Diagnosis of Pathological Speech with Efficient and Effective Features for Long Short-Term Memory Learning

Tuan D. Pham*, Simon B. Holmes, Lifong Zou, Mangala Patel, Paul Coulthard

Barts and The London Faculty of Medicine and Dentistry

Queen Mary University of London

Turner Street, E1 2AD, London, UK

*Corresponding author (email: tuan.pham@qmul.ac.uk)

Abstract

The majority of voice disorders stem from improper vocal usage. Alterations in voice quality can also serve as indicators for a broad spectrum of diseases. Particularly, the significant correlation between voice disorders and dental health underscores the need for precise diagnosis through acoustic data. This paper introduces effective and efficient features for deep learning with speech signals to distinguish between two groups: individuals with healthy voices and those with pathological voice conditions. Using a public voice database, the ten-fold test results obtained from long short-term memory networks trained on the combination of time-frequency and time-space features with a data balance strategy achieved the following metrics: accuracy = 90%, sensitivity = 93%, specificity = 87%, precision = 88%, F_1 score = 0.90, and area under the receiver operating characteristic curve = 0.96.

1 Introduction

The involvement of speech pathology in head and neck cancer, and oral-maxillofacial surgery has recently been highlighted by the Michigan Medicine [1]. It is known that voice disorders stem from a wide array of factors, either present since birth or acquired later in life. These factors encompass physiological or anatomical irregularities in the upper airway due to underlying neurological conditions, trauma, and head and neck cancer.

Roughly 30% of adults might encounter difficulties linked to voice disorders [1]. The consequences of voice disorders on quality of life can lead to social seclusion and hinder one's professional and personal pursuits. The field of speech-language pathology also encompasses the specialized management of head and neck disorders. Within this domain, individualized assessment, treatment, and education are provided to patients dealing with diseases or dysfunctions in the structures of the head and neck, which includes the mandible, maxilla, soft and hard palates, nose, cheek, lips, tongue, and throat [1].

The development of speech within the evolving craniofacial complex relies significantly on the structural and functional integrity of this complex [2]. While development initiates during prenatal stages, a substantial portion of communicative language potential unfolds postnatally and hinges on neurological and cognitive well-being. Both genetic and acquired abnormalities can lead to significant dental misalignments, potentially impacting speech production. Certain investigations indicated a connection between temporomandibular disorders, voice, and oral health-related quality of life in women [3]. Thus, addressing maxillofacial disorders typically necessitates a combined approach involving orthognathic surgery and orthodontic interventions [4]. Optimal patient care requires a truly interdisciplinary approach, ensuring comprehensive and well-coordinated treatment. Effective and efficient rehabilitation plans and schedules can be crafted when speech pathologists and maxillofacial surgeons collaborate closely to synchronize their treatment efforts

[2, 5, 6].

In fact, oral health challenges in the field of speech pathology were pointed out decades ago [7]. Assessment of the spectral characteristics of the vowels /a/, /i/, and /u/ was carried out in a group of 24 children [8]. The authors found statistically significant variations in fundamental frequency when analyzing the individual progress of patients with bilateral cleft lip, jaw, and palate conditions. The fundamental frequency and analysis of the first formant demonstrated their effectiveness in characterizing the vocal timbre of individuals with cleft conditions.

In pediatric dentistry, it was investigated that children with voice disorders tend to exhibit a greater prevalence and more pronounced malocclusions than those with ordinary speech development [9]. Moreover, dental medicine is deeply involved in modifying and repairing oral structures to combat the effects of disease and developmental irregularities. Given that a significant part of speech articulation occurs within the oral cavity, any modifications or restorations of these structures can impact speech, the extent of which depends on the location and extent of the alteration. A literature review highlighted the role of pediatric dentists in the early identification and management of speech impairments [10].

The association between speech disorders and dental medicine addressed above highlights the importance of developing a capacity for precise diagnosis of pathological speech. It has been realized that physicians currently face a deficit in diagnostic tools for voice disorders, and the incorporation of artificial intelligence (AI) tools into their toolkit has the potential to expedite diagnoses [11]. A recent review on machine learning methods for diagnosing voice disorders has also been reported with a focus on nonlaryngeal aerodigestive and neurological disorders [12].

Most recently, to tackle this issue, a recent work [13] has suggested the use of a transfer learning framework, which combined a pre-trained convolutional neural network called

OpenL3 with a support vector machine classifier for the automatic identification of multi-class voice disorders. Initially, the Mel spectrum of the provided voice signal was extracted, similar to traditional speech feature extraction methods. Subsequently, this Mel spectrum was fed into the OpenL3 network to generate high-level feature embeddings. Powered by deep learning in AI, to distinguish between healthy and pathological voices through the analysis of spectrogram images obtained from the recordings of the vowel /a/, a convolutional neural network model was integrated into a mobile health application, enabling a user-friendly and portable tool for evaluating voice disorders [14]. An alternative approach [15] combined three vocal attributes: chroma, mel spectrogram, and mel frequency cepstral coefficient. The researchers employed a deep neural network to detect voice disorders, utilizing the vowels /a/, /i/, and /u/ articulated at various pitch levels–high, low, and average.

This study presents a substantial extension of a recent work [16] on the selection of efficient and effective features for long short-term memory (LSTM) network learning on speech signals for diagnosing voice disorders presented at the *2023 IEEE Conference on Artificial Intelligence*. Here, the proposed approach explores various feature extraction techniques to address the computational challenges associated with the LSTM. These features encompass derivations from both time-frequency and time-space domains, as well as wavelet time scattering networks.

The rest of this paper is organized as follows. Section 2 outlines the voice dataset used, consisting of both healthy and pathological cases. Section 3 details the techniques applied for extracting features and classifying patterns in the diagnosis of voice disorders. The outcomes of testing multiple models for the diagnosis are presented in Section 4. Finally, Section 5 provides concluding insights from this research, along with potential avenues for future exploration.

2 Data

In this study, the VOICED (VOice ICar fEDerico II) database [17] was employed to demonstrate the effectiveness of the proposed approach in addressing a challenging biomedical classification problem. The database comprises acoustic recordings of the vowel "a" lasting five seconds, featuring two distinct groups of participants: healthy individuals and those with pathological conditions. Both groups consist of male and female subjects ranging in age from 18 to 70 years.

Specifically, the dataset contains 61 recordings from healthy voices and 147 from pathological voices, with variable signal lengths about 35,000 time points. Among these, 21 recordings pertain to healthy male voices and 51 to pathological male voices. For female voices, there are 40 healthy voice recordings and 96 pathological voice recordings. To ensure the reliability of the dataset, both healthy and pathological voices underwent rigorous clinical evaluation by medical experts. Diagnoses adhered to the guidelines outlined in the SIFEL protocol, which is a clinical protocol endorsed by the Italian Society of Phoniatrics and Logopaedics.

All recordings took place in a noise-free environment, using Vox4Health technology. The distance between the mobile device and the subjects was approximately 20 cm, at an angle of approximately 45 degrees. All participants received instructions to articulate the vowel in a natural manner, and the acquired signals were filtered to eliminate noise during the recording process. The database includes voice recordings featuring various forms of pathological conditions, which are classified as 1) hyperkinetic dysphonia, 2) hypokinetic dysphonia, and 3) reflux laryngitis. These types of pathology are briefly described as follows [18]:

- *Hyperkinetic dysphonia* represents a frequently encountered clinical condition, especially among individuals engaged in vocally demanding professions. This disorder is

marked by excessive muscular contractions within the pneumo-phonic apparatus. It results in a strained, high-pitched voice, diminished frequency modulation, and a noticeable harshness in vocal delivery. Additionally, the increased resistance of the vocal folds to the expiratory airflow intensifies the effort required for phonation, leading to disruptions in respiratory patterns. Numerous conditions fall under this category, including vocal fold nodules, Reinke's edema, chorditis, rigid vocal folds, polyps, and prolapse.

- *Hypokinetic dysphonia* is characterized by reduced vocal fold adduction during the respiratory cycle, especially during inhalation, resulting in airflow obstruction within the larynx. This incomplete vocal fold closure results in a weak and breathy voice. Interestingly, in cases of hypokinetic dysphonia, voice quality improves with increased vocal intensity, which can potentially lead to improper vocal strain. Conditions falling under the umbrella of hypokinetic dysphonia include dysphonia of the vocal fold groove, adduction deficits, presbiphonia, glottic insufficiency, vocal fold paralysis, conversion dysphonia, laryngitis, and extraglottic air leakage.

- *Reflux laryngitis* refers to an inflammation of the larynx triggered by the regurgitation of stomach acid into the esophagus. The primary symptom typically observed is persistent hoarseness, although additional symptoms may manifest to varying degrees, including pharyngitis, occasional coughing fits, nighttime coughing, asthma, nocturnal laryngeal spasms, and halitosis.

3 Methods

The process of extracting time-frequency and time-space features from time series data for LSTM-based classification was initially introduced in [19], and is further elaborated upon

here. Additionally, the fundamental concept of wavelet time scattering as a feature extraction technique is introduced to provide insights into the implementation of the proposed approach for diagnosing pathological speech signals.

3.1 Extraction of time-frequency features with instantaneous frequency and spectral entropy

The instantaneous frequency (IF) of a non-stationary signal is a time-dependent parameter that corresponds to the mean of the frequencies, denoted as f, within the evolving signal as it progresses through various time points t [20]. To estimate the IF of a signal at a given sampling rate, the IF function calculates the power spectrum of the spectrogram, denoted as $P(t, f)$, and then estimates the IF using the following expression:

$$IF(t) = \frac{\int_{-\infty}^{\infty} fP(t,f)df}{\int_{-\infty}^{\infty} P(t,f)df}. \tag{1}$$

The power spectrum quantifies the strength of a signal at a specific frequency f. In the case of a periodic signal, peaks are observed at the fundamental frequency and its harmonics within the spectrum. Quasiperiodic signals exhibit peaks at linear combinations of related frequencies, while chaotic signals result in the presence of broad-band components in the spectrum. However, in practical scenarios, it is impossible to determine the exact power spectrum because real signals are not infinitely long but instead measured over a finite time interval. Consequently, it becomes necessary to estimate the power spectrum numerically and was technically described in [19].

Spectral entropy (SE) of a signal provides insight into its spectral power distribution [20]. The SE treats the normalized power distribution in the frequency domain as a probability distribution and computes its Shannon entropy. In this context, the Shannon entropy is referred to as the spectral entropy of the signal. The probability distribution at a specific

time t, where $0 \leq t \leq T$, and frequency point z, denoted as $p(t,z)$, is computed as follows.

$$p(t,z) = \frac{P(t,z)}{\sum_f P(t,f)}. \tag{2}$$

The spectral entropy at time t, denoted as $SE(t)$, is determined as

$$SE(t) = -\sum_{z=1}^{Q} p(t,z) \log_2 p(t,z). \tag{3}$$

where Q is the total frequency points.

In this study, to extract the IF and SE features of the speech signals, the parameters were specified as follows: range of $f = [0, fs/2]$, and sampling rate fs = 300 Hz.

3.2 Extraction of time-space features with fuzzy recurrence plots and spatial entropy

A fuzzy recurrence plot (FRP) [21] is a visualization technique used in the analysis of time series data. It is derived from the concept of recurrence plots (RPs) [22], which are used to study nonlinear dynamics in time series. FRPs extend this idea by introducing a degree of fuzziness or uncertainty into the recurrence analysis.

A traditional RP is constructed by comparing each point in the phase space of a dynamical system with every other point and determining whether they are close enough based on some predefined distance metric. If two points are sufficiently close, a black dot that represents "recurrence" is marked on the 2D plot. This is typically represented as a binary plot, where a recurrence is denoted as a black point and non-recurrence as a white point. In an FRP, the binary nature of recurrence is relaxed. Instead of just marking points as either recurrent or non-recurrent, it assigns a degree of membership to each point, representing the degree of similarity or recurrence. This introduces a level of fuzziness or uncertainty into the analysis. The degree of membership takes real values between 0

and 1, where 0 means no recurrence (completely dissimilar) and 1 means a perfect match (completely similar). Real values between 0 and 1 indicate the level of partial similarity. FRPs can be useful for analyzing time series data when the distinction between recurrence and non-recurrence is inherently not clear-cut. This uncertainty modeling allows for a more natural understanding of the data and can reveal hidden patterns or relationships that might not be apparent in traditional binary RPs.

Given an embedding dimension d and a time delay β, a phase-space construction of the original time series or sequence $(x_1, x_2, \ldots, x_M)$ yields $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, where $N = M - (d-1)\beta$. The constructed elements of $\mathbf{Y}$ can be expressed as follows:

$$\mathbf{y}_i = (x_i, x_{i+\beta}, \ldots, x_{i+(d-1)\beta}),\ i = 1, \ldots, N - (d-1)\beta. \tag{4}$$

Using this constructed phase space $\mathbf{Y}$, an FRP, denoted as $\mathbf{R}$, can computed and represented as a grayscale image to visualize the recurrence patterns of a dynamical system. More precisely, an FRP is a square matrix containing membership grades that quantify the similarity between pairs of points in the constructed phase-space trajectory of the dynamical system. This similarity is mathematically expressed as:

$$R(i,j) = \mu(\mathbf{y}_i, \mathbf{y}_j),\ i,j = 1, \ldots, N, \tag{5}$$

where $\mu(\mathbf{y}_i, \mathbf{y}_j) \in [0,1]$ represents the membership of similarity between $\mathbf{y}_i$ and $\mathbf{y}_j$. A higher value of $\mu(\mathbf{y}_i, \mathbf{y}_j)$ suggests a stronger similarity between $\mathbf{y}_i$ and $\mathbf{y}_j$.

The elements of an FRP are determined using three fundamental properties of fuzzy inference:

$$\mu(\mathbf{y}_i, \mathbf{y}_i) = 1,\ i = 1, \ldots, N. \tag{6}$$

which expresses reflexivity.

$$\mu(\mathbf{y}_i, \mathbf{v}_k) = \mu(\mathbf{v}_k, \mathbf{y}_i),\ i = 1, \ldots, N, k = 1, \ldots, c, \tag{7}$$

which defines symmetry, where $\mathbf{v}_k$ represents the k-th cluster center, and $c > 1$ is the specified number of clusters to which each element $\mathbf{y}_i$ belongs with an estimated membership level.

$$\mu(\mathbf{y}_i, \mathbf{y}_j) = \max \left[\min_k \{\mu(\mathbf{y}_i, \mathbf{v}_k), \mu(\mathbf{v}_k, \mathbf{y}_j)\} \right],\ i \neq j, \tag{8}$$

which infers transitivity.

The membership grades $\mu(\mathbf{y}_i, \mathbf{v}_k)$, $i = 1, \ldots, N$, $k = 1, \ldots, c$, can be optimally determined using the fuzzy c-means (FCM) algorithm [23]. The objective of the FCM algorithm is to minimize the following function:

$$\mathcal{O} = \sum_{i=1}^{N} \sum_{k=1}^{c} [\mu(\mathbf{y}_i, \mathbf{v}_k)]^m \, \|\mathbf{x}_i - \mathbf{v}_k\|^2, \tag{9}$$

where $m \in [1, \infty)$, typically set to 2, represents the fuzzy weighting exponent. The FCM objective function is subject to the following constraint:

$$\sum_{k=1}^{c} \mu(\mathbf{y}_i, \mathbf{v}_k) = 1,\ i = 1, \ldots, N. \tag{10}$$

The objective function can be numerically minimized through iterative steps, as follows:

Using initial values for $\mu(\mathbf{y}_i, \mathbf{v}_k)$, $k = 1, \ldots, c$, $i = 1, \ldots, N$, the cluster centers and membership grades are iteratively updated until convergence or a predefined number of iterations is reached:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{N} \mathbf{y}_i [\mu(\mathbf{y}_i, \mathbf{v}_k)]^m}{\sum_{i=1}^{N} [\mu(\mathbf{y}_i, \mathbf{v}_k)]^m}, \tag{11}$$

and

$$\mu(\mathbf{y}_i, \mathbf{v}_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\mathbf{y}_i - \mathbf{v}_k\|}{\|\mathbf{y}_i - \mathbf{v}_j\|} \right)^{2/(m-1)}}. \tag{12}$$

For the computation of the FRPs in this study, the embedding dimension d= 1 and time delay β = 1 were used for the phase-space construction, and number of clusters c = 3 used for the FCM algorithm.

The Shannon entropy of an FRP (FRP-SE), which is a grayscale image, denoted as $H(\mathbf{R})$, is expressed as

$$H(\mathbf{R}) = -\sum_{l=1}^{G} p_l \log_2 p_l, \tag{13}$$

where G is set to 256, representing the number of gray levels in $\mathbf{R}$. This value is derived by converting real pixel values from the [0, 1] range into integers within [0, 255], p_l corresponds to the probability assigned to the intensity level l, which is calculated based on the normalized histogram for the l-th bin.

Drawing from the concept of non-probabilistic entropy as defined in [24], the fuzzy recurrence entropy of an $N \times N$ FRP (FRP-FE), denoted as $\hat{H}(\mathbf{R})$, which quantifies the level of uncertainty in recurrences within the constructed phase space of a dynamical system, is defined as [25]

$$\hat{H}(\mathbf{R}) = \sum_{i=}^{N} \sum_{j=1}^{N} -\mu(\mathbf{y}_i, \mathbf{y}_j) \log_2 \mu(\mathbf{y}_i, \mathbf{y}_j) - [1 - \mu(\mathbf{y}_i, \mathbf{y}_j)] \log_2 [1 - \mu(\mathbf{y}_i, \mathbf{y}_j)]. \tag{14}$$

3.3 Extraction of wavelet time scattering features

A network for performing wavelet time scattering decomposition employing the analytic Morlet wavelet [26, 27] can be constructed as described in [28]. This network leverages

wavelets and a lowpass scaling function to produce low-variance representations of real-valued time series data. The wavelet time scattering process results in representations that remain robust to translations within the input signal while retaining their ability to discriminate between different classes. These representations can be used as inputs to a classifier for pattern classification. The iterative computation of wavelet scattering coefficients across multiple layers is elucidated as follows.

Consider $\psi(t)$ as a band-pass filter, often referred to as the mother wavelet, which adopts the Morlet wavelet. Additionally, let $\psi_{\omega_u}(t)$ represent a wavelet filter bank, which can be constructed by dilating the mother wavelet as follows:

$$\psi_{\omega_u}(t) = \omega_u \psi(\omega_u t), \tag{15}$$

in which $\omega_u = 2^{(u/V)}$, $u \in \mathbb{Z}$, with $1 \leq u \leq U$ that is the maximum level of layers, and V representing the number of wavelets per octave.

Consider $\mathbf{s}$ as the input signal. The zeroth-order wavelet scattering coefficients are computed by calculating the average of the feature vector, expressed as

$$L_0(t) = \mathbf{s} * \phi(t), \tag{16}$$

where L_0 represents the zeroth-order scattering, ϕ stands for a low-pass filter, and the $*$ symbol denotes the convolution operator.

The coefficients of the first-order wavelet scattering, which belong to layer 1, are calculated by taking the average of the absolute values of the wavelet coefficients at this layer, as follows:

$$L_1(t, \omega_1) = |\mathbf{s} * \psi_{\omega_1}(t)| * \phi(t). \tag{17}$$

The second-order wavelet scattering coefficients are computed as follows:

$$L_2(t, \omega_1, \omega_2) = ||\mathbf{s} * \psi_{\omega_1}(t)| * \psi_{\omega_2}(t)| * \phi(t). \tag{18}$$

Similarly, the computation of the third-order wavelet scattering coefficients is expressed as

$$L_3(t, \omega_1, \omega_2, \omega_3) = |||\mathbf{s} * \psi_{\omega_1}(t)| * \psi_{\omega_2}(t)| * \psi_{\omega_3}(t)| * \phi(t). \tag{19}$$

In a general sense, wavelet scattering coefficients at layers u, where $u = 1, \ldots, U$, are established by applying a combination of convolution, modulus, and average pooling operators as follows:

$$L_j(t, \omega_1, \ldots, \omega_j) = |\ldots||\mathbf{s} * \psi_{\omega_1}(t)| * \psi_{\omega_2}(t)| \cdots * \psi_{\omega_j}(t)| * \phi(t). \tag{20}$$

The mother wavelet is defined as [26, 27]

$$\psi(t) = c \left[e^{-\frac{t^2}{2\sigma^2}} \right] e^{2\pi i f t}, \tag{21}$$

where, in this study, $c = 1$, σ is the wavelet duration set to 1, i is the imaginary unit, f is the central frequency, and $2\pi f$ is set to 5. Consequently, $e^{2\pi i f t} = \cos(5t)$, resulting in $\psi(t) = e^{-\frac{t^2}{2}} \cos(5t)$.

Other parameters for the wavelet time scattering in this study were specified as follows. Scale of time invariance = half of signal length, using wavelet scattering network with two filter banks, where V factor for filter bank 1 = 8 wavelets per octave, and V factor for filter bank 2 = 1 wavelet per octave, and sampling frequency = 1 Hz.

3.4 Classification with long short-term memory networks

Figure 1 illustrates the progression of an input time series as it passes through an LSTM layer [29]. When utilizing TF and TS features, the input at a given time point is a fusion

of four distinct features: IF, SE, FRP-SE, and FRP-FE, all of which are extracted from the corresponding time point segment. Learnable parameters of an LSTM layer encompass the input weights, denoted as **a**, the recurrent weights, denoted as **r**, and the bias, denoted as b. These parameters are organized into matrices and vectors as follows: **A** represents the concatenation of input weights, **R** embodies the concatenation of recurrent weights, and vector **b** encapsulates the concatenation of biases from each component. These concatenations are mathematically expressed as:

$$\mathbf{A} = [\mathbf{a}_i, \mathbf{a}_n, \mathbf{a}_g, \mathbf{a}_o]^T, \tag{22}$$

$$\mathbf{R} = [\mathbf{r}_i, \mathbf{r}_n, \mathbf{r}_g, \mathbf{r}_o]^T, \tag{23}$$

$$\mathbf{b} = [b_i, b_n, b_g, b_o]^T, \tag{24}$$

where the subscripts i, n, g, and o respectively denote the input gate, neglect (or forget) gate, cell candidate, and output gate.

The cell state at time step t is defined as

$$\mathbf{c}_t = n_t \circ \mathbf{c}_{t-1} + i_t \circ g_t, \tag{25}$$

in which $\circ$ is the Hadamard product.

The hidden state at time step t is given by

$$\mathbf{f}_t = o_t \circ \sigma_c(\mathbf{c}_t), \tag{26}$$

where σ_c represents the state activation function, which is typically computed using the hyperbolic tangent function.

At time step t, the input gate (i_t), neglect gate (n_t), cell candidate (g_t), and output gate (o_t) are formulated as follows:

$$i_t = \sigma_g(\mathbf{a}_i \mathbf{u}_t + \mathbf{r}_i \mathbf{n}_{t-1} + b_i), \tag{27}$$

$$n_t = \sigma_g(\mathbf{a}_n \mathbf{u}_t + \mathbf{r}_n \mathbf{f}_{t-1} + b_n), \tag{28}$$

$$g_t = \sigma_c(\mathbf{a}_g \mathbf{u}_t + \mathbf{r}_g \mathbf{f}_{t-1} + b_g), \tag{29}$$

$$o_t = \sigma_g(\mathbf{a}_o \mathbf{u}_t + \mathbf{r}_o \mathbf{f}_{t-1} + b_o), \tag{30}$$

where σ_g represents the gate activation function, commonly employing the sigmoid function.

Moreover, a bidirectional LSTM (bi-LSTM) [30] represents an extension of the LSTM, aiming to offer enhanced performance for sequence classification tasks. Unlike the LSTM, which is trained using a single sequence, the bi-LSTM architecture is trained using two simultaneous LSTM layers: one processes the input time series in its original order, and the other operates on a reversed version of the time series. This dual-layered architecture enables the model to capture bidirectional long-term dependencies between different time steps within the time series data, thereby providing additional contextual information to the network. Consequently, this approach is expected to facilitates more comprehensive learning from the input data.

In this study, the configuration of LSTM and biLSTM networks were chosen as follows. Number of hidden units = 100, output mode = last time step of the sequence, cell and state activation function = hyperbolic tangent, gate activation function = sigmoid. Training options for the networks were set as follow. Solver for training neural network =

Adam optimizer, bias learning rate = 1, input-weight L_2 regularizer = nonnegative scalar, recurrent-weight L_2 regularizer = 1, and bias L_2 regularizer = 0.

3.5 Measures of classification performance

To compute statistical measures of performance for classifiers, the following variables are defined as

- P = the number of samples of pathological voice
- N = the number of samples of healthy voice
- TP = the numer of correctly predicted samples as pathological voice
- TN = the number of correctly predicted samples as healthy voice
- FP = the number of falsely predicted samples as pathological voice
- FN = the number of falsely predicted samples as healthy voice

The measures of performance in the context of accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), and F_1 score are defined Table 1.

Another performance metric to consider is the area under the receiver operating characteristic (ROC) curve. The ROC curve is generated by plotting the true positive (TP) rate against the false positive (FP) rate across different thresholds. In some contexts, the TP rate is referred to as sensitivity or the probability of correct prediction, while the FP rate is synonymous with the probability of false alarm. Therefore, the area under the ROC curve (AUC) serves as an assessment of predictor quality, irrespective of the specific classification threshold chosen. Larger values of these performance measures indicate superior performance of the predictor.

4 Results

It was shown that the LSTM processing of long signals results in undesired computational complexity in association with the training phase [31]. The original speech signals were divided into short segments of L = 250 and 500 time points, where segments whose lengths were less than the specified length were discarded. To overcome the data imbalance, where samples of the pathological conditions are twice more than those of the healthy, partial samples of the healthy speech were replicated to match the sample size of the pathology. Thus, for L = 250 and 500, there are 21,746 and 10,861 short speech signals created for each cohort, respectively. Figures 2 and 3 show examples of healthy and pathological speech signals (L = 250) and their extracted features, respectively.

The dataset was randomly split into 10 folds, where 9 folds were used for training the networks and the remaining fold was used for testing the performance of the trained networks. Tables 2 and 3 show results of classifying healthy and pathological voices obtained from different LSTM-based classifiers in terms of accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), F_1 score, and area under the ROC curve (AUC) with respect to L = 500 and 250, respectively.

For speech segments with L = 500, the LSTM, biLSTM, Wavelet-LSTM (LSTM inputs are wavelet time scattering coefficients), and Wavelet-LSTM obtained 50% accuracy and AUC = 0.5. The bi-LSTM, wavelet-LSTM, and wavelet-biLSTM completely biased toward the healthy class (SPE = 100%), whereas the LSTM completely biased toward the pathological class (SEN = 100%). Both TFTS-LSTM (LSTM inputs are time-frequency and time-space features) and TFTF-biLSTM provided much better results (AUC = 0.93) than the other classification models. Accuracy rates obtained from the TFTS-LSTM and TFTS-biLSTM are 85% and 86%, respectively. Both TFTS-LSTM and TFTS-biLSTM could achieve a balanced classification between the healthy and pathological samples (SEN = 92% and 89%, SPE =

73% and 83%, respectively). The precision measure obtained from the TFTS-biLSTM (84%) is higher than the precision from the TFTS-LSTM (81%), whereas F_1 scores obtained from the two models are the same (0.86).

For shorter speech segments with L = 250, complete biases that are opposite to the case with L = 500 observed among LSTM, biLSTM, wavelet-LSTM, and wavelet-biLSTM models. The AUC values for LSTM, biLSTM, wavelet-LSTM, and wavelet-biLSTM are between 0.50 and 0.53; whereas the AUC for TFTS-LSTM and TFTS-biLSTM = 0.96 and 0.95, respectively. The use of the shorter signals could improve the performance of both TFTS-LSTM and TFTS-biLSTM models, whose accuracy rates (88% and 90% for TFTS-biLSTM and TFTS-LSTM, respectively) are significantly higher than those of the other four models (50%). Once again, both TFTS-LSTM and TFTS-LSTM could diagnose the pathological conditions better than the healthy samples (SEN = 93% vs. SEN = 87% for TFTS-LSTM, and SEN = 94% vs. SEN = 82% for TFTS-biLSTM). The precision and F_1 score of the TFTS-LSTM are higher than those of the TFTS-biLSTM.

Based on the results shown in Tables 2 and 3, the TFTS-LSTM for learning the speech signals of L = 250 provided the most favorable classification performance for differentiating between the healthy and pathological voices. To provide a visual comparison between the three different classification methods, Figure 4 shows the iterative machine learning for differentiating the two classes from the speech segments of L = 250 and confusion matrices using LSTM, WS-LSTM, and TFTS-LSTM models. It can be seen that both LSTM and WS-LSTM could not improve the training over 50 epochs, whereas the TFTS-LSTM could reach toward nearly the maximum training accuracy or minimum loss.

In comparison with using the speech signals of original lengths reported in a previous study [16], in this study, both LSTM and biLSTM increased the classification accuracy to about 2%; wavelet-LSTM and wavelet-biLSTM resulted in about 20% decrease in the classi-

fication accuracy; and TFTS-LSTM and TFTS-biLSTM increased the classification accuracy to about 14% and 12%, respectively. A recent study, which incorporated deep-time recurrence features [32] into the LSTM for classifying healthy and pathological speech signals using the same data set, obtained ten-fold results with accuracy = 86%, sensitivity = 100%, and specificity = 50%, showing an unbalanced differentiation between the two cohorts. In these studies, TFTS-LSTM and TFTS-biLSTM outperformed all other methods in terms of accuracy and balance in the diagnosis.

Furthermore, in terms of computational complexity, the use of both LSTM and biLSTM required significant longer training times than the Wavelet-LSTM, Wavelet-biLSTM, TFTS-LSTM, and TFTS-biLSTM. By training these models with a single processor (Intel(R) Core(TM) i7-6500U, CPU@2.50 GHz), it was noticed that the time taken for training either the LSTM or biLSTM with the original speech segments were 25 times longer than the time for training the wavelet-LSTM or wavelet-biLSTM, and 10 times longer than the training time for the TFTS-LSTM or TFTS-biLSTM.

5 Conclusion

The application of time-frequency and time-space features for LSTM learning was shown to be of high and better performance in classifying physiological signals than several other classification models [19]. In this study, the experimental results illustrated the superior performance of the LSTM that learned on time-frequency and time-space features of healthy and pathological speech signals to those of LSTM networks that learned on either the original data or wavelet-scattering coefficients of the signals.

Certain advantages offered with the incorporation time-frequency and time-space features as input into an LSTM network include better diagnosis accuracy and computational efficiency. A strategy for creating balanced data and augmenting training samples was

found to be effective by means of the results obtained from the TFTS-LSTM and TFTS-biLSTM classifiers. In this study, an empirical basis for the selection of short lengths of the speech signals was relied on. An analytical procedure for choosing an optimal signal length is worth developing in a future investigation. Furthermore, considering the inclusion of time-frequency and time-space features in other classifiers would be encouraged for potential improvement of the diagnosis.

Medical voice analysis systems utilize hardware, software, and human-computer interaction to achieve smart hospital facilities [33]. Technical elaborations on this study can contribute to endeavors concerning intelligent technology for the diagnosis of pathology in human acoustics and its potential applications in smart healthcare.

Data and Computer Code Availability

The data that were used in this study are publicly available from the PhysioNet website [18]. The Matlab codes implemented in this study are freely available at the first author's personal website: `https://sites.google.com/view/tuan-d-pham/codes`, under the title "TFTS-LSTM for pathological voice diagnosis".

Declaration of Interest Statement

The authors declare that they have no conflicts of interest.

Funding

Not applicable.

References

[1] Speech-language pathology's role in head and neck cancer, voice and oral-maxillofacial surgery, *Michigan Medicine*, May 31, 2023, https://medicine.umich.edu/dept/otolaryngology/news/archive/202305/speech-language-pathology%E2%80%99s-role-head-neck-cancer-voice-oral-maxillofacial-surgery.

[2] LeBlanc E.M., Shprintzen R.J., Speech and the maxillofacial complex: A structural-functional perspective for diagnosis and management, *Oral and Maxillofacial Surgery Clinics of North America*, 6 (1994) 113-120.

[3] Pereira T.C., Brasolotto A.G., Conti P.C., Berretin-Felix G., Temporomandibular disorders, voice and oral quality of life in women, *J Appl Oral Sci.*, 17 Suppl (2009) 50-56.

[4] Lathrop-Marshall H., Keyser M.M.B., Jhingree S., Giduz N., Bocklage C., Couldwell S., et al., Orthognathic speech pathology: Impacts of Class III malocclusion on speech, *Eur J Orthod.*, 44 (2022) 340-351.

[5] Pinsky T.M., Goldberg H.J. Potential for clinical cooperation between dentistry and speech pathology, *Int Dent J.*, 27 (1977) 363-369.

[6] Hassan T., Naini F.B., Gill D.S., The effects of orthognathic surgery on speech: A review, *J Oral Maxillofac Surg.*, 65 (2007) 2536-2543.

[7] Fawcus R., Dental problems in speech pathology, *Proc R Soc Med.*, 61 (1968) 619-622.

[8] Gugsch C., Dannhauer K.H., Fuchs M., Evaluation of the progress of therapy in patients with cleft lip, jaw and palate, using voice analysis–A pilot study, *J Orofac Orthop.*, 69 (2008) 257-267.

[9] Mogren A., Havner C., Westerlund A., Sjogreen L., Agholme M.B., Mcallister A., Malocclusion in children with speech sound disorders and motor speech involve-

ment: A cross-sectional clinical study in Swedish children, *Eur Arch Paediatr Dent.*, 23 (2022):619-628.

[10] Bommangoudar J.S., Chandrashekhar S., Shetty S., Sidral S., Pedodontist's role in managing speech impairments due to structural imperfections and oral habits: A literature review, *Int J Clin Pediatr Dent.*, 13 (2020) 85-90.

[11] Compton E.C., Cruz T., Andreassen M., Beveridge S., Bosch D., Randall D.R., Livingstone D., Developing an artificial intelligence tool to predict vocal cord pathology in primary care settings, *Laryngoscope*, 133 (2023) 1952-1960.

[12] Idrisoglu A., Dallora A.L., Anderberg P., Berglund J.S., Applied machine learning techniques to diagnose voice-affecting conditions and disorders: Systematic literature review, *J Med Internet Res*, 25 (2023) e46105.

[13] Peng X., Xu H., Liu J., Wang J., He C., Voice disorder classification using convolutional neural network based on deep transfer learning, *Sci Rep*, 13 (2023) 7264.

[14] Verde L., Brancati N., De Pietro G., Frucci M., Sannino G., A deep learning approach for voice disorder detection for smart connected living environments, *ACM Transactions on Internet Technology*, 22 (2021) 1-16.

[15] Zakariah M., Reshma B., Alotaibi Y.A., Guo Y., Tran-Trung K., Elahi M.M., An analytical study of speech pathology detection based on MFCC and deep neural networks, *Comput Math Methods Med.*, 2022 (2022) 7814952.

[16] Pham T.D., Efficient deep learning for pathological speech recognition, *2023 IEEE Conference on Artificial Intelligence (CAI)*, Santa Clara, CA, USA, 2023, pp. 103-104, doi: 10.1109/CAI54212.2023.00052.

[17] Cesari U., De Pietro G., Marciano E., Niri C., Sannino G., Verde L., A new database of healthy and pathological voices, *Computers & Electrical Engineering*, 68 (2018) 310-321.

[18] VOICED (VOice ICar fEDerico II) database, `https://archive.physionet.org/physiobank/database/voiced/`, accessed 03 March 2023.

[19] Pham T.D., Time-frequency time-space LSTM for robust classification of physiological signals, *Scientific Reports*, 11 (2021) 6936.

[20] Boashash B., Khan N.A., Ben-Jabeur T., Time-frequency features for pattern recognition using high-resolution TFDs: A tutorial review. *Digital Signal Processing*, 40 (2015) 1-30.

[21] Pham T.D., Fuzzy recurrence plots, *EPL*, 116 (2016) 50008.

[22] Eckmann J.P., Kamphorst S.O., Ruelle D., Recurrence plots of dynamical systems, *Europhysics Letters*, 5 (1987) 973-977.

[23] Bezdek J.C., Ehrlich R., Full W., FCM: The fuzzy *c*-means clustering algorithm, *Computers & Geosciences*, 10 (1984) 191-203.

[24] de Luca A., Termini S., A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, *Information and Control*, 20 (1972) 301-312.

[25] Pham T.D., Fuzzy recurrence entropy, *EPL* 130 (2020) 40004.

[26] Morlet J., Arens G., Fourgeau E., Giard D., Wave propagation and sampling theory–Part I: Complex signal and scattering in multilayered media, *Geophysics*, 47 (1982) 203-221.

[27] Morlet J., Arens G., Fourgeau E., Giard D., Wave propagation and sampling theory–Part II: Sampling theory and complex waves, *Geophysics*, 47 (1982) 222-236.

[28] Mallat S, Group invariant scattering, *Communications in Pure and Applied Mathematics*, 65 (2012) 1331-1398.

[29] Yu Y., Si X., Hu C., Zhang J., A review of recurrent neural networks: LSTM cells and network architectures, *Neural Computation*, 31 (2019) 1235-1270.

[30] Schuster M., Paliwal K.K., Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing*, 45 (1997) 2673-2681.

[31] Pham T.D., Wardell K., Eklund A., Salerud G., Classification of short time series in early Parkinson's disease with deep learning of fuzzy recurrence plots, *IEEE/CAA Journal of Automatica Sinica*, 6 (2019) 1306-1317.

[32] Pham T.D., Deep time-recurrence features, *EPL*, 142 (2023) 51001.

[33] Zhang J., Wu ., Qiu Y., Song A., Li W., Li X., Liu Y., Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review, *Computers in Biology and Medicine*, 153 (2023) 106517.

Table 1: Statistical performance measures of classification.

ACC	SEN	SPE	PRE	F_1 score
$\frac{TP+TN}{P+N}$	$\frac{TP}{P}$	$\frac{TN}{N}$	$\frac{TP}{TP+FP}$	$\frac{2TP}{2TP+FP+FN}$

Table 2: Diagnosis of healthy and pathological speech signals of length = 500 time points.

Classifier	$\%ACC$	$\%SEN$	$\%SPE$	$\%PRE$	F_1	AUC
biLSTM	50.00	0.00	100	NaN	0	0.52
LSTM	50.00	100	0.00	50.00	0.67	0.52
Wavelet-biLSTM	50.00	0.00	100	NaN	0	0.55
Wavelet-LSTM	50.00	0.00	100	NaN	0	0.54
TFTS-biLSTM	85.96	89.13	82.78	83.81	0.86	0.93
TFTS-LSTM	85.04	92.36	77.72	80.56	0.86	0.93

Table 3: Diagnosis of healthy and pathological speech signals of length = 250 time points.

Classifier	$\%ACC$	$\%SEN$	$\%SPE$	$\%PRE$	F_1	AUC
biLSTM	50.00	100	0.00	50.00	0.67	0.50
LSTM	50.00	0.00	100	NaN	0	0.50
Wavelet-biLSTM	50.00	100	0.00	50.00	0.67	0.53
Wavelet-LSTM	50.00	100	0.00	50.00	0.67	0.51
TFTS-biLSTM	88.02	93.75	82.30	84.12	0.89	0.95
TFTS-LSTM	90.14	93.38	86.90	87.69	0.90	0.96

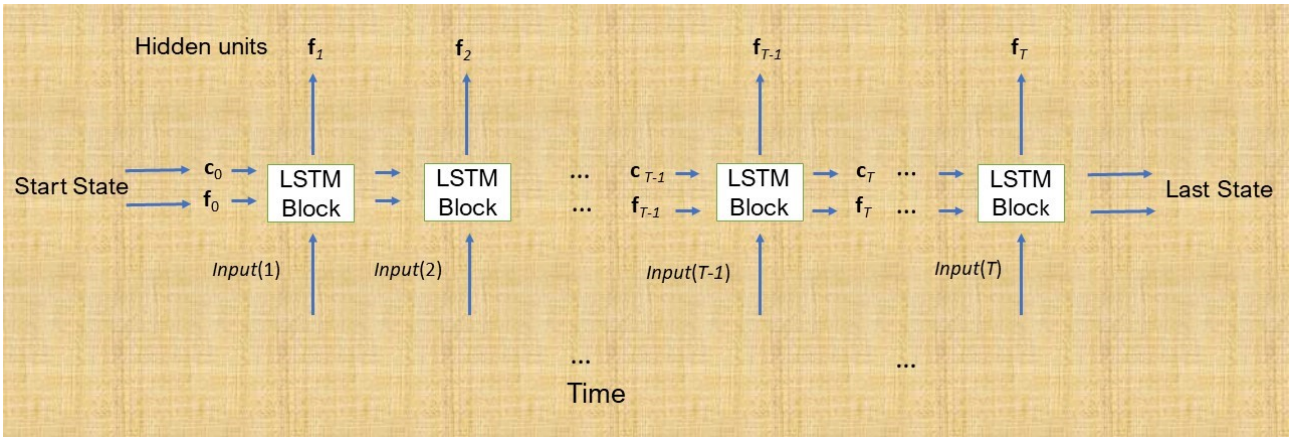
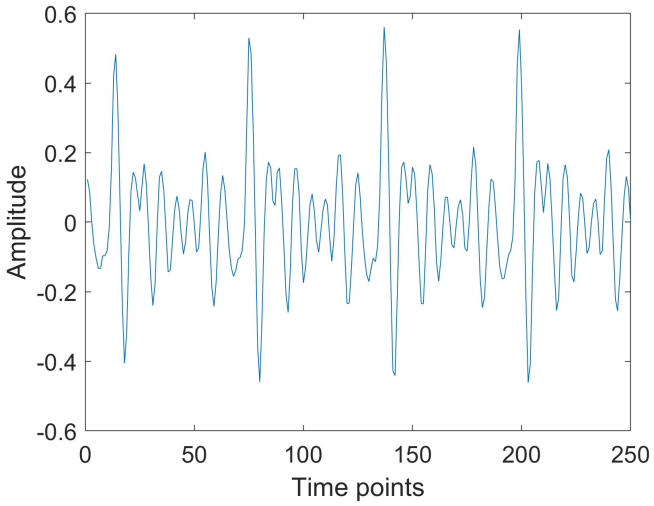
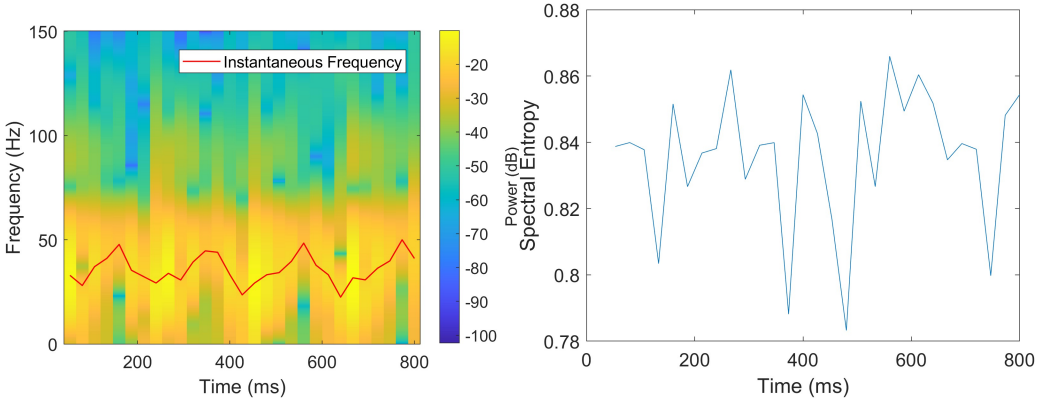


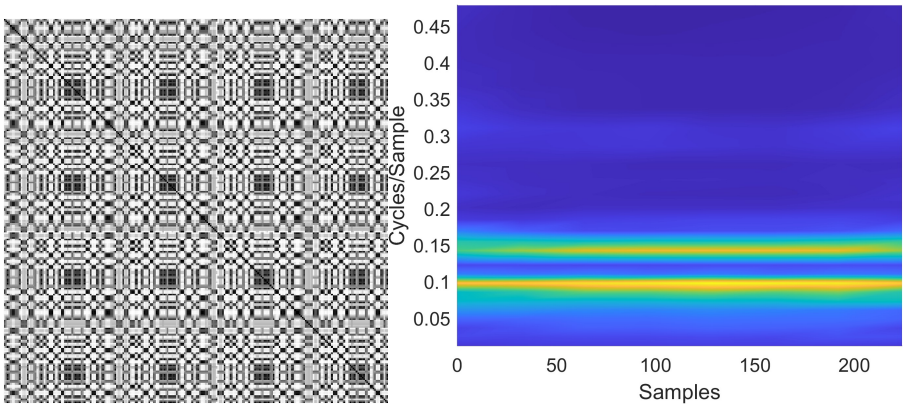
Figure 1: Architecture of a long short term memory (LSTM) network.



(a)

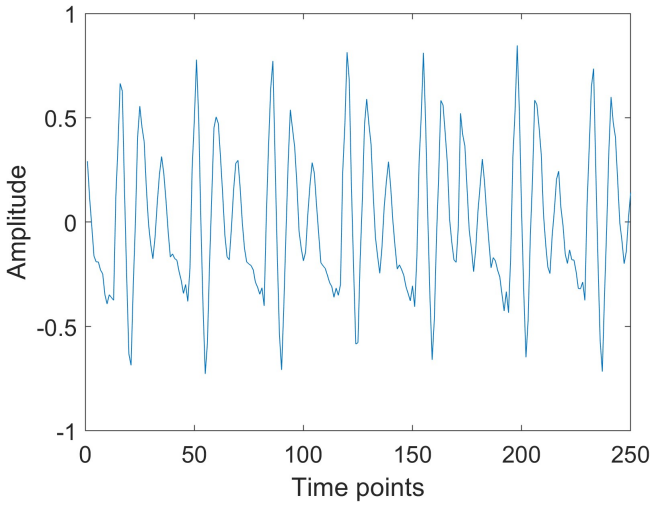


(b) (c)

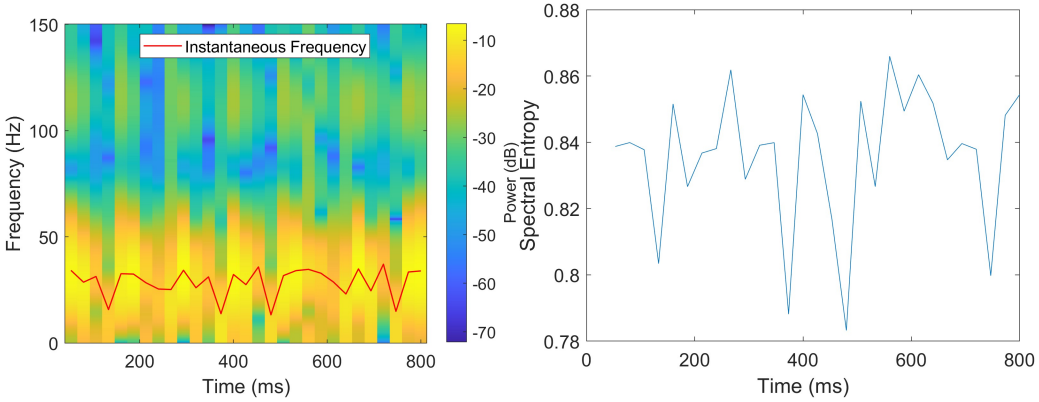


(d) (e)

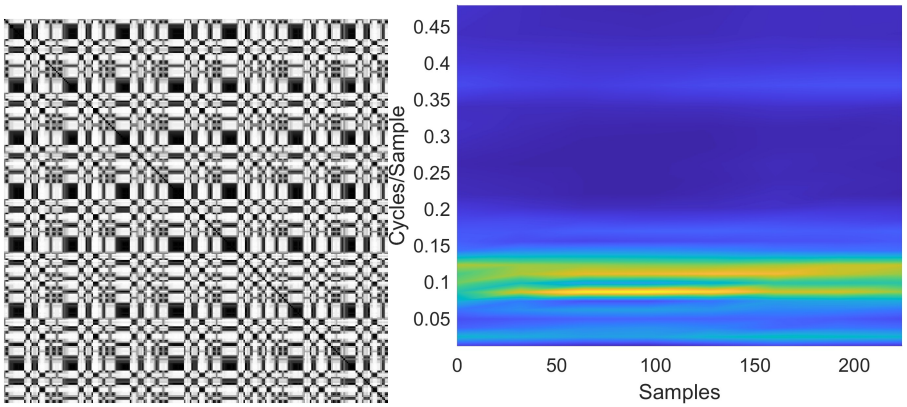
Figure 2: Healthy speech and features: (a) signal segment, (b) instantaneous frequency, (c) spectral entropy, (d) fuzzy recurrence plot, and (e) scattergram of first-order scattering coefficients.



(a)

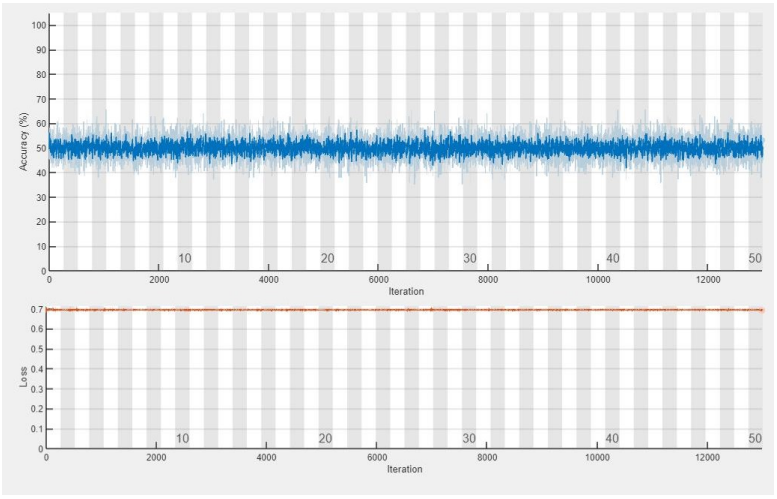


(b) (c)

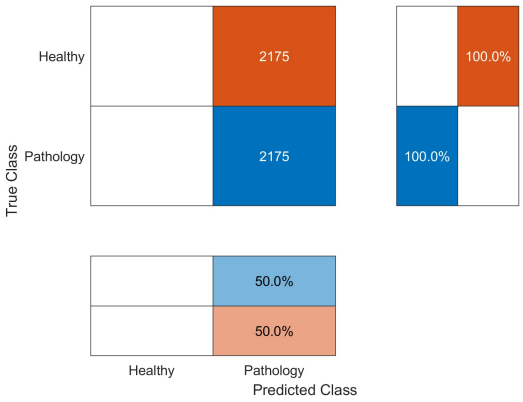


(d) (e)

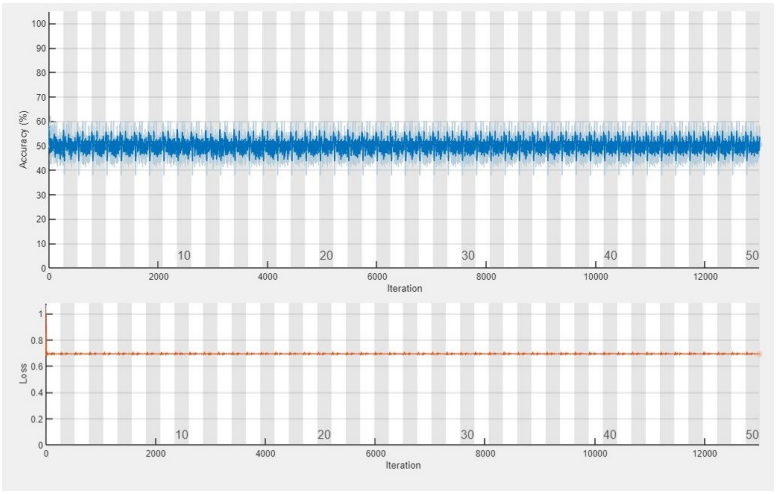
Figure 3: Pathological speech and features: (a) signal segment, (b) instantaneous frequency, (c) spectral entropy, (d) fuzzy recurrence plot, and (e) scattergram of first-order scattering coefficients.



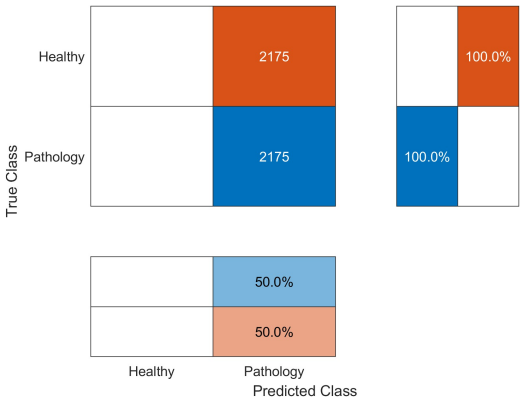
(a) LSTM



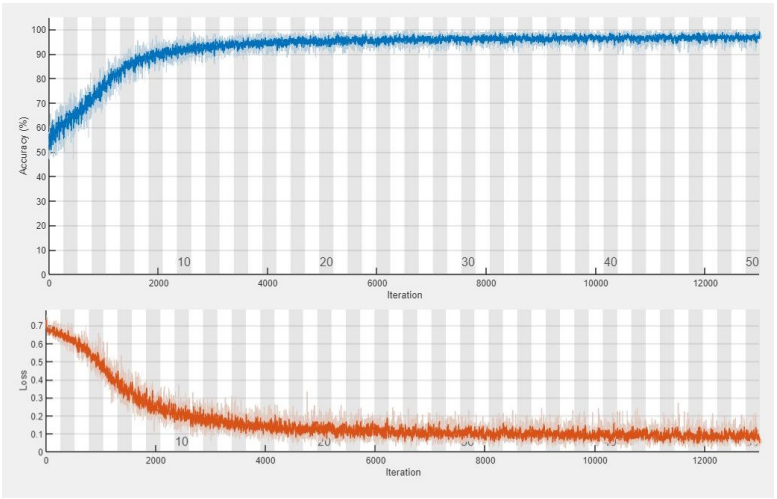
(b) LSTM



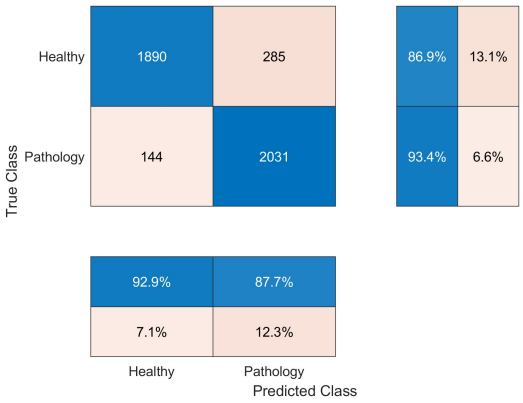
(c) Wavelet-LSTM



(d) Wavelet-LSTM



(e) TFTS-LSTM



(f) TFTS-LSTM

Figure 4: Training processes and confusion matrices using different classification models.