	Spatiotemporal approach, Maxfilter, Source Estimation Method	Spectral Peak = 0.017Cohen's d = 0.77	When evaluating the brain source result, caution should be exercised
Uddin and Nilsson 2020)	Audio speech database	Convolution Neural Network, Neural Structured Learning (NSL)	Recall = 93%	Additional deep learning techniques might be used in conjunction with NSL to improve the model's efficiency
Takao et al. 2018)	Saarbruecken Voice Database	Convolution Neural Network	Accuracy = 95.41% F1 score = 94.22%Recall = 96.13%	A single endoscopist assessed the system's speech recognition ability, taking into account the endoscopist's age, dialect, voice loudness and articulation, as well as individual speaking speeds

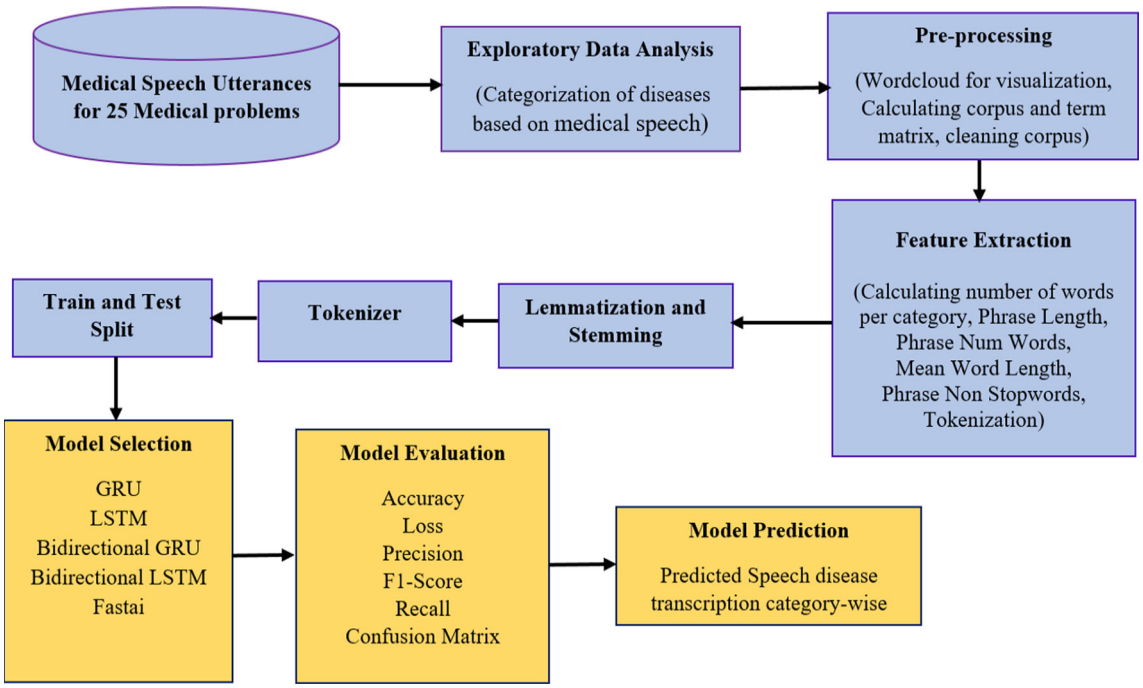


Fig. 2 Proposed framework of the system

stemming, tokenization and further train-test split, deep learning models selection, and evaluation for prediction of medical speech utterances (ref. Section 4).

4 Materials and methods

This part contains the libraries imported, information regarding the dataset, strategies utilized during preprocessing, feature extraction along with the used algorithms to assess the performances.

4.1 Platforms and library

Several Python libraries are utilized to effectively implement the suggested medical voice and transcription categorization algorithms, including TensorFlow, Keras, Wordcloud, FastAI, Soundfile, Pandas, Seaborn Matplotlib, Nltk, spaCy, and Librosa. The Fastai program simplifies the process of rapidly and accurately training neural networks using modern best practices. On the other hand, TensorFlow is a machine learning and deep learning software package that is free and open-source. Keras is an open-source software library with a Python interface for artificial neural networks and functions as a front end for the TensorFlow library. It is helpful for various applications but is notably well-suited for the training and inference of deep neural networks (Ramasubramanian and Singh 2019). Additionally, NLTK (Natural Language ToolKit) (Mehta et al. 2020) is a well-known Python framework for creating applications that interface with human language data. It includes an intuitive user interface for over 50 corpora and lexical resources, including WordNet, as well as a suite of text processing libraries for tokenization, classification, tagging, stemming, semantic reasoning and parsing, as well as wrappers for industrial-strength natural language processing libraries. Additionally, wordcloud (Jayashankar and Sridaran 2017) is used, as is Soundfile–an audio library based on libsndfile, CFFI (C Foreign Function Interface), and NumPy. To read or write sound files directly, use the read() and write() functions () (Lazzarini 2019). Matplotlib is a Python library for creating static, animated, and interactive visualizations. spaCy is a free and open-source Python package for natural language processing. Among other things, it incorporates NER (Named Entity Recognition), POS (Parts of Speech) tagging, dependency parsing, and word vectors. Additionally, librosa is a Python package that enables the analysis of music and audio. It contains the building blocks necessary for the development of music information retrieval systems (Zisad et al. 2020).

4.2 Dataset collections and representations

Audio utterances for general medical symptoms like "Acne" or "Ear ache" comprise the dataset, totaling more than 8 h. Individual human contributors produced each speech based on a particular symptom. In the medical

profession, such audio samples train the conversational agents. The audio utterances and transcriptions are also included in this collection. Because the dataset contains some inaccurate text labeling and poor-quality audio utterances, it was cleaned before using any deep learning or Fastai for prediction. The dataset contains the 6661.wav files and one.csv file, which includes the 13 different columns such as Audio clipping, Audio clipping: confidence, Background noise Audible, Background noise audible: Confidence, Overall audio quality, Speaker: Confidence and Id, Prompt, Phrases, Filename and Writer Id. For training, only Phrase, prompt, and filename are used because, Phrase tells us the text what Speaker Spoke, whereas Prompt tell us Medical Problem Category and File Name is used to link all the Audio files with the Phrase and prompt Column. Table 2 shows the disease-wise total count of the.wav file and percentage covered per category.

Table 2 Disease-wise medical speech utterance count

Category	Total count	Total percentage covered in dataset per category
Acne	328	4.924
Back pain	259	3.888
Blurry vision	246	3.693
Body feels weak	241	3.618
Cough	293	4.399
Ear ache	270	4.053
Emotional pain	231	3.468
Feeling cold	263	3.948
Leg pain	283	4.249
Foot ache	223	3.348
Hair falling out	264	3.963
Hard to breath	233	3.498
Head ache	263	3.948
Heart hurts	273	4.098
Infected wound	306	4.594
Injury from sports	230	3.453
Internal pain	248	3.723
Joint pain	318	4.774
Knee pain	251	4.579
Muscle pain	282	4.234
Neck pain	251	3.768
Open wound	208	3.123
Shoulder pain	320	4.804
Skin issue	262	3.933
Stomach ache	261	3.918

Figures 3 and 4 show the phrase length classes in which diseases are dispersed based on recorded medical speech for each condition in the EDA (Exploratory Data Analysis). The EDA emphasizes a different person's recorded voice utterance for each of the 25 diseases, including injury, shoulder discomfort, back pain, and others.

4.3 Preprocessing

The wordcloud has been generated (Fig. 5) based on the different type of disease-wise speech utterances which simplify the specific word appears in the sources of medical transcription and intent data file. The medical speech wordcloud consists of all the vocabulary words having different sizes based on the number of speech utterances in each Category. The bigger the word appears, the more often it is mentioned within a given text and the more critical it is. Using the wordcloud library generates the wordcloud for spoken medical utterances of different diseases in python.

Further cleaning data for preprocessing includes removing the Nan values, checking duplicate values, calculating corpus and term index. Whereas corpus is defined as the collection of all the text values in the dataset (as shown in Table 3), term index is the word count in a matrix format that calculates the number of words available in each row.

4.4 Feature extraction

During feature extraction, first, calculate the number of words per category with the percentage covering the whole dataset. Then find out the phrase length by counting the total number of lines and phrase number of words per line. In mean word length, computed the mean of the phrases length and phrase non-stop words shows the words only without any stopping signs like ! @, #, $, %, etc. Table 4 shows the complete process for extracted features. The library is used to gather together the multiple inflected forms of a word so that they may be examined as a single item to conduct the text cleaning lemmatizer from the WordNetLemmatizer(). As depicted in Fig. 6, inflection is the change of a word to communicate many grammatical categories such as tense case, voice, aspect, person, number, gender, and mood. With a prefix, suffix, or infix, or another internal modification such as a vowel shift, an inflection conveys one or more grammatical categories.

The text cleaning process also includes using a tokenizer to divide a large size sample of text into words. This is necessary for natural language processing jobs in which each word must be recorded and given to additional analysis such as classification and counting for a specific mood. Further, convert all text to lowercase and removed all sorts

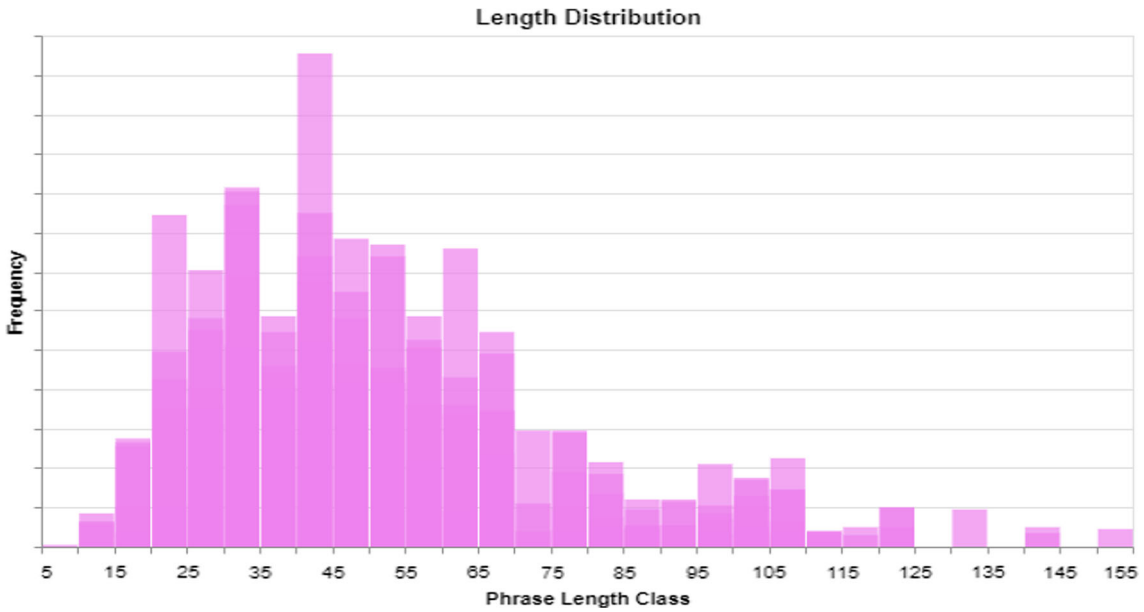


Fig. 3 Phrase length classes of medical speech utterances

of symbols in punctuations. And also, remove tokens that are not alphabetical with the filter of stop words and stem all the sentences—the outcomes of text cleaning as shown in Table 4**,** which helps to show cleaned columns phrases. The phrase is the exact line spoken by the patient, and the prompt is the same word we need to train based on the phrase words, phrase length. Whereas output of the cleaned column after the cleaning column phrases is different from the original terms, the tokenizer library provides a unique token to each word in the database. As a result, the total numbers of unique tokens generated are 1070. The dataset is split into a training set and a testing set as a result of tokenization. The training dataset comprises 75% of the data, whereas the testing dataset contains 25%.

4.5 Applied deep learning models

Various deep learning models such as long short-term memory, bidirectional LSTM, gated recurrent unit, bidirectional GRU, and Fastai have been applied to predict the medical speech utterances for different types of illness:

- ***LSTM***

The LSTM (Long Short-Term Memory) architecture is a type of RNN (Recurrent Neural Network) used in deep learning. LSTMs have feedback connections, unlike traditional feed forward neural networks. It can process not only single data points (such as photos), but also entire data streams (such as speech or video).

A typical LSTM unit (Fig. 7) consists of four components: a cell, an input gate, an output gate, and a forget gate. The equations are used to evaluate all of them Eqs. (1, 2, 3) The three gates control the entry and outflow of information, and the cell holds data across arbitrary time intervals (Mohamed et al. 2015). LSTM networks are ideally suited for categorizing, analyzing, and forecasting time series data, as there may be unpredictable delays between significant occurrences in a time series. LSTMs were created to address the vanishing gradient problem that can occur when standard RNNs are trained.

$$I_T = \sigma(W_I[h_{T-1}, X_t] + b_I) \quad (1)$$

$$F_T = \sigma(W_F[h_{T-1}, X_t] + b_F) \quad (2)$$

$$O_T = \sigma(W_O[h_{T-1}, X_t] + b_O) \quad (3)$$

here I_T input gate, F_T forget gate, O_T output gate, σ sigmoid function, W_X the weight of all gates, X neurons, h_{T-1} output of the previous LSTM block at timestamp (T-1), X_T input at current timestamp, and b_X biases for respective gates.

- ***Bidirectional LSTM***

Bidirectional LSTMs (Fig. 8) are a variant of standard LSTMs that can significantly enhance model performance when used for sequence classification tasks. Bidirectional LSTMs train two LSTMs on the input sequence rather than one. The first is performed on the input sequence in its entirety, while the second is performed on a reversed duplicate of the input sequence. This can offer extra context for the network, resulting in faster and more complete issue learning.

In the realm of voice recognition, rather than a linear interpretation, the context of the entire utterance is employed to understand what is spoken. As a result, the

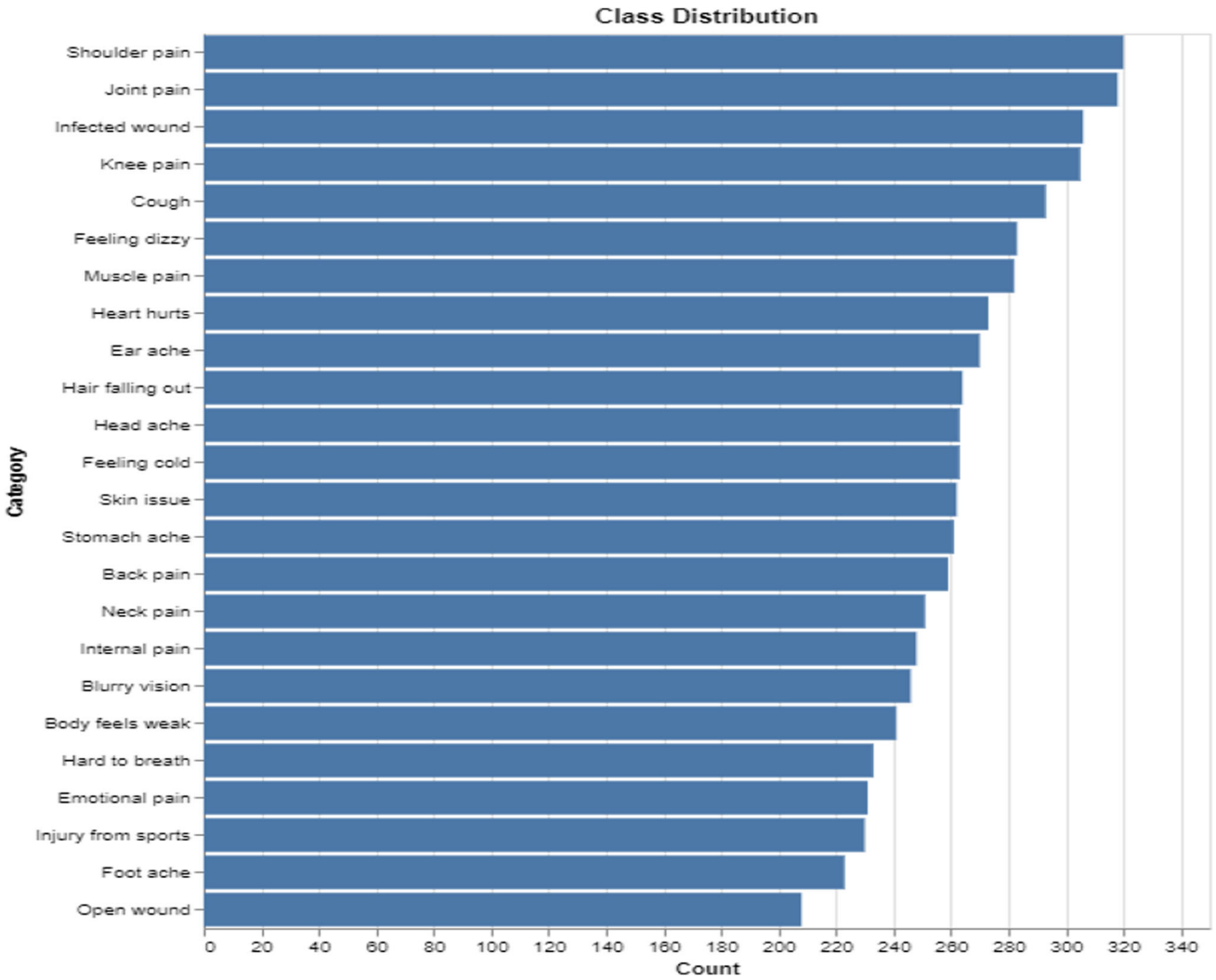


Fig. 4 Categorization of diseases based on medical speech

input sequence is supplied in both directions. To be exact, each time step in the input sequence is processed separately, yet the network traverses the sequence simultaneously in both directions (Graves et al. 2013).

- ***GRU***

GRUs is a gating technique used in recurrent neural networks. The GRU (Fig. 9) is similar to a long short-term memory with a forget gate, but requires fewer parameters due to the absence of an output gate. The performance of GRU was shown to be comparable to that of LSTM for specific tasks in speech signal modeling, and natural language processing. On some smaller and less frequent datasets, GRUs has been proven to perform better (Dey and Sale 2017).

Another interesting feature of GRU is that, unlike LSTM, it lacks a distinct cell state (Ct). It possesses solely a concealed state (HT). GRUs are faster to train due to their simplified design. GRU has three distinct gates: an Update Gate (Z), a Reset Gate (R), and a current memory gate which are defined in Eqs. (4, 5, 6, 7)

$$Z_t = \sigma(W_z.[H_{t-1}, X_t]) \tag{4}$$

$$R_t = \sigma(W_z.[H_{t-1}, X_t]) \tag{5}$$

$$\tilde{H_t} = \tanh(W.[R_t * H_{t-1}, x_t]) \tag{6}$$

$$H_t = (1 - Z_t) * H_{t-1} + Z_t * \tilde{h_t} \tag{7}$$

where Z and R represent the update and reset gates, respectively, while $\tilde{H}$ and H represent the intermediate memory and output, respectively.

- ***Bidirectional GRU***

Bidirectional GRUs, or BiGRUs, consist of two GRUs. One receives input in a forward direction, while the other receives information in a backward manner. Figure 10

Fig. 5 Wordcloud for visualizations of different disease

Table 3 Phrases with possible prompts

Text value	Phrase	Prompt
0	I feel down when I remember her	Emotional pain
1	I feel like breaking my back when I carry heavy things	Back pain
2	when I move my arm there is too much pain	Muscle pain
3	My son pierced his lip and the skin inside his lip is grey and swollen and looks infected	Infected wound
4	Lower back muscles are aching	Back pain
5	At my back leg I have muscle pain	Leg pain
6	In my left leg I have muscle pain	Leg pain
6652	I feel depressed when I see my hair falling out	Hair falling out
6653	I have shoulder pain when I try to carry my groceries	Shoulder pain
6654	I have a cut that has become red and oozes puss	Infected wound
6655	I have a ear ache when I go to the pool	Ear ache
6656	I feel a burning sensation in my guts about 2 h after each meal	Stomach ache
6657	I have a split on my thumb that will not heal	Open wound
6658	I feel a lot of pain in the joints	Joint pain
6659	The area around my heart doesn't feel good	Heart hurts
6660	I complain a lot with skin allergy	Skin issue

illustrates the architecture of bidirectional gated recurrent unit.

The bi-GRU model is composed of two unidirectional GRUs facing each other. One GRU advances forward, starting at the beginning of the data series, and the second GRU advances backward which starts from the end of the sequenced data (Abdelgwad et al. 2021). This technique enables knowledge to come from future as well as past to have an effect on present situations. In Eqs. (8, 9, 10), the bi-GRU is defined as follows:

Table 4 Text cleaning attributes

No	Phrase	Prompt	Phrase_length	Phrase_num_words	Mean_word_len	Phrase_non_Stopwords	Cleaned_phrase
0	I feel down when I remember her	Emotional pain	31	7	3.571	5	remember feel
1	I feel like breaking my back when I carry heavy things	Back Pain	54	11	4	9	carry heavy thing feel like breaking
2	when I move my arm there is too much pain	Muscle Pain	41	10	3.2	2	pain arm
3	My son pierced his lip and the skin inside his lip is grey and swollen and looks infected	Infected wound	103	18	3.727	11	son lip pierced swollen skin inside lip grey look infected
4	Lower back muscles are aching	Back Pain	38	5	3.875	4	muscle lower aching
5	At my back leg I have muscle pain	Leg Pain	65	8	3.4	7	muscle pain muscle pain leg
6	In my left leg I have muscle pain	Leg Pain	33	8	3.25	5	muscle pain left leg
7	I have to apply pain relief cream because I have cut my finger while playing football but it does not help	Injury from sports	107	21	3.909	11	cut finger playing football apply pain relief cream help
8	I have many problems in my derma like itching and acne in my face	Skin issue	66	14	3.786	7	acne face problem derma like itching
9	My arm has strange rash	Skin issue	31	6	3	4	strange rash arm
6656	I feel a burning sensation in my guts about 2 h after each meal	Stomach ache	68	14	3.928571	8	feel burning sensation gut hour meal
6657	I have a split on my thumb that will not heal	Open wound	46	11	3.272727	4	split thumb heal
6658	I feel a lot of pain in the joints	Joint pain	35	9	3	5	feel lot pain joint
6659	The area around my heart doesn't feel good	Heart hurts	43	8	4.5	6	area heart feels good
6660	I complain a lot with skin allergy	Skin issue	33	6	4.666667	5	complain a lot skin allergy

$$\overrightarrow{H_t} = \mathrm{GRU}_{\mathrm{fwd}}(X_t, \overrightarrow{H_{t-1}}) \quad (8)$$

$$\overleftarrow{H_t} = \mathrm{GRU}_{\mathrm{bwd}}\left(X_t, \overleftarrow{H_{t+1}}\right) \quad (9)$$

$$H_t = \overrightarrow{H_t} \oplus \overleftarrow{H_t} \quad (10)$$

where $\overrightarrow{H_t}$ is the state of the forward GRU, $\overleftarrow{H_t}$ is the state of the backward GRU, $\oplus$ indicates the operation of concatenating two vectors.

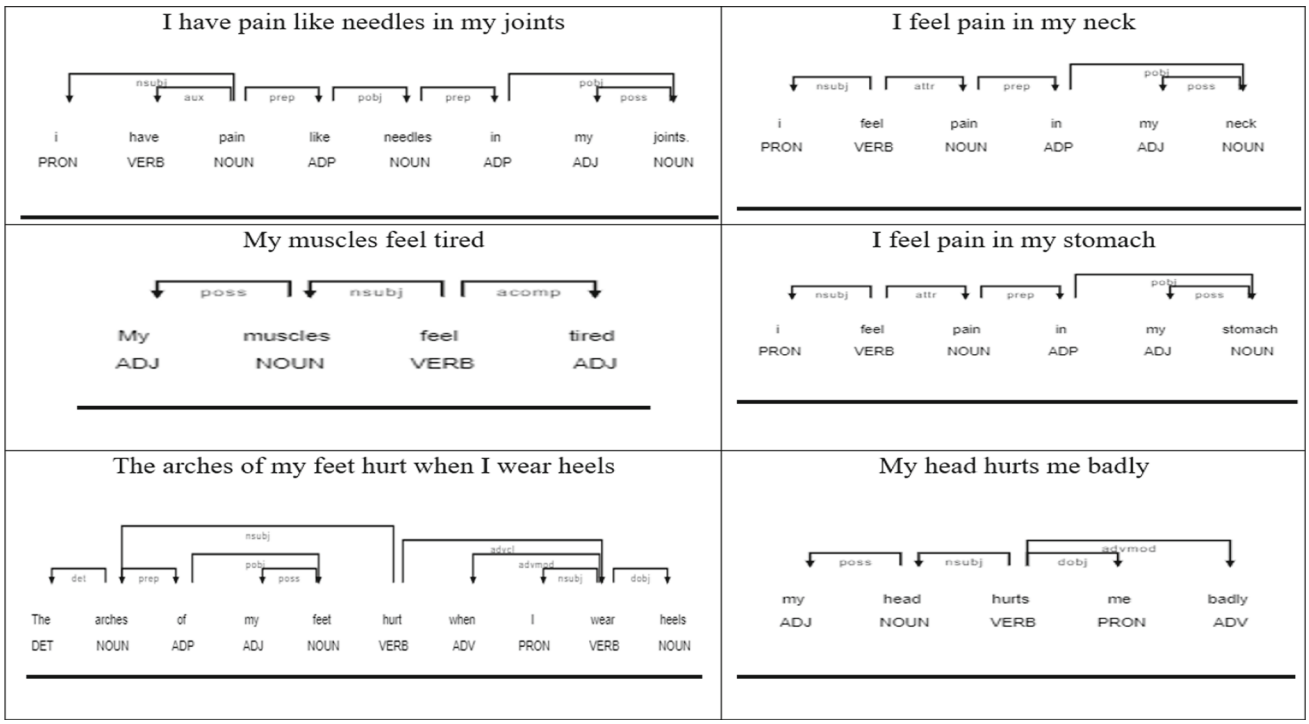
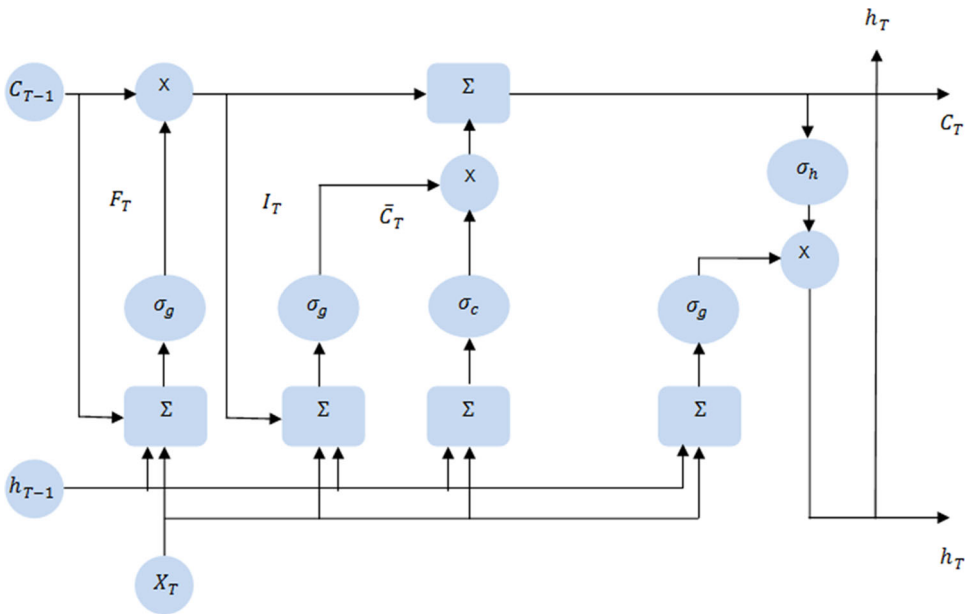


Fig. 6 Word lemmatizer for medical speech utterances

Fig. 7 LSTM architecture



- ***Fastai***

A deep learning-based library, Fastai offers high-level configuration that enables researchers to rapidly and efficiently get advanced results in common deep learning domains, and those who work with low-level features can combine as well as match to create novel techniques. Fastai is designed with two primary aims: to be friendly and dynamic. It tries to do both without making significant concessions regarding ease of use, flexibility, or performance. The Fastai library is centered on using pre-trained language models and their fine-tuning, which is accomplished in the following three steps: Data preprocessing with the least amount of code possible (Louinci et al. 2021). Create a language model that includes pre-trained weights that you can fine-tune for your dataset. Add other

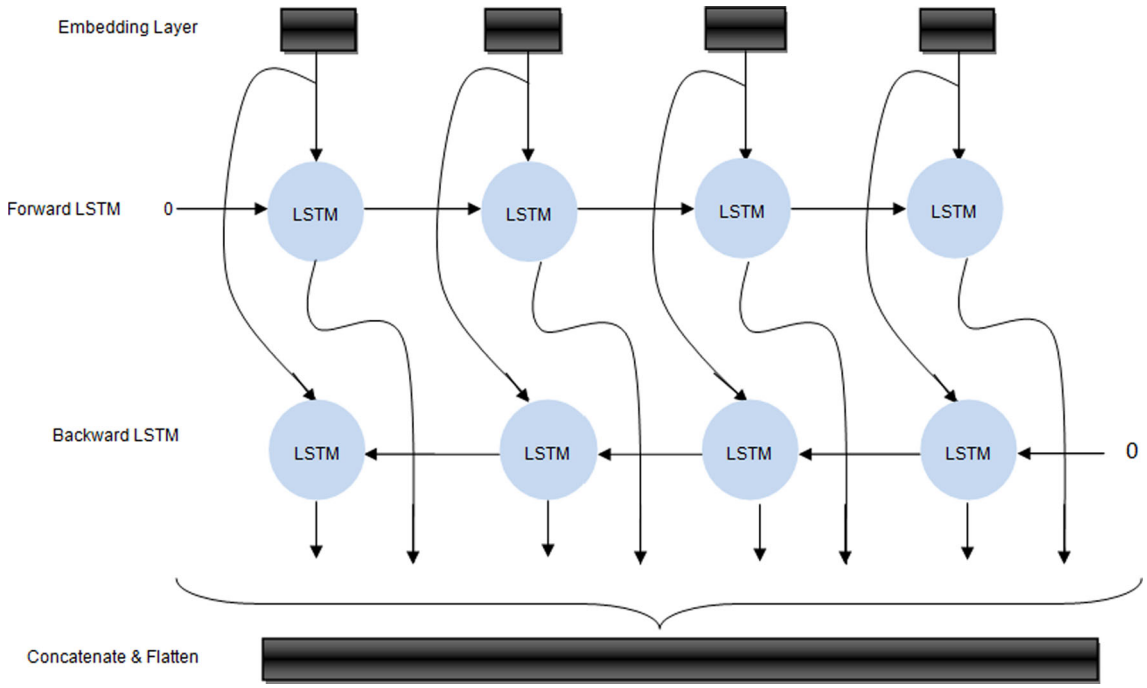
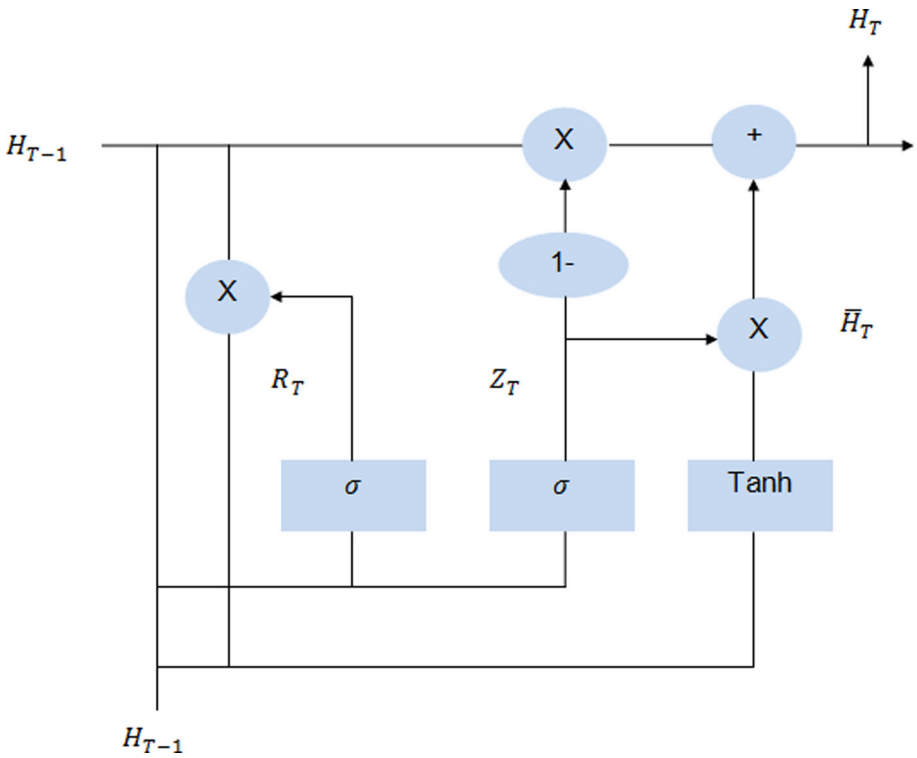


Fig. 8 Bidirectional LSTM architecture

Fig. 9 GRU architecture



models, such as classifiers, to the language model. Figure 11 outlines all of the classes and methods necessary for developing a text categorization model.

4.6 Evaluative parameters

Accuracy

It is the metric used to evaluate which model is the most effective in identifying connections and patterns among

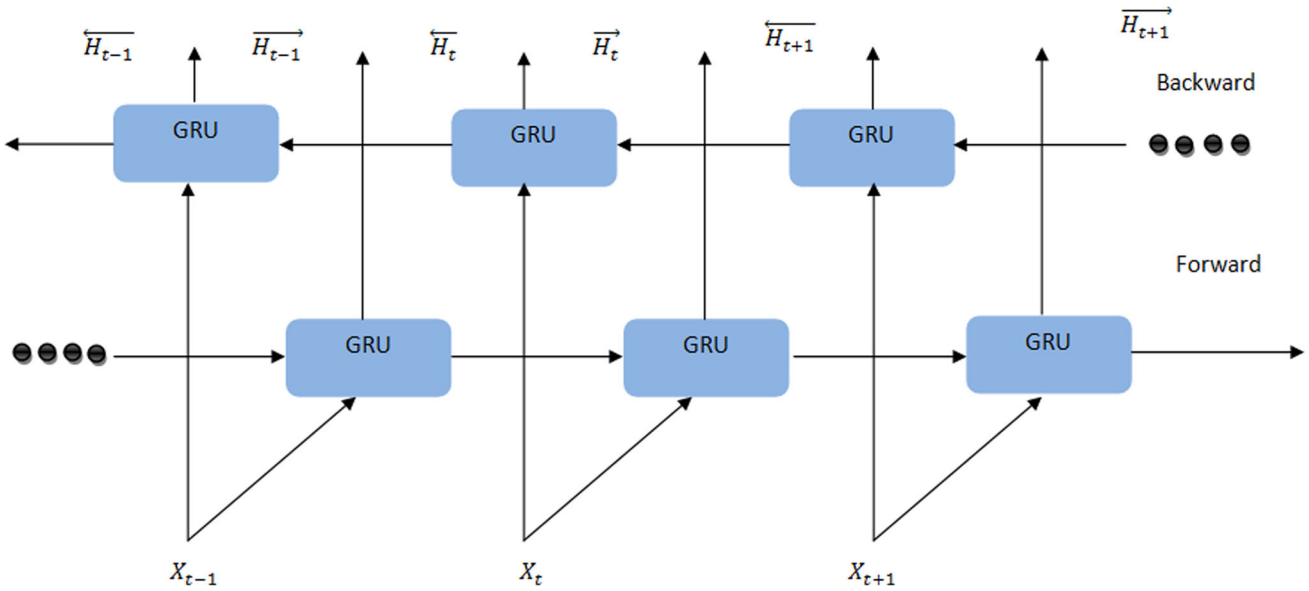


Fig. 10 Bidirectional GRU architecture

variables in a dataset using the input or training data. It is straightforward to compute by dividing correct guesses by the total number of forecasts (Shukla and Jain 2021). It is calculated using Eq. 11

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

where TP, TN, FP, FN stand for true positive, true negative, false positive, and false negative, respectively.

- *Loss*

Loss is defined as a number which indicates how wrong the model predicts. It is a kind of penalty for the wrong prediction. If the value of loss is zero, it means the model is perfect else it is imperfect. (Patil and Agashe 2021). It is calculated using Eq. 12.

$$\text{Loss} = \frac{(AV - PV)^2}{N} \tag{12}$$

where N stands for number of observations, AV stands for actual value and PV stands for predicted value.

- *Precision*

The number of actual positive results is divided by all positive effects, including those not identified correctly. Precision is also known as a positive predicted value (Al-Hassan and Al-Dossari 2021). It is calculated using Eq. 13:

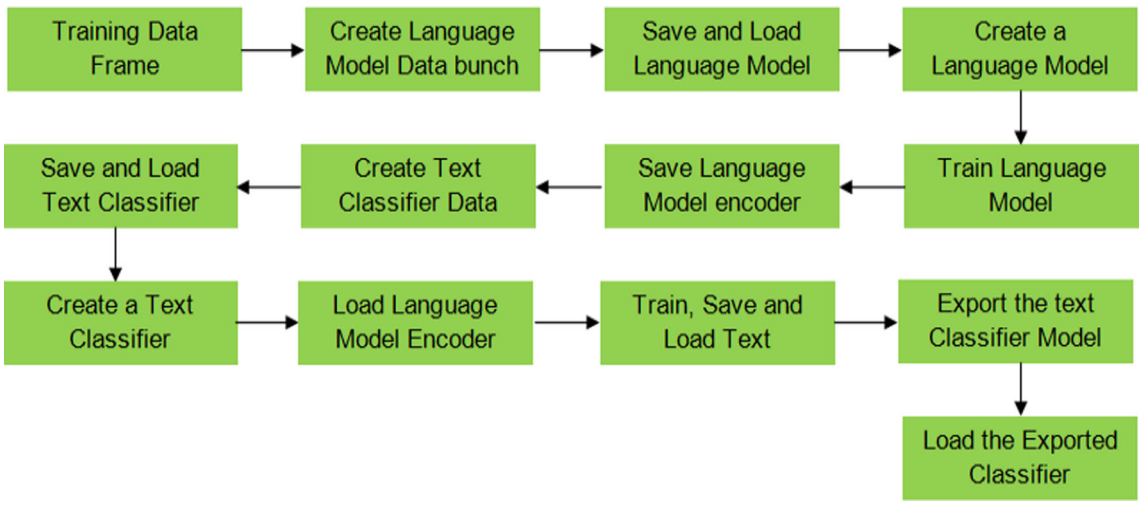


Fig. 11 Text classification model

Table 5 Accuracy versus loss scores

Algorithms	Accuracy	Loss	Validation accuracy	Validation loss
GRU	80.00	3.21	95.00	0.16
Bidirectional LSTM	92.78	0.27	96.39	0.06
LSTM	90.66	0.32	94.89	0.22
Bidirectional GRU	92.45	0.25	96.30	0.12
Fastai Text Classifier	93.32	0.169	98.00	0.09

Table 6 Average recall, precision and F1 score values

Algorithms	Precision	Recall	F1 score
GRU	86.6416	84.58	84.756
Bidirectional GRU	95.68	95.32	94.88
LSTM	94.116	93.8	93.52
Bidirectional LSTM	95.826	95.76	**95.6944**
Fastai text classifier	**96.8916**	**95.8**	95.48

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

where TP stands for true positive and FP stands for false positive.

- *Recall:*

The recall is the number of genuine positive findings divided by the total number of positive samples that should have been found. In binary classification, recall is sometimes referred to as sensitivity (Al-Hassan and Al-Dossari 2021). It is calculated with the help of Eq. 14:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

where TP stands for true positive and FN stands for false negative.

- *F1 score*

The F-score or F-measure is a statistic used in binary classification statistical analysis to determine a test's correctness. It is judged by the accuracy and recall of the exam. An F-score might be as high as 1.0, indicating perfect precision and recall, or as low as 0 (Krishnan et al. 2021). It is calculated as follows in Eq. 15:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{15}$$

where TP stands for true positive, FP stands for false positive, and FN stands for false negative.

5 Results and analysis

The results of using four models to identify speech to forecast disease are shown in this part. These models are long short-term memory, bidirectional long short-term memory, gated recurrent unit, bidirectional gated recurrent unit, and Fastai text classifier. The findings are summarized using various metrics, including accuracy, loss, precision, recall, and F1 score. Four subsections are given to illustrate the average accuracy and loss, the average precision, recall, and F1 score, confusion matrix for each deep learning model, and comparative analysis graphs for each disease category of a different model.

5.1 Average loss and accuracy

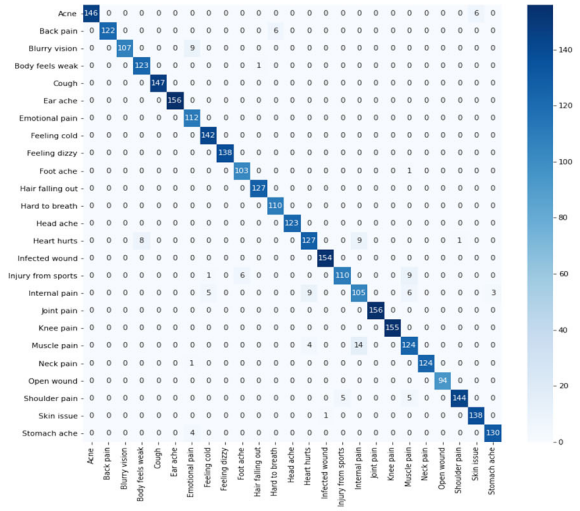
The accuracy and loss of four algorithms are given. Loss and accuracy scores are subdivided further into validation and training accuracy/loss. The classification models' training and validation accuracies and the losses they incur are listed in Table 5.

It was demonstrated that the Fastai text classifier achieved the maximum accuracy and the lowest loss value by 93.32% and 0.169, respectively. In addition to this, Fastai also emerged as the best-suited algorithm for identifying speech to forecast illnesses, with 98% accuracy and 0.09 loss when the models were tested again after being validated.

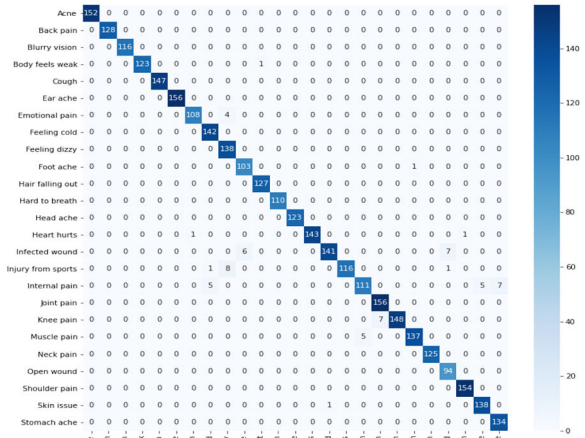
5.2 Average F1 score, precision, and recall

The average precision, recall, and F1 score of four classification methods for various aforementioned diseases (see Table 2) are listed in Table 6 as well as the best values are shown in bold font.

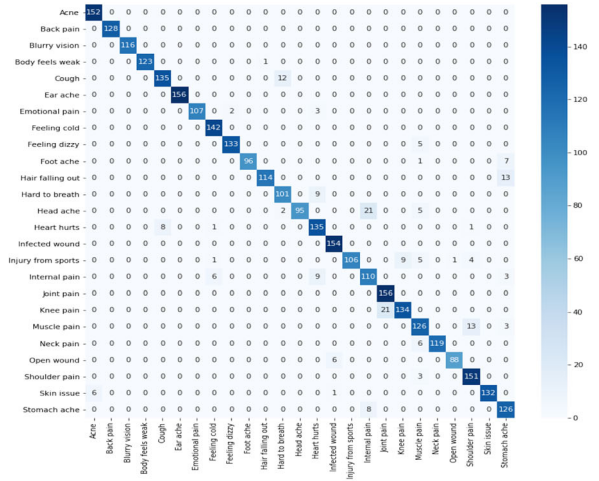
It is shown that Fastai has demonstrated the highest average precision and recall values, with 96.8% and 95.8%, respectively, while bidirectional LSTM has demonstrated the highest average F1 score value, with 95.69%. LSTM, bidirectional LSTM, GRU, Bidirectional GRU, and Fastai are used to compute the precision, recall, and F1 score value for each disease listed in Table 2.



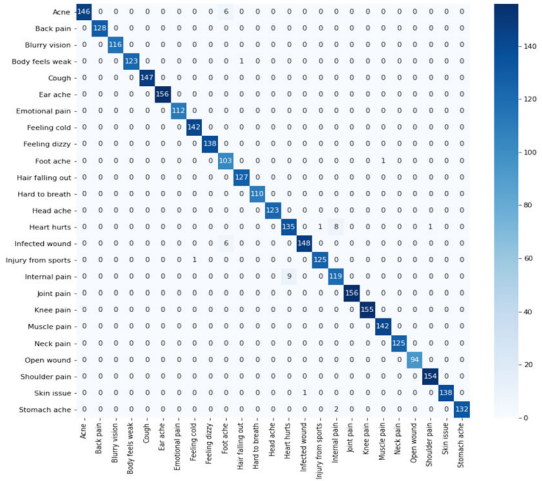
(a) LSTM



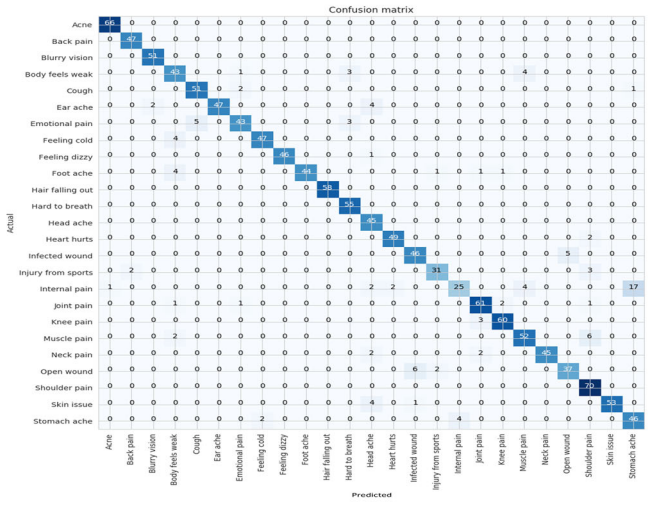
(b) GRU



(c) Bi-directional LSTM



(d) Bi-directional GRU



(e) Fastai

Fig. 12 Confusion matrix for medical diseases

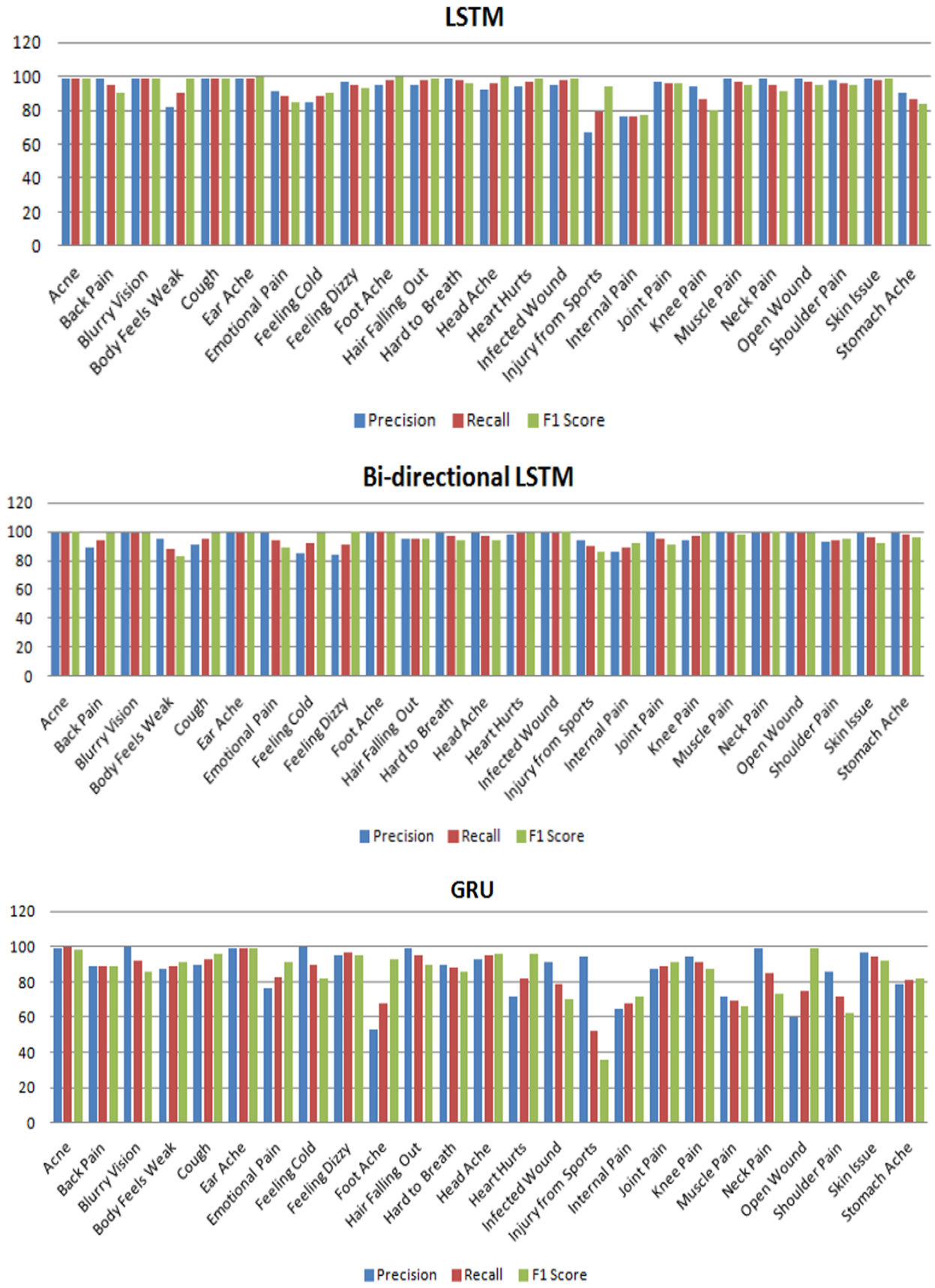


Fig. 13 Performance of deep learning models

5.3 Confusion matrix

Each row represents an instance of actual class, i.e., the number of samples, and each column represents an instance of predicted class, i.e., the predicted number of samples, in a 25 × 25 confusion matrix. The degree of successfully predicted classes is represented by the values of diagonal elements. Because they are wrongly categorized with another class, the false classified off-diagonal elements express the perplexity. The confusion matrix of four algorithms applied to the illnesses is depicted in Fig. 12. After calculating the confusion matrix, all of the evaluative metrics (accuracy, loss, precision, recall, and F1 score) have been computed and are presented in the following subsections.

5.4 Comparative analysis

In Fig. 13, comparison has been presented to predict the performance of four models such as LSTM, Bidirectional LSTM, GRU, bidirectional GRU, and Fastai which have been incorporated on 25 different types of diseases.

LSTM predicts acne, blurry vision, and cough. Bidirectional LSTM predicts acne, blurry eyes, and heart hurts. Neck pain, GRU predicts earache, bidirectional GRU predicts cough, earache, foot ache, headache, heart hurts, infected wound, neck pain, and an open wound with the highest precision, recall, and F1 score by 99% while as Fastai predicts feeling dizzy, knee pain, neck pain with 100% precision, and skin issue with 100% recall and F1 score.

6 Discussion

By modifying existing medical processes, speech recognition is predicted to disrupt the health sector. In health, speech technology has unrivaled prospects, which can improve the current healthcare system, which is constantly beset by an aging population and chronic diseases. For its promise and wide range of uses, speech recognition technology is finding a market in healthcare. Several voice-activated healthcare solutions are in the works, with the potential to improve the lives of tens of thousands of people. As a result, notwithstanding their novelty, unique deep learning architectures/models for detecting illnesses in speech using textual phrases are presented in this study. Because no existing state-of-the-art works provide significant amounts of data to validate other types of medical speech utterances, 6661 wav (waveform audio) files and one.csv (comma-separated value) file containing 13 distinct categorization columns were used to validate the different types of medical speech utterances. This solves the challenge of evaluating and quantifying deep learning models. Fastai is effective for voice recognition of illnesses, according to the experimental results, which show a significant correlation between model prediction performance and experimental data. As previously reported, no scientific work has contributed to the donations to the best of our knowledge. Despite the tremendous potential of speech technology in the health domain, large-scale adoption of speech-enabled technologies faces several challenges. As there are less noise reduction filters used in this study, the accuracy of speech recognition is not as fine as it should be and such noisy input also degrades the system performance rate. In addition to this, cultural and linguistic diversity limit the potential application of speech technologies in digital healthcare. To perform effectively in such circumstances, voice-based healthcare solutions must be trained in multiple languages. A vast amount of data is required to generalize DL models. The datasets available for voice processing and analysis are often very restricted. Despite the fact that there are more than 5,000 languages spoken worldwide, just 389 languages are being used by 94% of the world's population (Latif et al. 2021), implying that language as well as speech analytical research suffers from a data shortage.

7 Conclusion and future scope

The ability of computer software to recognize and transform spoken words into a machine-readable format is known as SR (speech recognition), or voice recognition. SR technology enables doctors to speak and amend notes directly into the medical records without any conventional transcription. Thanks to deep learning architectures, speech recognition accuracy has improved dramatically, and numerous automated learning algorithms have been created. LSTM (long short-term memory), Bidirectional LSTM, GRU (gated recurrent unit), bidirectional GRU, and Fastai deep learning frameworks are used in this study to understand and analyze disorders using medical voices and their associated textual words. Fastai produced the highest accuracy, precision, recall, and loss, while as LSTM has the highest F1 score. This research aims to evaluate and improve various data augmentation and classification models. It has been showed that the proposed technique enhanced the rate of predicting the speech considerably. In future, the system necessities the use of optimization techniques to improve the system's performance. Along with this, the recognition rate needs to be improved by using various noise reduction filters during the preprocessing step. In addition to this, the creation of reliable and standardized measures, questionnaires, and other

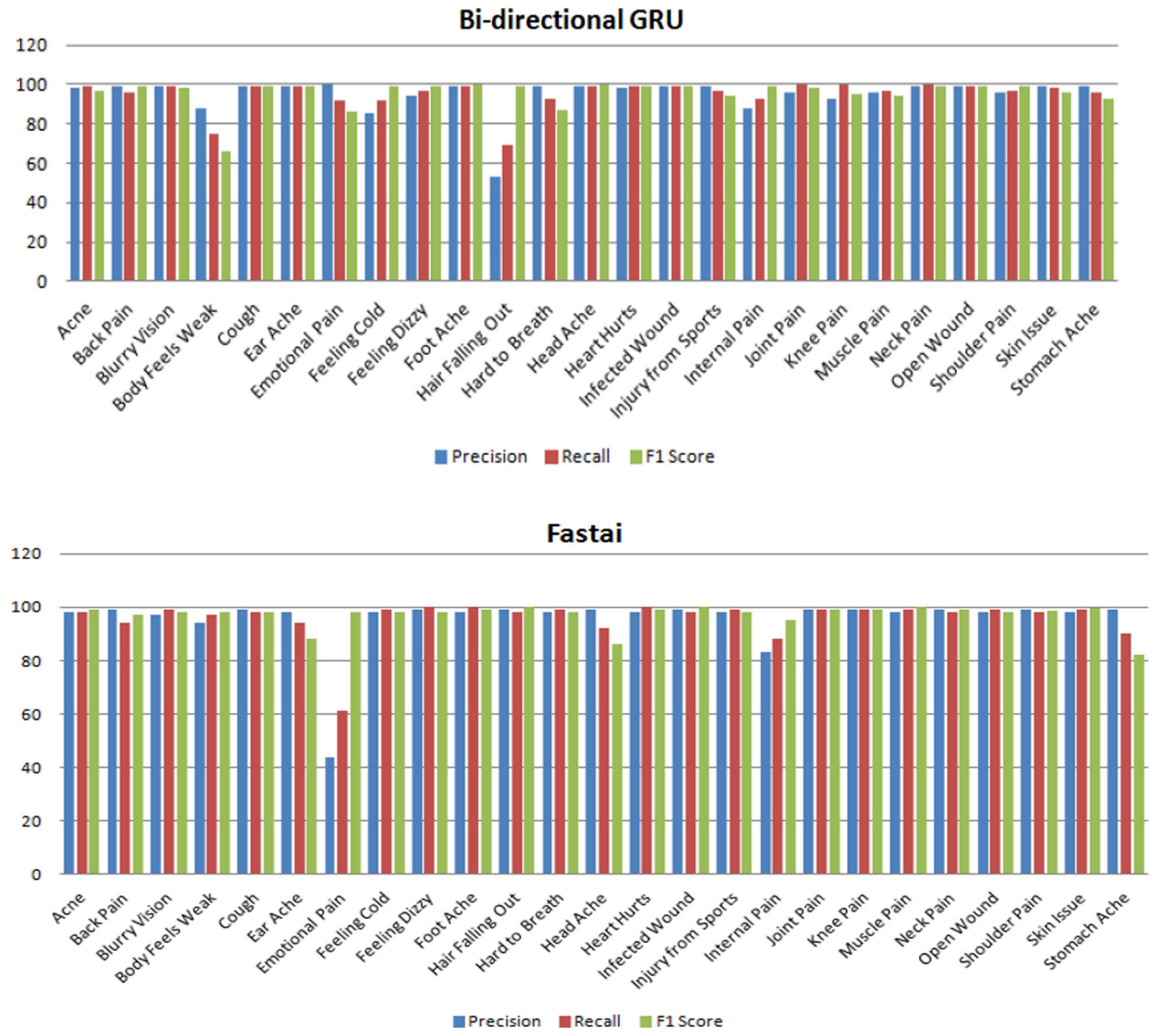


Fig. 13 continued

instruments specifically designed to assess SR usability and clinical workflow could aid in the identification of specific issues and potential remedies. Thusly, it is required to develop an automatic error detection system for the medical disease generated after speech recognition for the safety of the patient. It is also required that during and after adapting the speech recognition process, the study focus on the doctors' preferences, practices, and potential. Furthermore, in the coming years, the system recognition technique should be created in such a way that it does not encounter any language barriers.

Funding Not applicable.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdelgwad MM, Soliman THA, Taloba AI, Farghaly MF (2021) Arabic aspect based sentiment analysis using bidirectional GRU based models. J King Saud Univ–comput Inf Sci. https://doi.org/10.1016/j.jksuci.2021.08.030

Akinloye FO, Obe O, Boyinbode O (2020) Development of an affective-based e-healthcare system for autistic children. Sci African 9:e00514. https://doi.org/10.1016/j.sciaf.2020.e00514

Al-Hassan A, Al-Dossari H (2021) Detection of hate speech in Arabic tweets using deep learning. Multimedia Syst. https://doi.org/10.1007/s00530-020-00742-w

Alhussein M, Muhammad G (2018) Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access 6:41034–41041. https://doi.org/10.1109/ACCESS.2018.2856238

Blackley SV, Huynh J, Wang L, Korach Z, Zhou L (2019) Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc 26(4):324–338. https://doi.org/10.1093/jamia/ocy179

Dey R, Sale F (2017) Gate variants of Gated Recurrent Unit (GRU) neural networks. In: 60th International Midwest Symposium on Circuits and Systems, pp 1597–1600

Graves, A., Jaitly, N., Mohamed, A. (2013) Hybrid Speech Recognition with Deep Bidirectional LSTM. In: IEEE workshop on Automatic Speech Recognition and Understanding, pp 273–278

Ismail A, Abdlerazek S, El-Henawy IM (2020) Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping. Sustain (switz). https://doi.org/10.3390/su12062403

Jayashankar S, Sridaran R (2017) Superlative model using wordcloud for short answers evaluation in eLearning. Educ Inf Technol 22:2383–2402. https://doi.org/10.1007/s10639-016-9547-0

Johnson M, Lapkin S, Long V, Sanchez P, Suominen H, Basilakis J, Dawson L (2014) A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak. https://doi.org/10.1186/1472-6947-14-94

Krishnan PT, Joseph Raj AN, Rajangam V (2021) Emotion classification from speech signal based on empirical mode decomposition and non-linear features. Complex Intell Syst 7:1919–1934. https://doi.org/10.1007/s40747-021-00295-z

Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU (2018) Electronic health record interactions through voice: a review. Appl Clin Inform 9(3):541–552. https://doi.org/10.1055/s-0038-1666844

Kumar Y, Singh N, Kumar M, Singh A (2021) AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi language. Soft Comput 25(2):1617–1630. https://doi.org/10.1007/s00500-020-05248-1

Lam HY, Tang YM, Tang V, Wu CH (2020) An intelligent m-healthcare system for improving the service quality in domestic care industry. IFAC-PapersOnLine 53(2):17439–17444. https://doi.org/10.1016/j.ifacol.2020.12.2113

Latif S, Qadir J, Qayyum A, Usama M, Younis S (2021) Speech technology for healthcare opportunities challenges, and state of the art. IEEE Rev Biomed Eng 14:342–356. https://doi.org/10.1109/RBME.2020.3006860

Lazzarini V (2019) Soundfiles. In: Computer music instruments II. Springer, Cham. https://doi.org/10.1007/978-3-030-13712-0_10

Louinci K, Meziani K, Riu B (2021) Muddling label regularization deep learning for tabular datasets. arXiv, pp 1–36

Lu L, Sheng J, Liu Z, Gao JH (2021) Neural representations of imagined speech revealed by frequency-tagged magnetoencephalography responses. Neuroimage 229:117724. https://doi.org/10.1016/j.neuroimage.2021.117724

Luchies E, Spruit M, Askari M (2018) Speech technology in Dutch health care: A qualitative study. In: HEALTHINF 2018–11th international conference on health informatics, proceedings; part of 11th international joint conference on biomedical engineering systems and technologies, BIOSTEC, vol 5, pp 339–348. https://doi.org/10.5220/0006550103390348

Mehta RP, Sanghvi MA, Shah DK, Singh A (2020) Sentiment analysis of tweets using supervised learning algorithms. In: Luhach A, Kosa J, Poonia R, Gao XZ, Singh D (eds) First international conference on sustainable technologies for computational intelligence advances in intelligent systems and computing. Springer, Singapore. https://doi.org/10.1007/978-981-15-0029-9_26

Mohamed J, Zweig G, Gong Y (2015) LSTM time and frequency recurrence for automatic speech recognition. IEEE Workshop Autom Speech Recognit Underst (ASRU). https://doi.org/10.1109/ASRU.2015.7404793

Mohammed MA, Abdulkareem KH, Mostafa SA, Ghani MKA, Maashi MS, Garcia-Zapirain B, Oleagordia I, Alhakami H, Al-Dhief FT (2020) Voice pathology detection and classification using convolutional neural network model. Appl Sci (switz) 10(11):1–13. https://doi.org/10.3390/app10113723

Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks a systematic review. IEEE Access 7:19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Noort MC, Reader TW, Gillespie A (2021) The sounds of safety silence: interventions and temporal patterns unmute unique safety voice content in speech. Saf Sci 140:105289. https://doi.org/10.1016/j.ssci.2021.105289

Patil S, Agashe S (2021) Comparison of neural network architectures for speech emotion recognition. In: Biswas A, Wennekes E, Hong TP, Wieczorkowska A (eds) Advances in speech and music technology. advances in intelligent systems and computing. Springer, Singapore. https://doi.org/10.1007/978-981-33-6881-1_25

Paulett JM, Langlotz CP (2009) Improving language models for radiology speech recognition. J Biomed Inform 42(1):53–58. https://doi.org/10.1016/j.jbi.2008.08.001

Poder TG, Fisette JF, Déry V (2018) Speech recognition for medical dictation: overview in quebec and systematic review. J Med Syst. https://doi.org/10.1007/s10916-018-0947-0

Ramasubramanian K, Singh A (2019) Deep learning using keras and tensorflow. In: Machine learning using R. Apress, Berkeley. https://doi.org/10.1007/978-1-4842-4215-5_11

Santosh KC (2019) Speech processing in healthcare can we integrate. In: Intelligent speech signal processing. Elsevier. https://doi.org/10.1016/B978-0-12-818130-0.00001-5

Shukla S, Jain M (2021) A novel stochastic deep resilient network for effective speech recognition. Int J Speech Technol 24:797–806. https://doi.org/10.1007/s10772-021-09851-x

Sonal J, Dodiya T (2016) Speech recognition system for medical domain pdf. Int J Comput Sci Inf Technol 7(1):185–189

Suominen H, Zhou L, Goeuriot L, Kelly L (2016) Task 1 of the CLEF ehealth evaluation lab 2016 handover information extraction. CEUR Workshop Proceed 1609:1–14

Takao T, Masumura R, Sakauchi S, Ohara Y, Bilgic E, Umegaki E, Kutsumi H, Azuma T, Medicine A, Takao T (2018) New report preparation system for endoscopic procedures using speech recognition technology, pp 6–8. 10–1055-a-0579–6494.

Uddin MZ, Nilsson EG (2020) Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng Appl Artif Intell 94:103775. https://doi.org/10.1016/j.engappai.2020.103775

van Lente H, Boon WPC, Klerkx L (2020) Positioning of systemic intermediaries in sustainability transitions between storylines and speech acts. Environ Innov Soc Trans 36:485–497. https://doi.org/10.1016/j.eist.2020.02.006

Vij A, Pruthi J (2018) An automated psychometric analyzer based on sentiment analysis and emotion recognition for healthcare. Proced Comput Sci 132:1184–1191. https://doi.org/10.1016/j.procs.2018.05.033

Zhang F, Underwood G, McGuire K, Liang C, Moore DR, Fu QJ (2019) Frequency change detection and speech perception in cochlear implant users. Hear Res 379:12–20. https://doi.org/10.1016/j.heares.2019.04.007

Zisad SN, Hossain MS, Andersson K (2020) Speech emotion recognition in neurological disorders using convolutional neural network. In: Mahmud M, Vassanelli S, Kaiser MS, Zhong N (eds) Brain informatics bi 2020 lecture notes in computer science. Springer, Cham. https://doi.org/10.1007/978-3-030-59277-6_26