Received: 23 February 2023 | Revised: 5 April 2023 | Accepted: 18 April 2023

DOI: 10.1111/exsy.13327

ORIGINAL ARTICLE

Expert Systems WILEY

Voice pathology identification system using a deep learning approach based on unique feature selection sets

Nuha Qais Abdulmajeed | Belal Al-Khateeb | Mazin Abed Mohammed

Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq

Correspondence
Mazin Abed Mohammed, Computer Science Department, College of Computer Science and Information Technology, University of Anbar, 31001, Ramadi, Anbar, Iraq.
Email: mazinalshujeary@uoanbar.edu.iq

Abstract

Voice pathology diagnosis requires extracting significant features from voice signals, and classical machine learning models can overfit to the training data, which can cause difficult issues and pose challenges. The study aimed to develop a reliable and efficient system for identifying voice pathologies utilizing the long short-term memory (LSTM) method. The study combined unique feature sets such as the mel frequency cepstral coefficients (MFCCs), zero crossing rate (ZCR), and mel spectrograms, which have not been used together in previous works. Voice pathology identification improved the accuracy rate using the LSTM approach on the Saarbruecken voice database (SVD) samples. The best results achieved by the proposed system showed an accuracy rate of 99.3% for /u/ vowel samples in neutral pitch, 99.2% for /a/ vowel samples in high pitch, 99% for /i/ vowel samples in neutral pitch, and 99.2% for sentence samples. The experimental results were evaluated utilizing accuracy, precision, specificity, sensitivity, and F1 measures. Additionally, the study compared the performance of LSTM with that of artificial neural networks (ANNs) and found that LSTM achieved better outcomes.

KEYWORDS

artificial neural networks, LSTM, mel frequency cepstral coefficients, mel spectrogram, SVD, unique feature selection sets, voice pathology, zero crossing rate

1 | INTRODUCTION

Voice pathologies are a prevalent issue that considerably challenges effective social interaction. These pathologies stem from various factors, including vocal cord paralysis, infections of the voice tissue, abnormal voice use, swelling, and drug abuse (Al-Nasheri et al., 2017). Certain professionals, such as singers, actors, lawyers, and teachers who frequently use their voices at an elevated level, are more prone to developing voice disorders. Approximately one-quarter of these professionals are globally affected by such pathologies. The impacts of voice disorders can lead to social and personal complications, including anxiety, depression, and difficulties communicating with others (Islam et al., 2020). The standard method for diagnosing these pathologies involves invasive procedures such as laryngeal endoscopy, stroboscopy, and laryngeal electromyography, which require highly trained medical professionals and expensive specialized equipment. Unfortunately, these procedures can be painful and even traumatic. Moreover, inadequate medical services in remote areas can delay proper diagnosis and treatment. Therefore, there is an ongoing effort to explore alternatives to the traditional diagnoses of voice disorders as researchers seek more efficient and less invasive diagnostic techniques (Gómez-García et al., 2019; Hegde et al., 2019).

Voice pathology can be identified by utilizing traditional deep learning and machine learning techniques, as well as a hybridization of both (Al-Dhief, Latiff, Malik, Salim, et al., 2020; Islam et al., 2020). Machine learning techniques involve extracting features from voice signals and analysing them to classify the signals as healthy or pathological. These evaluation methods are objective and do not rely on human decision-making. They

https://doi.org/10.1111/exsy.13327

are also convenient to implement, as voice recordings can be obtained remotely through various internet recording applications (Mohammed et al., 2020). The main commonly utilized voice features for voice pathology detection include convolutional neural networks (CNNs), spectral techniques, linear prediction cepstral coefficients (LPCCs), mel frequency cepstral coefficients (MFCCs), glottal signal parameters, shimmer parameters, and jitter features. For classification purposes, various conventional algorithms are used to identify voice pathologies, such as K nearest neighbours, random forest technique, convolutional neural networks (CNNs), Gaussian mixture models (GMMs), support vector machines (SVMs), and online sequential extreme learning machine. The most widely utilized voice pathology databases are the Saarbruecken Voice Database (SVD), Massachusetts Eye and Ear Infirmary Database (MEEI), and Arabic Voice Pathology Database (AVPD) (Abdulmajeed et al., 2022; AL-Dhief, Latiff, Malik, Sabri, et al., 2020).

Existing methods for voice pathology detection face a number of limitations and difficulties, including the challenge of manually determining suitable voice features, the use of a small number of samples of voices for the majority of three pathologies, the exclusion of all but one vowel from the databases, and the difficulty of choosing a suitable classification method. To address these issues and improve the effectiveness of speech pathology identification systems in prior studies, we aimed to find a solution. Using the LSTM classifier with the SVD data set in a voice pathology identification system offers a powerful approach to automate the diagnosis of voice disorders. This system can aid speech-language pathologists and health care professionals in quickly and accurately identifying voice disorders in patients, leading to earlier and more effective treatment. The main contributions of this study are as follows:

1. A voice pathology identification system is proposed that uses the LSTM approach with the voice signals of the SVD data set to overcome the limitations of previous speech pathology identification techniques.
2. A combination of unique feature sets are extracted, including the zero crossing rate (ZCR), MFCCs, and mel-spectrogram, which have not been used together in previous studies.
3. The accuracy rate for voice pathology identification is improved using the LSTM approach with a unique feature set in which all available SVD samples are utilized, including all pathology subsets, even those that are uncommon.
4. A comparison of the performance of LSTM and artificial neural network (ANN) is conducted in the extracted features and selection stages, as well as in the classification stage of both pathological and healthy cases. The findings demonstrate that the LSTM method achieved better results.

The remaining sections of this research are organized as follows: Section 2 provides a review of the literature on the classification of vocal pathologies. Section 3 explains the proposed system for identifying voice pathology. Section 4 presents the experimental results and corresponding discussions. Finally, Section 5 provides the conclusion.

2 | RELATED WORK

The main challenges and issues with existing techniques for detecting and identifying voice pathology include a lack of standardization, high false positive rates, lack of generalization, high computational cost, and classification subjectivity. These challenges highlight the need for continued research and development in the field to develop more accurate, reliable, and generalizable techniques that can be widely adopted in clinical and research settings (Muhammad & Alhussein, 2021). Many studies have used machine and deep learning methods, such as Al-Nasheri et al. (Al-Nasheri et al., 2017), who discussed enhancing feature extraction methods for detecting and identifying voice pathology utilizing correlation functions in different frequency regions. The study extracted maximum peak values and their corresponding lag values as features and used an SVM technique to investigate the contribution of different frequency regions for detection and identification methods. The study used cases of sustained vowel /a/ from pathological and normal voices in three dissimilar data sets and performed a *t* test to determine important transformations between pathological and normal cases. The strength of this study lies in its use of a robust feature extraction method and examination of different frequency regions, while its limitations include a small data set size and the use of a single type of voice sample. Another study by Naikare et al. in (Naikare et al., 2018) focused on voice pathology classification utilizing noninvasive techniques with the support of ML methods. They used 75 voice samples of SVD for each of the three disorder classes (vocal fold paralysis, dysphonia, and laryngitis) in addition to normal speech cases. All these samples were taken from the vowel /a/ samples. This study focused on extracting spectral cepstral features (MFCC, LSF, and LPCC), short-term features (spectral centroid, short-term energy, and ZCR), and perturbation measures (shimmer parameters and jitter features). The extracted features were saved in a supervector utilizing GMM-UBM. This supervector was then projected on a low-dimensional feature vector known as ‘i-Vectors’. The parameters acquired afterward were utilized for training the model utilizing KNN, SVM, and naïve Bayes. The strength of this study was the use of I-vectors, which proved to be a strong benchmark method compared to speech features. The SVM algorithm provided good general outcomes compared to naïve Bayes and KNN due to its decision-making ability. However, the study had limitations, such as using only three classes of pathologies with a single vowel sample from each pathology and limited vowels in the SVD data set. The work of Dankovičová et al. (2018) utilized ML techniques to recognize pathological speech, mainly dysphonia. They used 194 samples from 94 patients with dysphonia and 100 healthy people from the SVD data set. A total of 130 features were extracted from the voice signals. They used principal component analysis as a feature selector. SVM, KNN, and RFC algorithms were used for classifying voice samples. Using feature selection was a good step in improving the study results while choosing only dysphonic pathology samples limited the study results. Wu et al. (Wu et al., 2018) used a spectrogram of voice signals as the CNN input for automated extracted features and identifying normal and disordered voices. The authors used six pathology subsets as the pathological set (recurrent laryngeal nerve paralysis, leukoplakia, laryngitis, Rinke's edema, vocal fold polyps, and vocal fold carcinoma). The strengths of this study were using spectrograms as input for classifying normal and pathological voice without manual feature extraction. However, the study had limitations, such as using only sustained vowel /a/ samples from the SVD data set with six pathologies. Kadiri and Alku (2019) conducted a methodical examination of glottal source features and their capability to detect voice pathology. The researchers utilized the QCP glottal inverse filtering method to extract glottal flows, as well as the ZFF method to compute approximate glottal source signals and direct acoustic voice signals to extract glottal source features. To conduct the pathology detection experiments, they used SVM and two databases: the HUPA and the SVD database. Combining glottal source features with conventional MFCCs and PLP features yielded the best detection performance. The strengths of this study were the robust features used and the techniques chosen for extracting features, while using a single classifier may be a limitation to the study performance. Al-Dhief, Latiff, et al. (2021) introduced a novel voice pathology detection system that utilizes various voice signals to distinguish between healthy and pathological classes. Specifically, voice signals for the vowel /a/ from both healthy and pathological classes were obtained from the SVD database. Subsequently, the features of these voice signals were extracted using the MFCC method. Classifying voice signals into healthy or pathological was then carried out using SVM. The strengths of this study were achieving 84% accuracy with the MFCC features, while its limitations were that the results obtained from SVM showed a degradation in performance as the number of voice cases increased. The work of Omeroglu et al. (2022) suggested that a new multimodal architecture was developed to improve the automatic detection of voice pathology using both speech and Electrogastrogram (EGG) signals from the SVD database. To obtain deep features, the researchers proposed a multimodal framework consisting of two parallel CNNs, one for voice signals and one for EGG signals. They also extracted classical handcrafted features in the same way and concatenated these features to form a more prominent feature set. A feature selection method was used to eliminate redundant features. Finally, an SVM classifier was employed to detect voice pathology. One of the strengths of this study is the careful selection and fusion of both handcrafted and deep features, which enabled a more comprehensive and accurate representation of the voice signals, while its limitation may be in using EGG signals with the voice. The related studies of voice pathology using AI with SVD are presented in Table 1.

It can be seen in the table above that a wide variety of machine and deep learning techniques have been applied to the problem of voice pathology identification and classification using a wide variety of extracted features. However, class imbalance is a common problem in vocal pathology databases because some pathologies are more common than others. This can lead to skewed models that do well across the board except for a select few courses. Changes in tone, volume, and sound of the voice are only some of the methods that vocal pathologies can make themselves known. To detect these alterations and construct useful representations of the data for machine learning methods, efficient feature extraction is needed. In addition, the high dimensionality and complexity of speech signals make it difficult to isolate the most important features for determining a vocal pathology diagnosis. Overfitting occurs when the model becomes overly specific to the training data and cannot accurately predict outcomes from novel, unknown data. To the best of our knowledge, all prior studies in the literature have concentrated on feature selection in an effort to reduce the feature dimension; however, the same features have been used in each of these studies. In contrast, our suggested method takes the most widely used features in voice pathology classification and uses them to choose a novel collection of features that have not been implemented in existing classifiers. Our method is unique among studies because of this novel approach, which has the potential to greatly increase vocal pathology classification precision.

3 | THE PROPOSED VOICE PATHOLOGY IDENTIFICATION SYSTEM

To accurately diagnose vocal disorders from voice samples, we have developed a voice pathology detection system that combines an LSTM classifier with the SVD data set's voice signals. The system is trained using speech patterns captured from people suffering from vocal diseases such as hoarseness, dysphonia, and others found in the SVD data set. This data set is ideal for use in a deep learning-based system because it includes speech signals that have been decomposed into SVDs to produce a compact depiction of the speech signals. The SVD representations of the speech signals, along with the labels designating the type of vocal disorder found in each signal, are passed into the LSTM classifier during training. The classifier then 'learns' to recognize patterns in the speech data that indicate the presence of particular vocal disorders. Voice disorders in untrained speaking signals can be identified using the vocal pathology detection method. When a new speech input is received, it is SVD-decomposed and then fed into a previously learned LSTM classifier. During the training phase, the LSTM model is trained using the training set, while the testing set is used to evaluate the proposed system's performance. The data set contains a total of 26,468 samples, with 8878 healthy and 17,590 pathological samples. The data set includes three vowels (/i/, /a/, /u/) and sentences, with each vowel having four intonations. Each intonation has 2042 voice signals, while 1988 signals are sentences. The signals are split into 70% for training and 30% for testing. Figure 1 depicts the major stages of the LSTM classifier-based speech pathology detection system that we suggest. There are roughly four distinct stages

TABLE 1 Summary of the related works in voice pathology classification based on AI techniques.

Author and year	Method	Features	Data set	Accuracy	Strengths	Limitations
Al-Nasheri et al. (2017)	SVM	Entropy, peak, and lag	SVD	92.27%	Used a robust feature extraction method and examined different frequency regions.	Used a small data size and single type of voice samples.
Hossain et al. (2017)	GMM	Glottal signal, shape, and cepstral features	SVD	93%	Fused the two input modalities (voice and EGG).	EGG signals alone were not very effective in detecting speech pathology.
Naikare et al. (2018)	SVM, Naïve Bayes and KNN	MFCC, LSF, LPCC, short time energy, spectral centroid, ZCR, jitter and shimmer parameters	SVD	94%	I-vectors proved to be a robust speech feature, while SVM provides better results than the KNN and Naïve Bayes.	The analysis only utilized 3 classes of pathologies with one vowel sample from the entire set of pathologies and vowels in the SVD data set.
Dankovičová et al. (2018)	SVM, KNN and RFC	Spectral Roll-off, Shimmer, MFCC, and Jitter	SVD	85.71%	Extracted a large number of features and used feature selection to choose the best features.	Chose only dysphonia pathology samples which meant a small data set size.
Dahmani and Guerti (2018)	KNN	Glottal signals	SVD	93%	Reached a high accuracy in voice pathology classification with the KNN classifier and 12 features.	Used a small data size that contained /a/ vowel, Parkinson's disease and spasmodic dysphonia samples.
Wu et al. (2018)	CNN	CNN-based features	SVD	77%	Classified pathological and normal voice using spectrograms needed for manual feature extraction.	Used only the sustained vowel /a/ samples from the SVD data set with the six pathologies.
Kadiri and Alku (2019)	SVM	Glottal source features	SVD	74.32%	The combination used glottal source features with the conventional MFCCs and PLP.	Used a single classifier in their experiments with a large data size.
Mohammed et al. (2020)	CNN	CNN-based	SVD	95.41%	The CNN offered fast and accurate diagnoses and treatments in just 3 s	Small data size of only the three common pathologies based on /a/ vowel samples.
Al-Dhief, Latiff, et al. (2021)	SVM	MFCC	SVD	84.37%	Achieved 84% accuracy with the MFCC features.	SVM results degraded when the number of voice samples increased.
Al-Dhief, Baki, et al. (2021)	OSELM	MFCC	SVD	91.17%	A good capability system to detect and classify outcomes using OSELM classifier.	The voice samples used for analysing three common pathologies (paralysis, polyp, and cyst) based on the vowel sound /a/.
Syed et al. (2021)	CNN, RNN	spectral centroid, MFCC, pitch, spectral flux, roll-off, energy entropy, ZCR, and energy.	SVD	87.11	A good performance was reached with only pathology sample subsets using the CNN classifier.	Classification errors made by CNNs were difficult to investigate due to their 'black box' nature and size of the training data sets.
Omeroglu et al. (2022)	SVM	MFCC, LPC, Pitch, spectral slope, CNN-based	SVD	90.10%	Used techniques for the selected and fused handcrafted features.	Used EGG signals with voice signals.

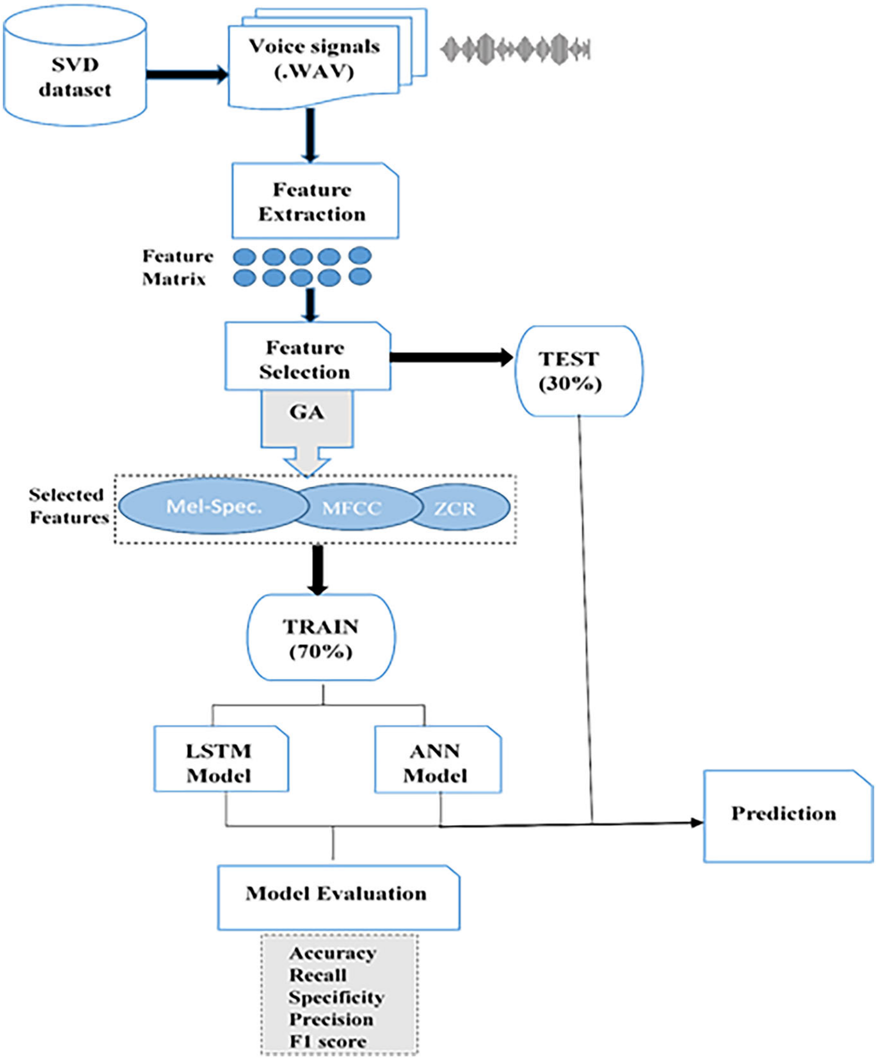

FIGURE 1 The voice pathology identification system.

to the research. Collecting speech recordings from the SVD library is the first step. Voice pathology identification relies heavily on feature extraction from speech signals, which is the emphasis of the second part. The MFCC technique was used to derive features from the vocalizations. In the final stage, features were selected using a feature selection method. In the final stage, characteristics were selected using a feature selection method. This process is crucial to improve the classification model's efficiency, as it allows us to eliminate unnecessary features that are unnecessary or unnecessary to the task. Finally, in the fourth phase, the study used LSTM to detect and classify the voice signals into healthy or pathological classes. The features were used to train the classification model, and the model's recall, precision, and accuracy were measured to determine how well it performed. The voice pathology identification method was broken down in depth here.

3.1 | Data set

The SVD data set was used for this investigation because it contains recordings of voices from healthy and disordered individuals. Sustained /a/, /i/, and /u/ with various intonations (low-high-low, low, normal, and high) and a spoken sentence in German, 'Guten Morgen, wie geht es Ihnen?' (translates to 'Good morning, how are you?'), are included in this publicly available database maintained by the Institute of Phonetics at Saarland University. The SVD database stores audio captured at 50 kHz with 16-bit precision, and each file has both the speech and EGG impulses stored. Six hundred and eighty-seven fit individuals (428 females and 259 males) and thirteen hundred and fifty-six cases (727 females and 629 males) with one or more of the 71 diseases are represented by audio samples. The data files are available for distribution from the (Barry & Pützer, 2007) website; the speech and EGG signals can be saved in their original file forms (NSP and EGG, respectively) or as WAV files, which can then be compressed on the computer. The export process may take several minutes to complete. The unique features of the SVD database, including the

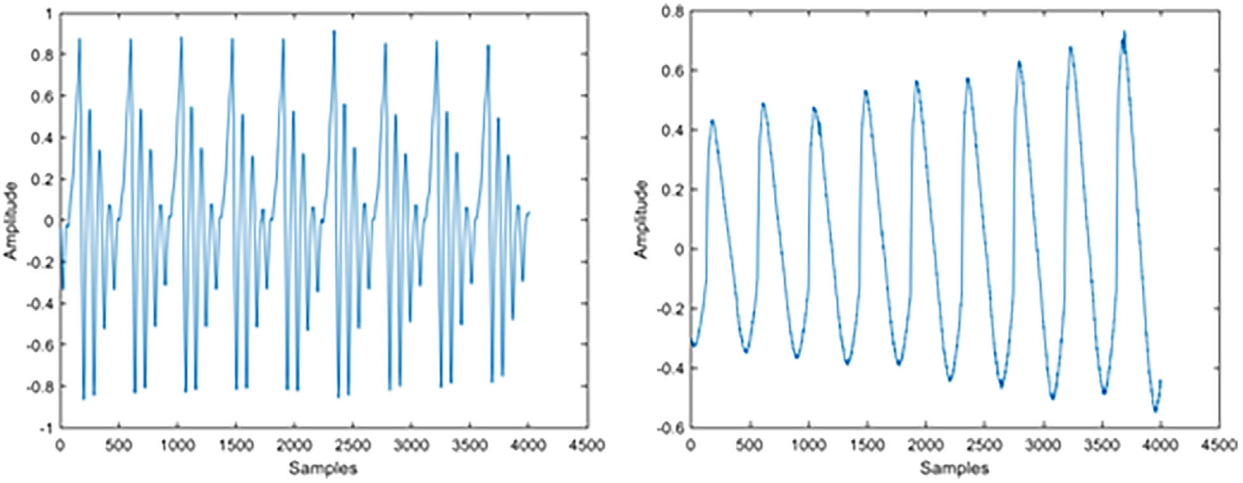

FIGURE 2 Voice and Electrogastrogram (EGG) signal samples taken in the Saarbruecken voice database (SVD) data set.

TABLE 2 The attributes of the Saarbruecken voice database (SVD) data set.

Language	Sampling frequency	Vowels
German	50 KHz	1. /a/ 2. /u/ 3. /i/ 4. Sentences

variety of recorded voice samples, make it an excellent resource for research on voice pathology detection (Barry & Pützer, 2007). The samples and attributes of the SVD data set are presented in Figure 2 and Table 2.

3.2 | Feature extraction

Features were extracted from individual cases and a feature matrix was built to represent the key features of human speaking. To be more specific, we derived a total of 192 features for a particular vowel, 48 for each of its possible intonations. There were calculated to be 624 features total, including those for each subject's /a/, /i/, and /u/ vocal samples and their associated phrases. Multiple varieties can be found in some of the features. (e.g., MFCC1 and MFCC2) Below, describe these specifics:

3.2.1 | Spectral centroid

The spectral centroid denotes the frequency at which the energy of a spectrum is centred, or in other words, where the 'centre of mass' of a sound is placed. Equation (1) shows the calculation of the spectral centroid.

$$f_C = \frac{\Sigma_k s(k) f(k)}{\Sigma_k s(k)}, \tag{1}$$

where *S* (*k*) is the spectral magnitude at frequency bin *k*, and f (*k*) is the frequency at bin *k*.

3.2.2 | Spectral roll-off

Spectral roll-off refers to a measure of the shape of a signal, indicating the rate at which high frequencies decrease in amplitude to 0. Mathematically, it is represented as:

$$W(\omega)=\vartheta\left(\frac{1}{\omega^{n+1}}\right),(\text{as}\,\omega\rightarrow\infty), \tag{2}$$

where $W(\omega)$ is said to be of order $\frac{1}{\omega^{n+1}}$ if there exist $\omega 0$ and some positive constant M< ∞ such that $|W(\omega)|< \frac{M}{\omega^{n+1}}$ for all $\omega > \omega 0$.

3.2.3 | Spectral bandwidth

Spectral bandwidth is a measurement that describes the width of the band of light at half the peak maximum, also known as the full width at half maximum (FWHM). This measurement is often used to analyse signals, such as voice and EGG signals, and represents the range of frequencies present in the signal. It is represented by Equation (3):

$$v=\sqrt{\left(1-\frac{(m_2)^2}{m_{0^*}m_4}\right)}, \tag{3}$$

where spectral breadth (v) and spectral moment (m) describe the wave's features.

3.2.4 | Zero crossing rate

Measuring the total amount of zero-crossings in a given time window is a simple but effective way to measure a signal's regularity. By tracking where the signal goes through the zero line, we can learn more about the signal's general irregularity and consistency. Mathematically, it is represented in Equation (4):

$$zcr=\frac{1}{T-1}\sum\nolimits_{t=1}^{T-1}II\{s_t s_{t-1}<0\}. \tag{4}$$

Therefore s_t is the length of signal t and II {x} is the pointer function {=1 if x is true, else =0}.

3.2.5 | Mel frequency cepstral coefficients

The spectral spread of a signal can be represented concisely by the MFCCs, which are a collection of features comprised of around 10 to 20 components. Speech and analysis of voice benefit greatly from the coefficients, which are derived through a sequence of mathematical procedures meant to mimic the way the human ear processes sound. The MFCCs can be used to effectively describe and compare various speech samples by summarizing the spectral properties of a signal within a small number of features. Its mathematical representation is given by the following equation:

$$c_i=\sum\nolimits_{n=1}^{Nf}Sn\cos\left[i(n-0.5)\left(\frac{\pi}{Nf}\right)\right],i=1,2,\ldots,L. \tag{5}$$

The number of triangular filters in the filter bank (Nf), the log energy produced of the nth filter coefficient (Sn), and the overall number of MFCC coefficients (L), are all inputs to the algorithm ci = MFCC coefficient. The signal is first filtered using a bank of triangular filters; the resulting energy is then logged; the log filter bank energies are subjected to a discrete cosine transform (DCT); and finally, the preliminary L coefficients of the DCT output are retained as part of the MFCC calculation. The L MFCC coefficients obtained in this way offer a condensed spectrum depiction of the original signal.

3.2.6 | Chroma feature

The chroma of an audio output is represented by a vector of 12 features, one for every one of the 12 pitch classes. It's frequently utilized in music data retrieval tasks like genre classification, melody identification, and chord estimation because it offers a way to describe the resemblance between various parts of music. Standard methods for determining chroma features involve initial transforming the audio data into a spectrogram and then adding the power of each frequency range associated with a given pitch class. The resulting vector reflects the signal's chrominance features. The equivalent equation in mathematics is Equation (6).

$$S(t,c) = \sum_k S\left(t, 2^{c+k}\right), \tag{6}$$

where k is a number within a suitable range.

3.2.7 | Mean

The mean measures how typical the data in a collection is. By summing all the data elements and splitting by the entire amount of data elements in the collection, we can calculate a numerical average. A formula for mean is given by Equation (7).

$$\text{Mean} = \frac{\Sigma x}{n}. \tag{7}$$

where x = values of data and n = number of data values.

3.2.8 | Root-mean-square

The root-mean-square (RMS) refers to the total magnitude of the signal, which in layman terms can be interpreted as the loudness or energy parameter of the audio file. It is defined mathematically by Equation (8):

$$E(x) = \sum_n |x(n)|^2. \tag{8}$$

3.2.9 | Tempogram

A tempogram is a representation of the time-tempo relationship in a signal, measured in beats per minute (BPM), which denotes the pace or speed of the underlying rhythm. It is essentially a matrix of features that describes the prevalence of different tempos at different points in time. It can be expressed as a function T, mapping the time parameter t (in seconds) and tempo parameter τ (in BPM) to a nonnegative real number as in Equation (9):

$$C(t,[T]) = \sum_{\lambda \in [T]} T(t,\lambda) \tag{9}$$

3.2.10 | Mel spectrogram

Mel spectrograms can be obtained by multiplying frequency domain values by a filter bank.

$$M = F \times \text{Fb}, \tag{10}$$

where F is the frequency domain of the signal and Fb is a filter bank.

The details of 48 features that were extracted with their names and numbers are presented in Table 3.

3.3 | Feature selection

Feature selection has been shown to enhance forecast performance in some areas by weeding out superfluous or useless features (Cai et al., 2018). With the help of feature selection, we were able to zero in on the optimal collection of features for separating normal from abnormal

TABLE 3 Features extracted from voice signals.

Feature name	Number of features
ZCR	1
MFCC	20
Mel spectrogram	20
Spectral centroid	1
Spectral roll-off	1
Spectral bandwidth	1
Chroma	1
Tempogram	1
Mean	1
RMS	1

TABLE 4 Genetic algorithm pseudocode.

```
Input: All features
Output: Best features
Begin
Step 1: Randomly create population P of pop size individuals.
Step 2: While (stop condition) FALSE:
    $P^* = \emptyset$
    While (size of ($P^*$) $\neq$ popsize)
        Select $p_1$ and $p_2$ from P.
        Crossover $p_1$ and $p_2$ to obtain $h_1$ and $h_{2.}$
        Mutate $h_1$ and $h_2$ to obtain $c_1$ and $c_{2.}$
        If [Dis ($p_1$, $c_1$) + Dis ($p_2$, $c_2$)] $\leq$ [Dis ($p_1$, $c_2$) + Dis ($p_2$, $c_1$)]
            If $c_1 < p_1$ then $P^* = P^* \cup \{c_1\}$ else $P^* = P^* \cup \{p_1\}$
            If $c_2 < p_2$ then $P^* = P^* \cup \{c_2\}$ else $P^* = P^* \cup \{p_2\}$
        Else
            If $c_1 < p_2$ then $P^* = P^* \cup \{c_1\}$ else $P^* = P^* \cup \{p_2\}$
            If $c_2 < p_1$ then $P^* = P^* \cup \{c_2\}$ else $P^* = P^* \cup \{p_1\}$
    Step 3: End While
    Step 4: $P = P^*$
    Step 5: Evaluate individuals in P using diagnostic scheme
Step 6: End While
END
```

speech recordings. We used the genetic algorithm (GA) to select the optimal set of k features. Models were built using a small set of features, and their cross-validation scores were analyzed; these feature groups were created with the help of evolutionary algorithms. The goal of this evolutionary method was to maximize classification success with a minimum number of features. Table 4 shows the pseudocode for the GA. The following steps summarize the mechanism of the algorithm, as shown in Figure 3:

To perform feature selection using GA:

1. The first step involved creating a population of subsets of possible features.
2. These subsets were then evaluated using a predictive model for the desired task.
3. After evaluating each member of the population, a tournament was conducted to select the subsets that progressed into the next generation.
4. The succeeding generation was composed of the tournament winners, with some crossover between the feature sets of the winners and mutations (random addition or removal of features).

The metrics obtained in each iteration (generation);

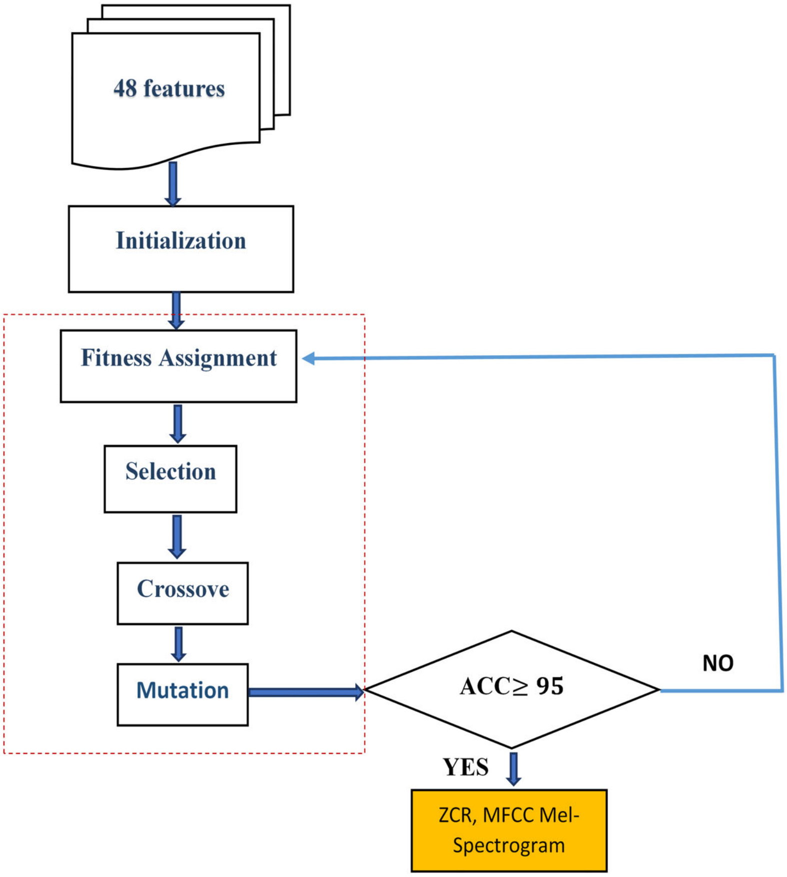

FIGURE 3 Genetic algorithm diagram.

1. **gen:** The generation index.
2. **nevals:** Number of features installed by this generation's.
3. **fitness:** The cross-validation average score measure (validation set).
4. **fitness_std:** The standard deviation of the cross-validation precision.
5. **fitness_max:** The best model from this iteration, as measured by the fitness metric
6. **fitness_min:** The generation's lowest individual model result.

The algorithm runs for a set number of generations (iterations). Then, the optimal number of the population is the selected features. The selected features are ZCR in Equation (4), MFCC in Equation (5) and the mel spectrogram in Equation (10). Names and numbers of the selected features are shown in Table 5.

3.4 | Classification stage

Voice pathology classification is the application of AI systems to distinguish between normal and abnormal examples of human speech. Specifically, this is accomplished by putting speech samples or features extracted from voice recordings into an AI model, which then predicts whether the voice is healthy or pathological. The AI model can be 'trained' on a database of speech samples that have been annotated as 'healthy' or 'pathological' and then used to make forecasts about unseen samples. This classification could help medical doctors detect voice pathologies more quickly, accurately, and objectively, which should lead to better patient care. In this study, we used two AI approaches, ANN and LSTM, to increase the reliability of our findings. We used two different classifiers based on distinct principles and have demonstrated strong performance in various domains. To ensure the accuracy of our results, we employed two AI methods: LSTM and ANN. We employed two different classifiers based on distinct principles that have demonstrated strong performance in various domains. The algorithm's parameters, including layers, learning rate, epochs, test, dropout, and train, were chosen through experimentation and achieved the best performance.

TABLE 5 Features selected by GA.

Selected feature	Number of features
ZCR	1
MFCC	3
Mel spectrogram	5

TABLE 6 The best parameters utilized in the long short-term memory (LSTM) and artificial neural network (ANN) models.

Parameters	Value
Activation function	Sigmoid
Gradient descent optimizer	Adam
Learning rate	0.00001
Epoch	200
Dropout	0.5

3.4.1 | Artificial neural network

The ANN is made up of interconnected artificial neurons that are inspired by biological neurons. These neurons take in data and send it to other neurons via a singular output (Hardesty, 2017). Both the outputs of additional neurons in the neural structure and feature values extracted from external data (such as documents or images) can serve as inputs. The function is carried out by using the information released by the neural network's terminal neurons. The ANN model presented in our study contains the same number of parameters as shown in Table 6: neurons, dropout, optimizer, learning rate, and the sigmoid activation function as the LSTM model. Figure 4 shows the architecture of the ANN model.

3.4.2 | Long short-term memory

Long short-term memory (LSTM) is optimized for modelling time series and their long-range relationships. For voice detection purposes, it has proven to be more accurate than DNNs and more effective than the gold standard of acoustic modelling. The LSTM architecture uses model parameters efficiently, converges quickly, and outperforms other models. It is a three-layer deep LSTM with dropout and dense layers. LSTM is particularly effective for modelling dynamic signals with time fluctuations and complex associations on multiple timescales. The LSTM architecture is represented in Figure 5, and the pseudocode is provided in Table 7. LSTM has proven effective in sequence marking and prediction activities such as language detection and handwriting (Yang & Horie, 2015).

The mathematical equations of the LSTM classifier are explained as follows:

For $X_t \in R^N$, where N is the feature length of each time step, f_t, o_t h_t, h_{t-1}, c_t, c_{t-1}, $b \in R^H$, where H is the hidden state dimension (Brownlee, 2017):

1. Input gate: $i_t = (W_{xi}\, x_t + W_{hi}\, h_{t-1} + W_{ci}\, c_{t-1} + b_i)$.
2. Forget gate: ft = (Wxf xt + Whf ht-1 + Wcf ct_1+ bf).
3. Output gate: ot = (Wxo xt + Who ht-1 + Wco ct + bo).
4. Memory cell: ct = ft ⊙ ct-1 + it ⊙ tanh (Wxc xt + Whcht-1 + bc).
5. Cell output: y_t =ht = ot ⊙ tanh (ct).

The activation function used in this model for the output layer is the sigmoid function for binary classifications since only two classes are utilized in this thesis (healthy and pathological). The Adam optimizer is used for its efficiency as a loss function optimization with a (0.00001) learning rate. It is used to update network weights iteratively based on training data. The parameters used in the LSTM model are explained in Table 6.

3.5 | Performance measures

Evaluating models involves determining how effectively they provide proper classification and calculating the value of each prediction in every situation. Sometimes, we only care about how often a model gets any prediction right, while at other times, it is critical that the model makes a

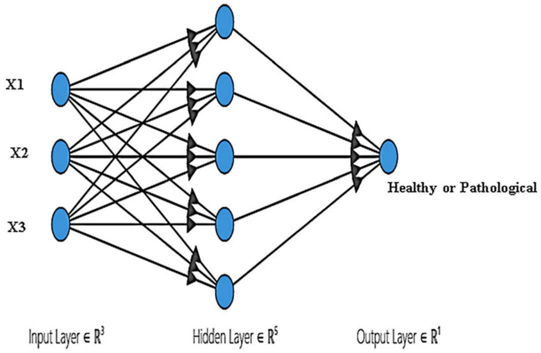

FIGURE 4 Artificial neural network (ANN) architecture.

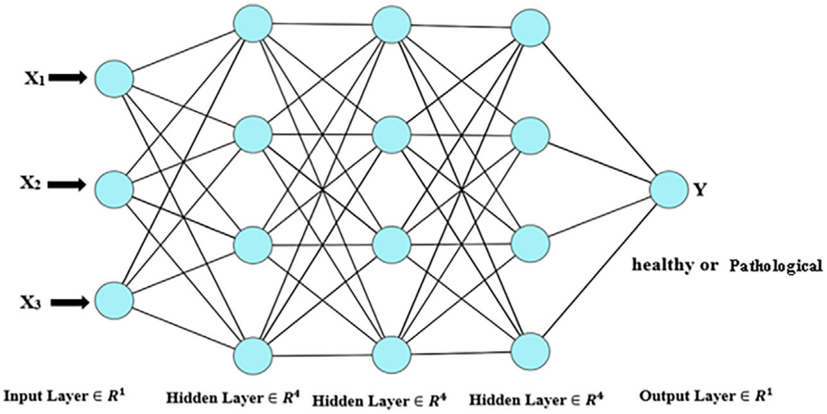

FIGURE 5 Long short-term memory (LSTM) architecture.

TABLE 7 Pseudocode for the long short-term memory (LSTM) algorithm.

```
Input: Data of selected features
Output: 0 or 1 (healthy/pathological). // in case of testing individual case
Begin
Step 1: Initialize values for all parameters in Equations (1, 2, 3): {W, b} $\in R$.
Step 2: At time t, $x_t$ is the input and $y_t$ is the output of the node.
Step 3: $f_t = \sigma (W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f)$ is the output of the forget gate at time t.
Step 4: $i_t = \sigma (W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i)$ is the output of the input gate at time t.
Step 5: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh (W_{xc} x_t + W_{hc} h_{t-1} + b_c)$ is the cell structure of the input gate at time t.
Step 6: $o_t = \sigma (W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o)$ is the output of the output gate.
Step 7: The final output $y_t$ of the node is: $y_t = o_t \odot \tanh (c_t)$
END.
```

certain type of prediction more accurately than others (Patterson & Gibson, 2017). The suggested model's effectiveness for classifying voice pathology was evaluated based on certain criteria, such as precision, accuracy (ACC), F1-score, specificity (SP), and sensitivity (SN) measures:

1. $ACC = (TP + TN)/(TP + FP + FN + TN)$
2. $SN\,(Recall) = TP/(TP + FN)$
3. $SP = TN/(TN + FP)$
4. $Precision = TP/(TP + FP)$
5. $F1 = 2TP/(2TP + FP + FN)$

In this context, the terms false negative (FN) and true positive (TP) indicate that the model correctly identified pathological samples as pathological or incorrectly identified them as healthy, respectively. In contrast, false positive (FP) and true negative (TN) refer to the model correctly identifying healthy samples as healthy or incorrectly identifying them as pathological, respectively. These outcomes are determined by the classifier.

4 | RESULTS AND DISCUSSION

In this section, the results of our proposed system of voice pathology detection and classification utilizing the LSTM and ANN classifiers on the SVD data set are presented. The prime objective of this study is to build a system to identify voice pathology by classifying features that have been extracted from voice samples into healthy and pathological. Different types of features have been explored to obtain high accuracy and the best performance of the classification process of voice samples. The following subsections present and discuss the details of the system's stage results.

4.1 | Feature extraction results

The voice signal data are in a waveform (WAV.) format, which is a kind of lossless file format that captures the closest mathematical representation of the original audio with no noticeable audio quality loss. Each signal has a frequency, amplitude, and speed of sound properties. Audio features refer to the characteristics or attributes of an audio signal that can be used as input for a classifier. The information to be extracted from audio files is only transformations of these main properties. These transformations are made to features according to mathematical deep processes on the voice signals to be in the form shown in Tables 8 and 9, so it can be easier to deal with by the later phases. Signals are broken down into windows to extract features, and then features can be extracted per window.

In this study, 10 sets of features were extracted from healthy and pathological voice signals. Some sets of spectral features extract the time domain from waveforms of the raw audio signals. Some methods of audio feature extraction focus on the frequency components of the voice signal by converting the signal into the frequency domain using the Fourier transform. Other techniques may use the short-time Fourier transform (STFT) to obtain a time-frequency representation of the signal. One specific type of time-frequency representation is the mel spectrogram, which combines both time and frequency components of the voice signal. The mel spectrogram is obtained by applying the STFT on the time domain waveform and then transforming the resulting frequency spectrum to a mel scale that is more closely aligned with human auditory perception. Our data contain healthy and pathological samples. Therefore, features are extracted from both kinds of samples in this phase to be then fed to the classifier. Table 8 shows an example of the 10 sets that contain 48 features that have been extracted from healthy samples, and Table 9 shows an example of the 10 sets that contain 48 features that have been extracted from pathological samples. In Tables 8 and 9 below, we see that the values of features that have been represented as numbers (positive and negative) vary across samples because of the type of feature type, differences in pathologies, the age of the person and the speech stream.

4.2 | Feature selection results

GA was used as a feature selector to choose the optimal group of features that would improve the system's accuracy. The 48 features from the previous phase pass through multiple steps using this algorithm. In each iteration, a number of features are selected to check its fitness. Fitness is an assignment operator used to measure the efficiency of the input features. Individuals with higher fitness are more likely to be selected for recombination. In this system, two fitness techniques are employed: ANN and LSTM algorithms. These two algorithms are chosen specifically because they are the classifiers in our system that obtain the desired features. This cycling of steps in the process is frequently repeated for each feature until reaching the best possible features that have the highest fitness (accuracy). A ZCR, MFCC, and mel spectrogram are the best three sets of features selected by the GA. It includes nine features, and these selected features are then fed to the classifier to complete the classification process of the proposed system. Examples of the selected features from the healthy and pathological samples of the SVD data set are shown in Tables 10 and 11. Values of these features differ between samples according to the difference in pathologies and each voice signal. These values include negative and positive numbers, but their difference does not affect the results because the selection is made according to features that achieve the best accuracy without linking these values and dealing with them separately. This process of selection leads to the evolution of our population features that are more suited for the system than its original form. The tables below show that feature values in the range of 300 and less with negative signs achieved better results than higher values, that is, for both classes (healthy and pathological) of samples.

TABLE 8 Features extracted from healthy voice samples.

Chroma	RMS	Spec. centroid	Spec. BW	Roll-off	ZCR	Mean	MFCC1	MFCC2	MFCC3
0.327791	0.104342	1099.3	1429.012	1295.767	0.062037	−0.00011	−253.756	167.5653	−32.8067
0.221732	0.217424	1015.261	1279.904	1341.702	0.046085	1.20E−05	−223.437	140.788	−33.0335
0.19356	0.108813	1231.622	1570.737	1615.355	0.05319	0.000254	−239.812	137.8353	−34.7449
0.232826	0.218659	955.2549	1033.97	1191.324	0.07124	4.85E−05	−268.34	136.7561	−42.2017
0.271252	0.25124	1561.16	1477.879	2894.316	0.080902	−0.00015	−140.078	105.5573	−72.6827
0.160771	0.33605	1218.67	1314.781	1541.718	0.072835	0.000367	−162.116	116.4655	−55.691
0.283559	0.113218	881.0405	932.8862	1035.837	0.056966	6.04E−05	−284.633	193.9038	−8.44746
MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13
−16.9501	−8.75988	−26.8307	42.87798	15.18643	−17.7057	−1.07925	10.57883	9.597101	−6.78289
−13.509	−28.8946	−4.68574	−9.10137	−2.38972	5.866603	−11.9318	5.502722	4.301877	−15.0677
−23.4948	−5.0129	4.674178	3.102747	14.29613	−37.1387	23.22799	−6.68988	1.991436	−15.083
−15.1609	−39.6664	−20.717	24.76455	−2.44779	−18.3952	6.132476	−6.26194	−1.81916	−19.8636
0.81972	−29.1871	−10.1285	15.65221	−5.52704	−16.9358	30.03207	1.875585	8.553572	−36.492
−14.351	−35.6387	−1.68933	11.10605	−1.82033	−19.3563	17.19264	4.39779	−12.3494	−17.3609
−25.6027	−55.172	−10.1832	5.0836	5.245618	25.10755	−6.74781	4.714509	−2.7717	5.897943
MFCC14	MFCC15	MFCC16	MFCC17	MFCC18	MFCC19	MFCC20	Spec1	Spec2	Spec3
4.286481	−13.163	−1.73596	−13.2923	5.385896	−4.27508	−1.90195	−0.18034	−0.64474	3.416104
−2.37148	2.030572	−0.19573	−6.42003	−6.60649	9.533967	−12.6451	3.238024	2.220109	5.193846
−4.12399	1.155768	2.400702	0.952423	1.387813	−4.67759	1.52274	5.644342	2.399247	−1.03599
1.687575	−22.4562	−7.94493	4.958881	−4.61655	−9.91073	−15.3312	3.25371	1.318945	−0.05469
−0.56166	−13.6113	−0.37003	4.93954	−7.31896	−1.71012	8.673904	−1.01915	−2.53609	3.856145
8.071336	−1.55437	−1.28964	0.227918	12.54687	−11.4627	8.53384	−1.37492	−2.1072	3.171114
22.83238	−26.8459	12.61108	−2.9935	−19.2605	−1.14923	18.2179	−5.90645	−1.21177	4.240194
Spec4	Spec5	Spec6	Spec7	Spec8	Spec9	Spec10	Spec11	Spec12	Spec13
1.678173	−0.51585	−11.339	−27.6663	−26.8167	−25.7519	−34.1078	−47.083	−50.4273	−47.724
1.788389	−6.5911	−18.0993	−24.3151	−22.9578	−29.2119	−36.8011	−39.1608	−44.548	−50.6271
−1.52599	−3.99987	−6.11825	−20.3779	−26.2865	−25.4443	−35.097	−47.1506	−40.8525	−38.8307
3.857064	−2.58735	−13.4919	−31.3674	−36.8487	−31.7949	−34.8524	−45.1791	−54.7439	−52.4918
3.157142	−2.58166	−1.32899	−10.5515	−19.6723	−14.9634	−20.5586	−27.896	−39.3081	−40.7676
5.283242	−3.99217	−9.66201	−22.6698	−31.5613	−25.3167	−33.2844	−40.3376	−42.6585	−48.5664
−0.1486	−7.19628	−26.3546	−32.5725	−35.5991	−37.4756	−41.1064	−42.7453	−51.6727	−59.0662
Spec14	Spec15	Spec16	Spec17	Spec18	Spec19	Spec20	Tempo	Label	
−41.9998	−56.9039	−73	−73	−73	−73	−73	130.0469	Healthy	
−52.0991	−62.3228	−73	−73	−73	−73	−73	130.0469	Healthy	
−47.1206	−67.3674	−73	−73	−73	−73	−73	142.9992	Healthy	
−57.6298	−72.6718	−73	−73	−73	−73	−73	130.0469	Healthy	
−45.1997	−71.4308	−73	−73	−73	−73	−73	142.9992	Healthy	
−52.2439	−72.195	−73	−73	−73	−73	−73	136.1992	Healthy	
−58.7446	−72.5802	−73	−73	−73	−73	−73	150.5547	Healthy	

TABLE 9 Features extracted from pathological voice samples.

Chroma	RMS	Spec. centroid	Spec. BW	Roll-off	ZCR	Mean	MFCC1	MFCC2	MFCC3
0.326471	0.193271	1378.926	2035.53	2141.635	0.044922	−0.00014	−148.173	139.2354	2.666287
0.367958	0.155119	1128.365	1613.979	1386.575	0.048154	−0.00014	−179.388	169.432	1.654767
0.330934	0.223882	962.0107	1243.838	1039.327	0.056019	−0.00035	−191.274	160.4902	1.80577
0.186065	0.133795	1571.758	1699.979	2595.397	0.068594	0.000286	−236.277	76.69358	−56.7444
0.240538	0.215664	1041.906	1146.96	1643.545	0.040931	0.000157	−183.346	177.1701	−53.6317
0.27688	0.22064	2274.011	2741.875	5757.383	0.049555	−0.00059	−84.6238	95.45142	−14.6741
0.24562	0.096182	1217.533	1677.179	2164.334	0.045309	0.000145	−218.882	154.7535	−13.4007
0.26878	0.153336	1090.507	1586.693	1361.975	0.036519	0.000453	−181.134	166.3624	−6.86547
−12.4911	−42.761	−13.4523	25.4388	−3.10384	9.867577	−11.2856	−3.4691	12.82758	−6.96764
−13.2824	−34.011	2.676587	30.21303	−9.96837	−14.505	7.898098	−6.59886	−3.19865	20.8519
MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13
−35.9988	−53.7027	−1.98838	−13.8754	1.439452	−13.2279	15.24505	−8.24172	−2.0263	−7.30409
18.30088	−28.4802	−37.7898	31.70245	13.93553	−24.9918	19.15936	−10.1958	−1.98186	16.2728
−34.116	−10.1306	6.61433	−6.87195	13.532	−3.32001	3.600418	0.243658	19.58135	−19.7451
−8.19718	−4.92989	−0.57303	−18.8871	34.30622	−28.0648	3.554334	13.87413	9.726306	−16.8976
−30.2593	−24.5015	−8.92421	14.93766	−3.60574	−9.53563	−2.18808	−7.33628	−11.4932	−22.8725
14.65853	−24.1742	−6.26272	−10.1576	6.986848	−5.58783	−3.60731	1.703899	−2.98178	−1.47048
5.239217	−28.366	21.93204	−27.8113	12.93183	−5.45313	−1.67856	0.187608	−2.38449	−3.23691
10.5563	−15.9546	4.002951	−13.7493	−7.46375	2.324546	0.805079	−5.8417	−6.0674	−2.03292
−9.29569	−1.72314	−6.91903	−6.61717	17.97591	17.79311	40.68086	−13.3633	−11.8268	−0.82571
−5.31795	−5.30262	−10.9322	−23.1766	12.93836	−7.31624	0.663352	−0.17051	−1.09409	0.200233
MFCC14	MFCC15	MFCC16	MFCC17	MFCC18	MFCC19	MFCC20	Spec1	Spec2	Spec3
0.370852	−7.05427	−7.91139	2.747212	−2.60962	−21.9139	5.892895	−0.82771	−4.7218	−6.24503
0.138012	−2.61924	−14.1366	−1.72628	−5.0316	−8.54475	9.899142	−3.98484	−10.1356	−11.6917
−6.75033	−20.8197	−30.6067	−30.9999	−31.6949	−33.0297	−33.7531	−40.0817	−42.1449	−37.3904
−6.43231	−17.4841	−32.2955	−40.8935	−31.4476	−29.1698	−42.0452	−41.2337	−47.0126	−45.4805
−15.7248	−36.7066	−39.9976	−41.7383	−35.5277	−35.9946	−43.356	−49.0158	−56.1221	−61.7659
−9.4004	−13.9169	−24.9786	−34.0091	−33.3628	−33.236	−39.4132	−40.6059	−44.2546	−53.892
−4.85668	−16.4601	−28.5199	−26.8905	−24.0428	−27.7175	−44.1379	−52.2388	−55.9699	−58.9831
−9.65389	−14.2428	−23.9681	−20.8773	−28.231	−32.1214	−33.305	−33.6762	−31.5048	−31.2556
Spec4	Spec5	Spec6	Spec7	Spec8	Spec9	Spec10	Spec11	Spec12	Spec13

(Continues)

TABLE 9 (Continued)

Chroma	RMS	Spec. centroid	Spec. BW	Roll-off	ZCR	Mean	MFCC1	MFCC2	MFCC3
−16.0389	−25.0061	−33.8282	−42.1498	−41.2525	−46.7969	−48.5313	−52.5558	−52.5116	−51.0178
−6.75033	−20.8197	−30.6067	−30.9999	−31.6949	−33.0297	−33.7531	−40.0817	−42.1449	−37.3904
−6.43231	−17.4841	−32.2955	−40.8935	−31.4476	−29.1698	−42.0452	−41.2337	−47.0126	−45.4805
−15.7248	−36.7066	−39.9976	−41.7383	−35.5277	−35.9946	−43.356	−49.0158	−56.1221	−61.7659
−9.4004	−13.9169	−24.9786	−34.0091	−33.3628	−33.236	−39.4132	−40.6059	−44.2546	−53.892
−4.85668	−16.4601	−28.5199	−26.8905	−24.0428	−27.7175	−44.1379	−52.2388	−55.9699	−58.9831
−9.65389	−14.2428	−23.9681	−20.8773	−28.231	−32.1214	−33.305	−33.6762	−31.5048	−31.2556
−16.1107	−19.794	−29.8607	−27.9693	−26.6096	−44.1883	−44.7217	−39.0713	−42.6948	−52.8526
−16.0389	−25.0061	−33.8282	−42.1498	−41.2525	−46.7969	−48.5313	−52.5558	−52.5116	−51.0178
Spec14	Spec15	Spec16	Spec17	Spec18	Spec19	Spec20	Tempo	Label	
−54.5658	−77	−77	−77	−77	−77	−77	120.4538	Pathological	
−64.2134	−77	−77	−77	−77	−77	−77	138.9992	Pathological	
−76.702	−77	−77	−77	−77	−77	−77	126.0469	Pathological	
−72.3042	−77	−77	−77	−77	−77	−77	154.9991	Pathological	
−75.4054	−77	−77	−77	−77	−77	−77	164.499	Pathological	
−54.1431	−77	−77	−77	−77	−77	−77	138.9992	Pathological	
−73.8013	−77	−77	−77	−77	−77	−77	120.4538	Pathological	
−65.0099	−77	−77	−77	−77	−77	−77	132.1992	Pathological	

TABLE 10 Samples of the selected features from healthy cases.

ZCR	MFCC1	MFCC5	MFCC11	Spec4	Spec9	Spec16	Spec17	Spec20	Label
0.062037	−253.756	−8.75988	10.57883	1.678173	−25.7519	−73	−73	−73	Healthy
0.046085	−223.437	−28.8946	5.502722	1.788389	−29.2119	−73	−73	−73	Healthy
0.05319	−239.812	−5.0129	−6.68988	−1.52599	−25.4443	−73	−73	−73	Healthy
0.07124	−268.34	−39.6664	−6.26194	3.857064	−31.7949	−73	−73	−73	Healthy
0.080902	−140.078	−29.1871	1.875585	3.157142	−14.9634	−73	−73	−73	Healthy
0.072835	−162.116	−35.6387	4.39779	5.283242	−25.3167	−73	−73	−73	Healthy
0.056966	−284.633	−55.172	4.714509	−0.1486	−37.4756	−73	−73	−73	Healthy

TABLE 11 Samples of the selected features from pathological cases.

ZCR	MFCC1	MFCC5	MFCC11	Spec4	Spec9	Spec16	Spec17	Spec20	Label
0.044922	−148.173	−42.761	−3.4691	−2.40104	−31.6949	−77	−77	−77	Pathological
0.048154	−179.388	−34.011	−6.59886	−7.50851	−31.4476	−77	−77	−77	Pathological
0.056019	−191.274	−66.1849	−8.20723	−5.57937	−35.5277	−77	−77	−77	Pathological
0.068594	−236.277	−53.7027	−8.24172	−6.51871	−33.3628	−77	−77	−77	Pathological
0.040931	−183.346	−28.4802	−10.1958	−2.66781	−24.0428	−77	−77	−77	Pathological
0.049555	−84.6238	−10.1306	0.243658	−6.27221	−28.231	−77	−77	−77	Pathological
0.045309	−218.882	−4.92989	13.87413	−10.2642	−26.6096	−77	−77	−77	Pathological
0.036519	−181.134	−24.5015	−7.33628	−11.0566	−41.2525	−77	−77	−77	Pathological

4.3 | Voice pathology classification results based on ANN

In our suggested classification system, the ANN was used as the classifier. The data were divided into two separate groups. The training group consists of ~70% of the 2041 samples for each of the data groups (vowels) used to train the ANN model. The testing group represents ~30% of the data for model testing. After that, the training data are given to the ANN classifier. Following the training phase, the training and test groups enter the prediction step. The training group has 1428 samples, whereas the testing group contains 613 samples. In terms of classification, two distinct experiments were run on the data using the identified features (all extracted features and the selected special features). The ANN model was trained first with a data set of all features and then with the features that were chosen.

Accuracy, ACC, F1-score, SP, and SN metrics were used to evaluate the results. The greatest achieved accuracies for the SVD data set while using various types of obtained features are shown in Table 12. Even though the same features were used, the accuracy levels varied for various vowels. For /a/, sentences, /i/, and /u/ vowels, the highest achievable accuracies before selection were 84.7%, 80.3%, 84.5%, and 83.7%, respectively. Table 13 displays the best accuracy obtained for the SVD data set using the nine features chosen by the GA. The accuracy, sensitivity, specificity, precision, and F1 values varied among vowel data for the same features. For /i/, /a/, /u/, and phrases, the best accuracies obtained were 78.3%, 77.3%, 76.3%, and 73.7%, respectively. The accuracy curves of the proposed ANN model's results are shown in Figure 6 for the data of all extracted features in which the accuracy of the trained samples was ~80% for different vowels, as shown in the blue curve below, while in Figure 7, the accuracy curves of ANN with the selected features did not show significant improvement when compared with samples before feature selection. The confusion matrix was used to evaluate the ANN model. Figure 8 displays the suggested method's performance utilizing four different measures. True positive (TP) cases are those in which the procedure accurately identifies a pathological sample as pathological. True negative (TN) cases are those in which the algorithm accurately identifies a true healthy sample as healthy. False-positive (FP) refers to situations in which the procedure incorrectly labels a healthy sample as unhealthy. Finally, false-negatives (FNs) refer to situations in which the system incorrectly labels an actual pathological sample as healthy.

4.4 | Voice pathology classification results based on LSTM

The results of the previous stage involved using an ANN for classifying voice pathology, but it did not provide the desired accuracy compared to old existing studies. As the results did not meet the required level of performance, another algorithm was used as the classifier in the proposed identification system to achieve better results. The new algorithm is LSTM. The data were divided into two groups. Training was ~70% of the

TABLE 12 Results of the artificial neural network (ANN) model before feature selection.

Features	No of features	SVD	ACC %	SN %	SP %	Precision %	F1%
LSZCR+ Spectral + Chroma+ Tempogram+ MFCC+ RMS+ Mel spectrogram+ Mean	48	/a/	**84.7**	66.5	93.8	84.6	74.5
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	Sentences	80.3	67.0	86.9	72.3	69.5
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	/i/	84.5	69.4	92.1	81.7	75.0
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	/u/	83.7	59.2	96.0	88.4	70.9

Note: Bold indicated best result or value.

TABLE 13 Results of the artificial neural network (ANN) model after feature selection.

Features	No of features	SVD	ACC%	SN%	SP%	Precision%	F1%
ZCR+ MFCC+ Mel spectrogram	9	/a/	77.3	74.2	78.8	64.0	68.7
ZCR+ MFCC+ Mel spectrogram	9	Sentences	73.7	33.5	94.1	74.2	46.2
ZCR+ MFCC+ Mel spectrogram	9	/i/	**78.3**	88.3	73.2	62.5	73.2
ZCR+ MFCC+ Mel spectrogram	9	/u/	76.3	48.5	90.4	71.9	57.9

Note: Bold indicated best result or value.

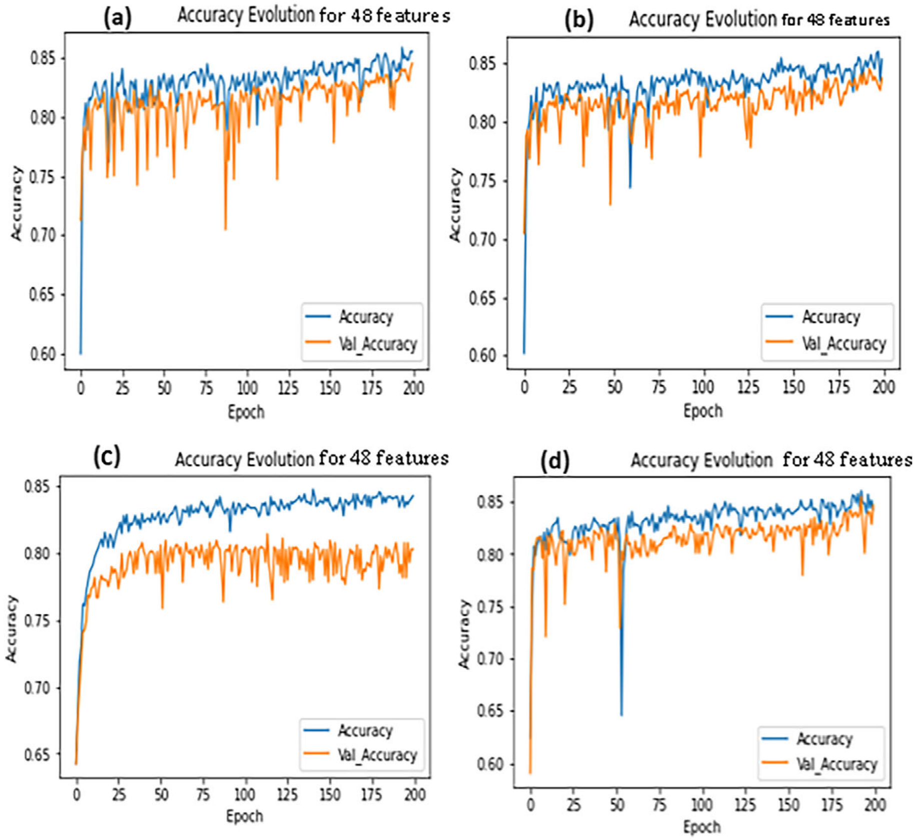

FIGURE 6 Accuracy curves of artificial neural network (ANN) with 48 features: (a) /i/ vowel samples. (b) /u/ vowel samples. (c) Sentence samples. (d) /a/ vowel samples.

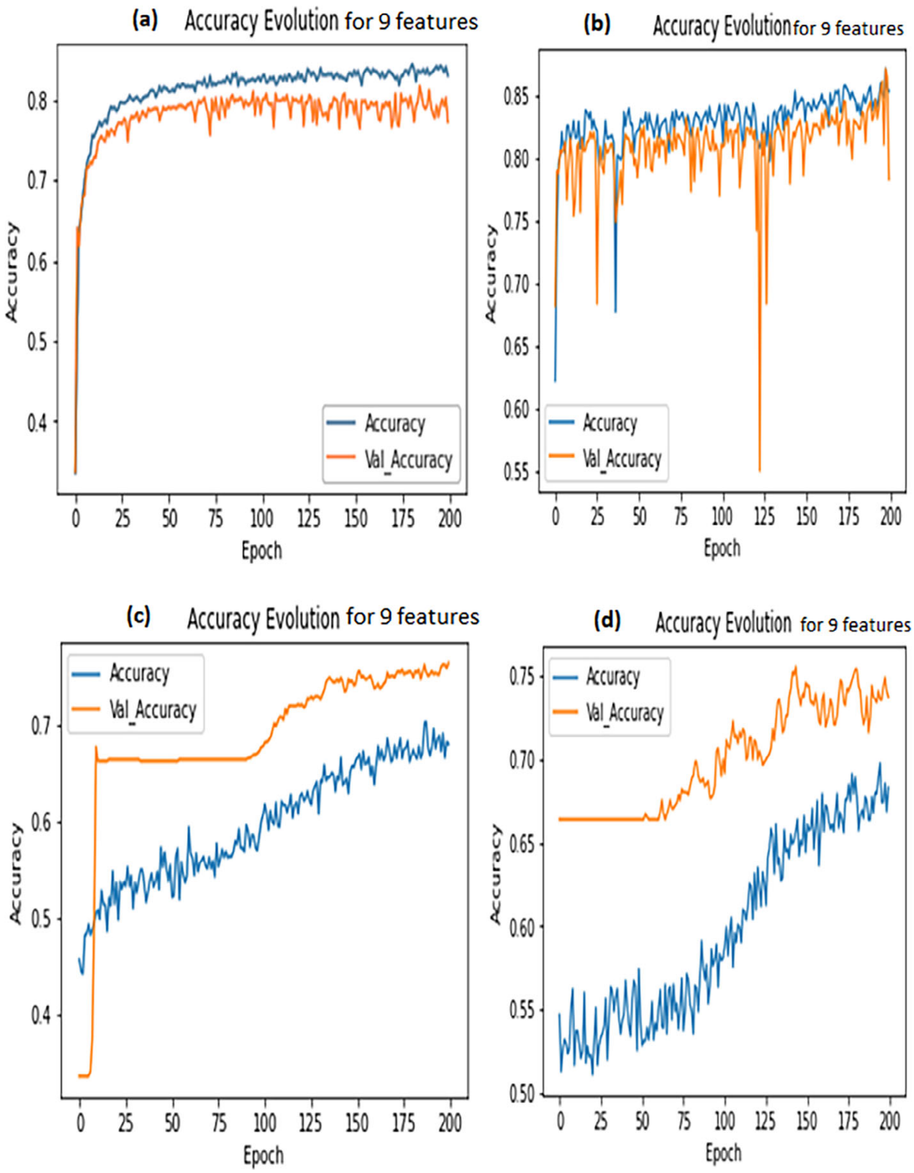

FIGURE 7 Accuracy curves of artificial neural network (ANN) with selected features: (a) /a/ vowel samples. (b) /i/ vowel samples. (c) /u/ vowel samples. (d) Sentence samples.

2041 samples for each one of the data groups (vowels) to train the LSTM model. Testing was ~30% of the data for model testing. Training data were then fed to the LSTM classifier. After the training phase, the training and testing groups entered the prediction stage. The training group included 1428 samples, while 613 are the testing samples. For the classification stage, we assessed two diverse experiments on data depending on the features (all the extracted features and the best-selected feature). The LSTM model was trained first with a data set of all features and then with the selected features. The results were evaluated according to precision, ACC, SP, SN, and F1 measures. Table 14 displays the highest obtained accuracies on the SVD dataset from the various feature extraction methods. This chart shows that for the same set of features, the accuracy rates for different vowels ranged widely. Generally, the highest achieved accuracies for features before the selection were 73.4%, 71.5%, 71.8%, and 71.9% for /a/, sentences, /i/ and /u/ vowels, respectively. The performance of the system improved after feature selection by the GA, and the best accuracies for the SVD data set were achieved with the nine selected features. Accuracy, sensitivity, specificity, precision and F1 measures also differed among data of each vowel for the same features. The highest accuracy achieved was 99.3%, 99.0%, 99.3%, and 99.2% for /a/, /i/, /u/, and sentences, respectively, as shown in Table 15. Accuracy curves of the presented LSTM model results are shown in Figure 9 for data of all features extracted and with the best-selected features. The blue curve shows the training data accuracy, and the orange curve shows the testing data accuracy. A confusion matrix was used in evaluating the LSTM, as shown in Figure 10, where TP refers to cases where the proposed method correctly identifies a real pathological sample as pathological. TN denotes cases where the proposed method correctly classifies a real healthy sample as healthy. FP represents cases where the proposed method incorrectly identifies a real healthy sample as pathological. FN

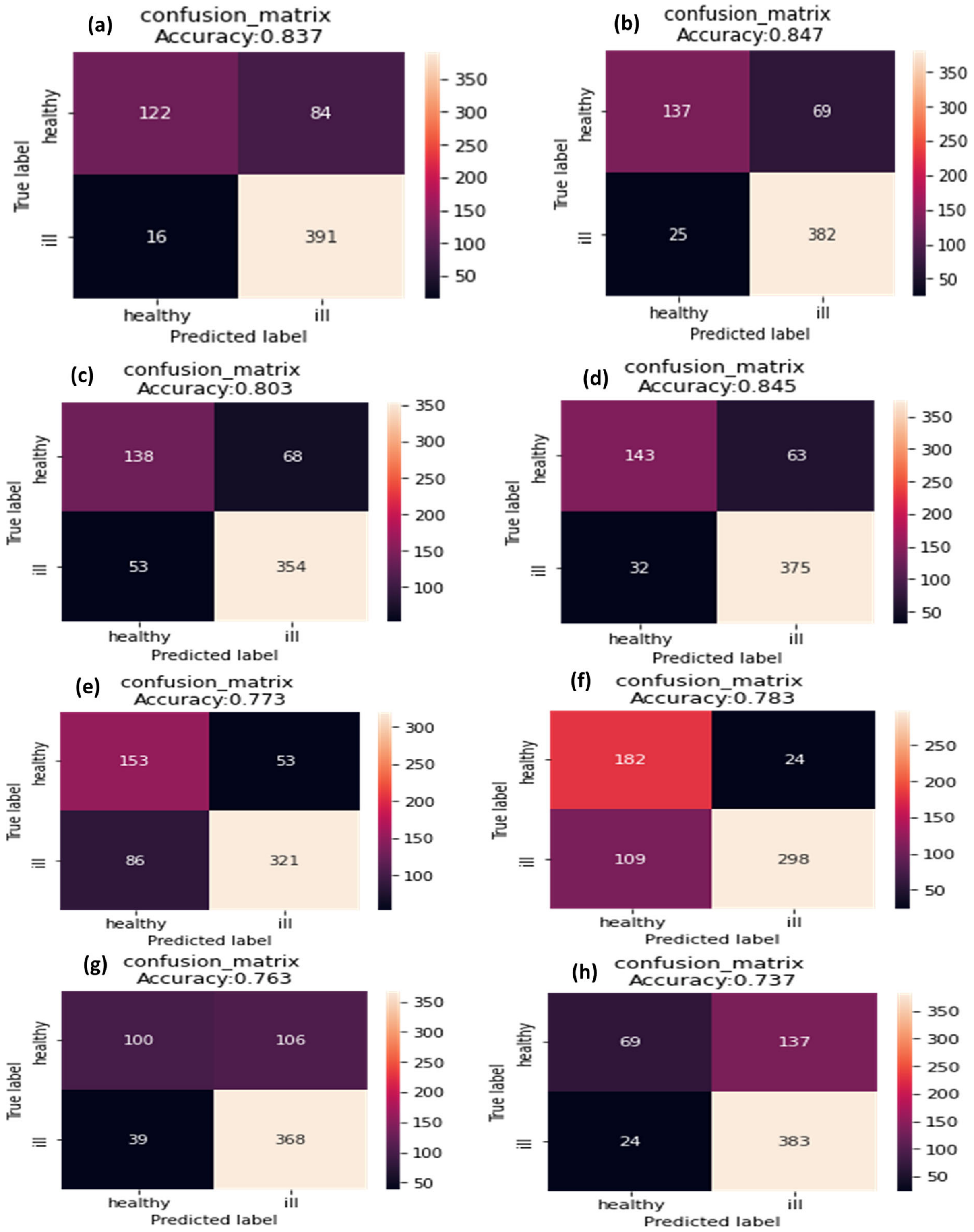

FIGURE 8 Confusion matrices of artificial neural network (ANN): (a–d) (48 Features). (e–h) [9 features].

signifies cases where the proposed method incorrectly classifies a real pathological sample as healthy. From the confusion matrix figure, we determined that the best identification accuracies have high values of TP and TN parameters with fewer values of FN and FP parameters. This combination of features was used for the first time together in our proposed system. The LSTM model outperformed the ANN model, as shown in Tables 14 and 15.

TABLE 14 Results of the long short-term memory (LSTM) model before feature selection.

Features	No of features	SVD	ACC %	SN %	SP %	Precision %	F1%
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ RMS+ Mel spectrogram+ Mean	48	/a/	**73.4**	26.2	97	83	39.8
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	Sentences	71.5	18.4	98	84.4	30.2
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	/i/	71.8	25.2	95.3	73.2	37.5
ZCR+ Spectral + Chroma+ Tempogram+ MFCC+ Mel spectrogram+ Mean	48	/u/	71.9	20.9	97.7	82.6	33.4

Note: Bold indicated best result or value.

TABLE 15 Results of the long short-term memory (LSTM) model after feature selection.

Features	No of features	SVD	ACC%	SN%	SP%	Precision%	F1%
ZCR+ MFCC+ Mel spectrogram	9	/a/	99.3	99.5	99.2	98.5	98.9
ZCR+ MFCC+ Mel spectrogram	9	Sentences	99.2	99.4	99.0	97.8	98.6
ZCR+ MFCC+ Mel spectrogram	9	/i/	99.0	99.5	98.7	97.6	98.5
ZCR+ MFCC+ Mel spectrogram	9	/u/	**99.3**	99.0	99.5	99.0	99.0

Note: Bold indicated best result or value.

4.5 | Discussions

Voice disorders are common and significant because they disrupt people's ability to communicate with others. Paralysis of the vocal cords, drug abuse, swelling, and improper use of the voice are some of the most common causes of voice pathologies. The challenge of voice pathology identification can be tackled with either classic machine learning or deep learning techniques, or a combination of the two. To determine whether a voice transmission is healthy or not, it must first undergo extraction of features and analysis, two phases of machine learning. The technique has some disadvantages, such as requiring human assistance in picking the right speech features to use and/or deciding on the best classification approach. Deep learning-based techniques that autonomously pull features for better classification performance may be preferable when creating speech pathology detection systems to avoid these problems.

Our system in this study used only voice signals from the SVD data set, unlike other similar studies. Some of them converted voice signals into spectrograms and dealt with them as photos in classification and feature extraction. However, others used both voice and EGG signals available in SVD and combine them for only a limited number of pathologies. This is because not all vowel samples of SVD were provided in EGG form. Therefore, the number of voice and EGG signals were extremely small for the same vowels and pathology. The proposed system in this study passed through multiple modifications until reaching the final pattern that overcomes the previously used methods. These modifications included using different ratios for training and testing groups during splitting and statistical-based feature selection methods, but the results did not improve. Therefore, the GA via ANN and LSTM was tested to select many groups of features in each cycle until obtaining the features with the best outcomes. These selected features (MFCC, ZCR, mel spectrogram) had a unique combination that was not tested in earlier related works. The best results achieved by our system are (99.3%, 99%, 99.5%, 99%, and 99%) for /u/ vowel samples in neutral pitch, (99.3%, 99.5%, 99.2%, 98.5%, and 98.9%) for /a/ vowel samples in high pitch, (99%, 99.5%, 98.7%, 97.6%, and 98.5%) for /i/ vowel samples in neutral pitch and (99.2%, 99.4%, 99%, 97.8%, and 98.6%) for the sentences samples in terms of accuracy, sensitivity, specificity, precision, and F1 measures, respectively. The performance of the proposed system using the LSTM classifier and the unique set of features was compared with other methods in terms of precision, ACC, SP, F1-score, and SN for voice pathology classification. The systems that we considered for comparison were proposed by Al-Nasheri et al. (2017), Dankovičová et al. (2018), Al-Dhief, Latiff, et al. (2021), Omeroglu et al. (2022). These systems did not use all data in SVD; rather, they used only the /a/ vowel samples. The system by Al-Nasheri et al. (2017) utilized entropy, peak and lag features, the system by Al-Dhief, Latiff, et al. (2021) used MFCC features, the work by Omeroglu et al. (2022) used deep and handcrafted features, and by Dankovičová et al. (2018), shimmer, jitter, spectral roll-off, and MFCC features were used. Our proposed method with LSTM outperformed all methods, where it achieved the highest accuracy, sensitivity, specificity, precision and F1 in the identification of voice pathology with a set of features that were not previously used in any other system. However, the ANN classifier did not show any further improvement. Table 16 shows the best accuracies of the systems.

In this study, extracting and selecting good features from voice signals that lead to better results was a complicated and time-consuming procedure. This process required an unlimited number of trials to obtain the most compatible features with the classifier. Another form of the

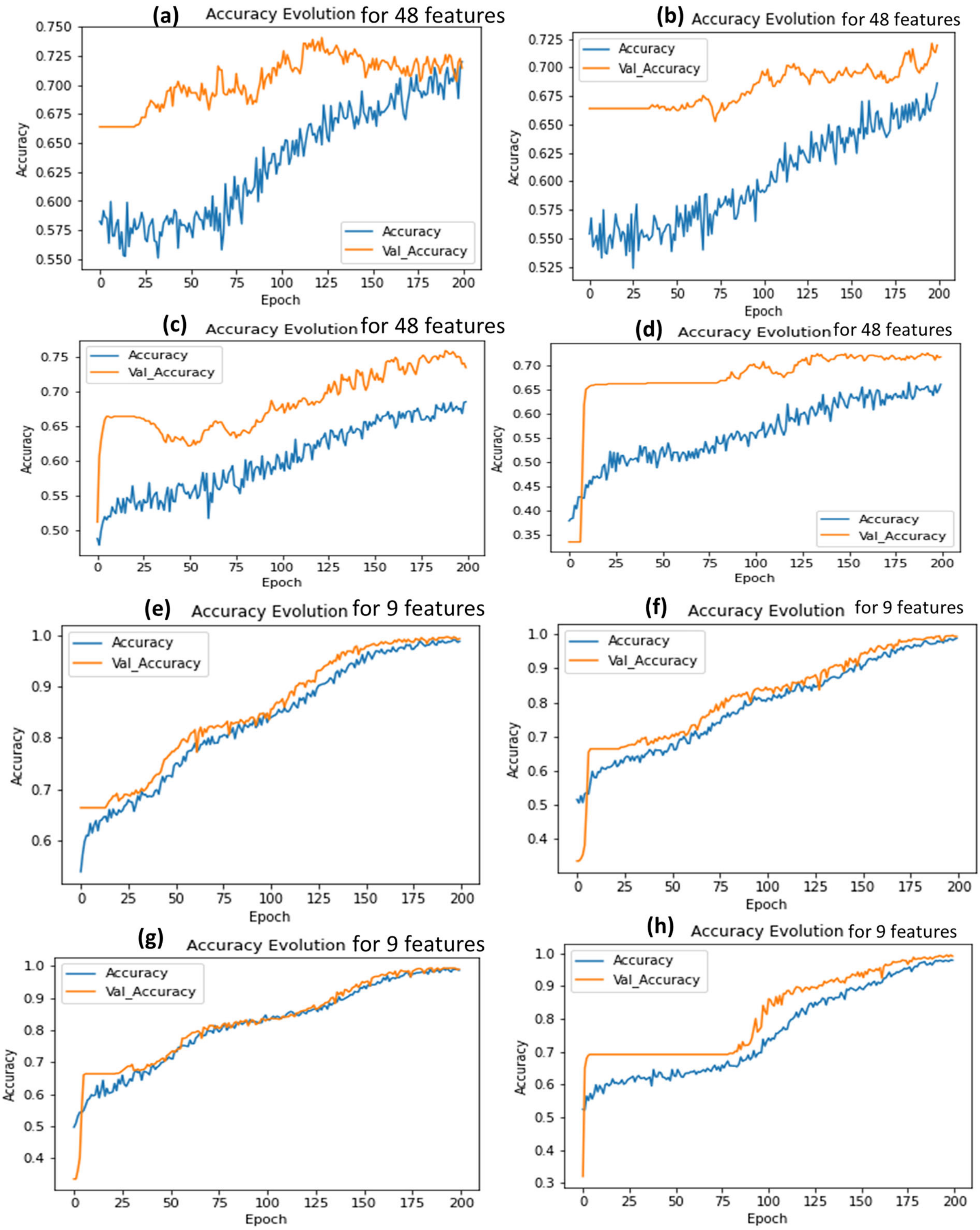

FIGURE 9 Accuracy curves of long short-term memory (LSTM): (a) Sentence samples. (b) /u/ vowel samples. (c) /a/ vowel samples. (d) /i/ vowel samples. (e) /a/ vowel samples. (f) /u/ vowel samples. (g) /i/ vowel samples. (h) Sentence samples.

difficulties we faced was associated with the SVD data set. It had an unequal number of samples (healthy and pathological). Another limitation was the presence of some uncommon pathology samples, making it impossible for the classifier to identify them and leading to unsatisfactory test results. It is hoped that by increasing the size of the data set, the performance of the proposed method can be enhanced in future research. Additionally, the feasibility of incorporating this framework into a mobile health system for monitoring and treating patients exhibiting voice pathology symptoms will be explored. In conclusion, to increase the proposed approach's practicality in clinical practice, subjective tests should be carried out by medical specialists.

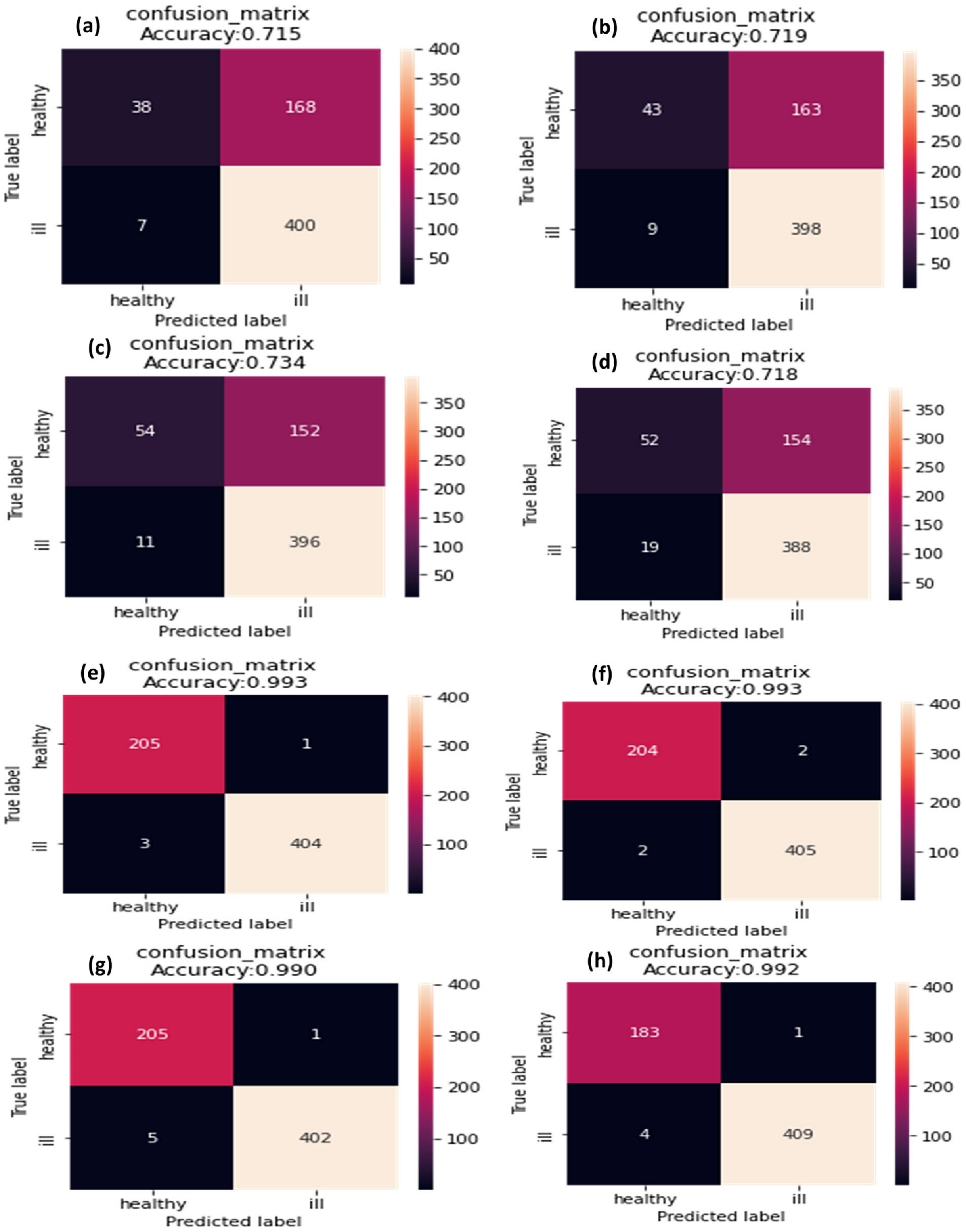

FIGURE 10 Confusion matrices of long short-term memory (LSTM): (a–d) (48 features). (e–h) (9 features).

5 | CONCLUSION

In this study, an automated system for voice pathology identification using an SVD data set was proposed. A total of 48 features (ZCR, spectral centroid, spectral bandwidth, spectral roll-off, chroma, MFCC, mel spectrogram, mean, RMS, and thermogram) were extracted from voice signals to build a system to identify healthy voice samples from pathological voice samples by ANN and LSTM classification systems. Vowel (/a/, /i/, and

TABLE 16 Comparison of our proposed system with state-of-the-art methods.

Author (s), year	Data set	Method	Feature selection	ACC%	SN %	SP %	Precision%	F1%
Al-Nasheri et al. (2017)	SVD	SVM	-	92.79	91.22	94.27	-	-
Al-Dhief, Latiff, et al. (2021)	SVD	SVM	-	84.37	80.95	90.90	-	-
Omeroglu et al. (2022)	SVD	SVM	InfoGainAttributeEval function in WEKA	90.10	92.9	84.6	-	92.57
Dankovičová et al. (2018)	SVD	RFC	PCA	91.3	-	-	-	-
Proposed Approach	**SVD**	**ANN**	**GA**	**84.7**	**66.5**	**93.8**	**84.6**	**74.5**
Proposed Approach	**SVD**	**LSTM**	**GA**	**99.3**	**99**	**99.5**	**99**	**99**

Note: Bold indicated best result or value.

/u/) and sentence voice samples in the SVD data set were tested in the presented system. Feature selection by using the genetic algorithm (GA) enhanced the results and obtained higher accuracies. Due to the insufficiently high initial results obtained using an ANN as a classifier, the LSTM classifier was used as the classifier in the system. The proposed LSTM system with the optimal features selected by GA (ZCR, mel spectrogram, and MFCC) achieved 99.3%, 99%, 99.5%, 99%, and 99% accuracy, sensitivity, specificity, precision, and F1 measures, respectively. The feature selection step using GA improved the results because it increased accuracy and other measures with sets of features that have never been used in previous voice pathology identification systems. The practical advantage of this method is the ability to distinguish healthy from unhealthy individuals using fewer features, which is sufficiently accurate depending on the sample voice signals, making it a less invasive method for diagnosing voice pathologies. We suggest that future studies include a larger sample size, more variable pathologies, an assessment of the severity of the disease, and a specific diagnosis for the pathological samples in addition to involving other types of classifiers or certain features in the tested system that may improve the results.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

In this study, the SVD is used with the healthy and disordered voices samples are included in this data set. SVD is an available free database (Barry & Pützer, 2007) maintained by the Institute of Phonetics at Saarland University. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Belal Al-Khateeb https://orcid.org/0000-0003-3066-0790

Mazin Abed Mohammed https://orcid.org/0000-0001-9030-8102

REFERENCES

Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2022). A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. *Journal of Intelligent Systems*, *31*(1), 855–875.

Al-Dhief, F., Latiff, N., Baki, M., Malik, N., Sabri, N., Naseer, N., & Albadr, M. (2021). Voice pathology detection using support vector machine based on different number of voice signals. In *26th IEEE Asia-Pacific conference on communications (APCC)* (pp. 1–6). IEEE.

Al-Dhief, F. T., Baki, M. M., Latiff, N. M. A. A., Malik, N. N. N. A., Salim, N. S., Albader, M. A. A., Mahyuddin, N. M., & Mohammed, M. A. (2021). Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, *9*, 77293–77306.

AL-Dhief, F. T., Latiff, N. M. A. A., Malik, N. N. N. A., Sabri, N., Baki, M. M., Albadr, M. A. A., Abbas, A. F., Hussein, Y. M., & Mohammed, M. A. (2020, November). Voice pathology detection using machine learning technique. In *In 2020 IEEE 5th international symposium on telecommunication technologies (ISTT)* (pp. 99–104). IEEE.

Al-Dhief, F. T., Latiff, N. M. A. A., Malik, N. N. N. A., Salim, N. S., Baki, M. M., Albadr, M. A. A., & Mohammed, M. A. (2020). A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms. *IEEE Access*, *8*, 64514–64533.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Malki, K. H., Mesallam, T. A., & Ibrahim, M. F. (2017). Voice pathology detection and classification using autocorrelation and entropy features in different frequency regions. *IEEE Access*, *6*, 6961–6974.

Barry, W. J., & Pützer, M. (2007). *Saarbrucken voice database, institute of phonetics*. University of Saarland. http://www.stimmdatenbank.coli.uni-saarland.de/

Brownlee, J. (2017). Long short-term memory networks with python, v1.0. *Machine Learning Mastery*, 245.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79.

Dahmani, M., & Guerti, M. (2018). Glottal signal parameters as features set for neurological voice disorders diagnosis using k-nearest neighbors (knn). In *2018 2nd international conference on natural language and speech processing (ICNLSP)* (pp. 1–5). IEEE.

Dankovičová, Z., Sovák, D., Drotár, P., & Vokorokos, L. (2018). Machine learning approach to dysphonia detection. *Applied Sciences*, *8*(10), 1927.

Gómez-García, J. A., Moro-Velázquez, L., & Godino-Llorente, J. I. (2019). On the design of automatic voice condition analysis systems. Part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, *51*, 181–199.

Hardesty, L. (2017, 14 April). Explained: Neural networks. *MIT News The Office*.

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, *33*(6), 947.e11–947.e33.

Hossain, M., Muhammad, G., & Alamri, A. (2017). Smart healthcare monitoring: A voice pathology detection paradigm for smart cities. *Multimedia Systems*, *25*, 565–575.

Islam, R., Tarique, M., & Abdel-Raheem, E. (2020). A survey on signal processing based pathological voice detection techniques. *IEEE Access*, *8*, 66749–66776.

Kadiri, S. R., & Alku, P. (2019). Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 367–379.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Ghani, K. A., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., & Al-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, *10*(11), 3723.

Muhammad, G., & Alhussein, M. (2021). Convergence of artificial intelligence and internet of things in smart healthcare: A case study of voice pathology detection. *IEEE Access*, *9*, 89198–89209.

Naikare, K., Nirmal, J. H., & Lad, N. (2018). Classification of voice disorders using i-vector analysis. *International Conference on Communication Information and Computing Technology (ICCICT)*, *2018*, 1–7.

Omeroglu, A., Mohammed, H., & Oral, E. (2022). Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion. *Engineering Science and Technology, an International Journal*, *36*, 101148.

Patterson, J., & Gibson, A. (2017). *Deep learning-a Practitioner's practice* (First ed.). O'Reilly Media.

Syed, S., Rashid, M., Hussain, S., & Zahid, H. (2021). Comparative analysis of CNN and RNN for voice pathology detection. *BioMed Research International*, *2021*, 1–8.

Wu, H., Soraghan, J., Lowit, A., & di Caterina, G. (2018). Convolutional neural networks for pathological voice detection. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1–4). IEEE.

Yang, J., & Horie, R. (2015). An improved computer Interface comprising a recurrent neural network and a natural user Interface. *Procedia Computer Science*, *60*, 1386–1395.

AUTHOR BIOGRAPHIES

Nuha Qais Abdulmajeed received the B.Sc. degree in computer science from the University of Anbar, Iraq and currently, master student at College of Computer Science and Information Technology, University of Anbar, Iraq. Her research interests include artificial intelligence, biomedical computing, and optimization.

Prof. Dr. Belal Al-Khateeb (first class) degree in computer science from Al-Nahrain University, Baghdad, IRAQ, in 2000, and the M.Sc. degree in computer science from Al-Nahrain University, Baghdad, IRAQ, in 2003, and the Ph.D. degree from the School of Computer Science, University of Nottingham, Nottingham, U.K., in 2011. He is currently a professor at the College of Computer Science and Information Technology, University of Anbar. He has published over 85 refereed journal and conference papers. His current research interests include deep learning particularly in health care, evolutionary and adaptive learning particularly in computer games, expert systems, and heuristics and meta/hyper-heuristics. He has a particular interest in computer games programming. Prof. Al-Khateeb is a reviewer of twenty-five international journals (including many clarivate journals) and thirty conferences.

Mazin Abed Mohammed received the B.Sc. degree in computer science from the University of Anbar, Iraq, in 2008, the M.Sc. degree in information technology from UNITEN, Malaysia, in 2011, and the Ph.D. degree in information technology from UTeM, Malaysia, in 2019. He is currently a Lecturer with the College of Computer Science and Information Technology, University of Anbar. His research interests include artificial intelligence, biomedical computing, and optimization.

How to cite this article: Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2025). Voice pathology identification system using a deep learning approach based on unique feature selection sets. *Expert Systems*, *42*(1), e13327. https://doi.org/10.1111/exsy.13327