Received May 31, 2021, accepted June 15, 2021, date of publication June 17, 2021, date of current version June 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3090317

Convergence of Artificial Intelligence and Internet of Things in Smart Healthcare: A Case Study of Voice Pathology Detection

GHULAM MUHAMMAD, (Senior Member, IEEE), AND MUSAED ALHUSSEIN, (Member, IEEE)
Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Ghulam Muhammad (ghulam@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia, under Grant RG-1436-016.

ABSTRACT The integration of artificial intelligence (AI) and the Internet of Things (IoT) has tremendous prospects in smart healthcare. The advancement of AI in the form of deep learning brought a revolution in automatic classification and detection systems. In addition, next-generation wireless communications such as 5G networking brought speed and the seamless transmission of data. With the convergence of these elements, the smart healthcare sector is currently booming. Particularly during the post-COVID-19 pandemic, the necessity of smart healthcare has come to light more than before. A significant number of people suffer from voice pathology. This pathology can be easily cured if detected early. In this study, a voice pathology detection system within a smart healthcare framework is proposed. The inputs are obtained by the IoT, namely microphones and electroglottography (EGG) devices to capture voice and EGG signals, respectively. Spectrograms are obtained from these signals and fed into a pretrained convolutional neural network (CNN). The features extracted from the CNN are fused and processed using a bi-directional long short-term memory network. The proposed system is evaluated using a publicly available database, called the Saarbruecken voice database. The experimental results show that bimodal input performs better than a single input. An accuracy of 95.65% is obtained for the proposed system.

INDEX TERMS Smart healthcare, deep learning, convolutional neural network (CNN), long short-term memory (LSTM), voice pathology detection.

I. INTRODUCTION

Owing to the excessive use of their voice, numerous individuals today suffer from voice pathologies. Specifically, teachers, students, musicians, attorneys, and the like are among those who commonly experience these problems [1]. Human voices provide a vast amount of informative material; as a result, they convey much about human wellbeing. This knowledge is applicable in the fields of automated speech pathology diagnosis and speaker identification, as well as other areas. Thus, the area of speech pathology draws a large number of scholars who are interested in investigating and studying voice disorders. The irregular growth of masses or tissue in the vocal folds results in a voice tone that differs from the norm [2]. Examples of abnormal growths in vocal folds resulting in voice pathologies include polyps, nodules, cysts, and sulci [3]. Most speech pathology signs include persistent hoarseness, scratchy throat, abnormal volume, and reduced capacity to speak clearly.

The methods used to investigate speech or voice problems may be analytical or empirical. The assessment of a disordered expression, such as "dysphonia," which is a medical term for voice disorders that prevent an individual from making a sound using the vocal organs, may be referred to as voice disorders. Perceptual and auditory evaluation methods are critical to the therapeutic treatment of dysphonia, as is the endoscopic evaluation of the larynx and vocal folds.

Although the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) and GRBAS (grade, roughness, breathiness, asthenia, strain) scales use rating scales to evaluate

The associate editor coordinating the review of this manuscript and approving it for publication was Diana Patricia Tobon.

speech (including frequency, quality, roughness, breathiness, exhaustion, and stress), the average grade of dysphonia, degree of roughness, breathiness, asthenia, and strain is assessed using the CAPE-V [4]. Because these assessment methods are often used in clinical practice, however, there are caveats for the analytical test. The clinician's expertise, the degree of the patient's dysphonia, the form of the auditory perceptual rating system, and the stimulation or speech task are all factors in the limits imposed on a given treatment. As a result of failure in auditory appraisal, physicians and researchers have designed a more quantitative voice recognition metric to measure dysphonia levels in patients [5]. Numerical values for pathology severity, along with a treatment plan, can be created through acoustic examination, and this knowledge can be made accessible to other stakeholders. Often, patients' acoustic analyses are performed by voice clinicians using sustained vowel recordings rather than constant speech samples [6]. While it has been established that the continuous vowel is the best method for producing a voice sample that captures various rhythms of expression, this does not reflect how people talk every day [7]. Just as voice onset, voice cessation, and voice breaks are essential in the assessment of voice consistency, variations in vocal characteristics with respect to these three phenomena are not entirely captured in continuous signals, such as vowels. The third thing to note is that dysphonia signs are more apparent in dialogue voice development than in sustained vowels. In comparison, adductor spasmodic dysphonia is distinct from natural voice due to differences in the duration of sound output. Additionally, adductor spasmodic dysphonia is due to the effect of the phonetic and supra-segmental structure of speech that is unrepresented in the continuous vowel that some of the acoustic correlates of an individual's voice are formed [8].

To successfully identify and treat voice issues, it is absolutely imperative to be able to detect certain problems. To help physicians, an automatic voice pathology detection (VPD) system can be used. The VPD system only works for sustained vowels where the speech signal remains constant for about 6–9 s However, in day-to-day conversations, people should not use prolonged speaking, but continuous expression instead. Conceptually, a practical VPD system must be able to detect pathology from continuous sentences, which would mean that a practical VPD system would be able to detect pathology from continuous sentences [9]. According to [9], a VPD's speech recognition technology uses continuous speech, however, it is far from ideal. There are some basic reasons for this. The main reason is that we are unable to obtain proper voice impairment features, so many of the features come from speech processing and speaker recognition. There are many forms of voice dysfunction, each of which is defined as a separate class. In short, it is difficult to obtain information from human voice signals, and having reliable, effective, and scalable features with discriminative capacity is of primary importance.

In voice disorders, the quality, volume, or pitch of the sound made by the larynx is irregular [10]. Voice disabilities may result from several conditions, including emotional problems, traumatic experiences, physical illnesses, and diseases. By and large, speech disorders are not life-threatening and are easy to correct. Vocal abuse is one of the most frequent forms of speech impairment. If we experience vocal trauma, we can look for long-term vocal symptoms such as nodules, polyps, cysts, and edema (swelling) of the vocal folds [11]. Parkinson's disease, endocrinological (hormonal) abnormalities, and surgical procedures such as thyroidectomy or cardiac bypass both may result in speech disorders.

Based on the above discussion, we understand that there is a need to assess voice pathology at its earliest occurrence. Until now, many VPD systems have been proposed in the literature. A recent trend is to incorporate VPD systems in smart healthcare [12]. Smart healthcare takes advantage of developments in artificial intelligence (AI), machine learning (ML), deep learning, edge and cloud computing, and next-generation wireless communications. Various pathology detection systems have been embedded in smart healthcare [13], [14].

Smart healthcare frameworks are becoming popular because they bring comfort and ease to our lives. A person can get his disease diagnosed while remaining at home, can get advice from multiple physicians across the world, and can avoid the hassle of obtaining an appropriate appointment at the hospital. The precision and reduced latency of smart healthcare are made possible by integrating the Internet of Things (IoT), edge and cloud computing, and 5G networks. Furthermore, sophisticated ML algorithms such as deep learning algorithms have increased the accuracy of healthcare systems [15].

In this paper, we develop a VPD system within a smart healthcare framework. The system involves two modalities: voice signals and electroglottograph (EGG) signals. We design the system using a convolutional neural network (CNN) model to extract features from these two modalities. The features are fused using a long short-term memory (LSTM) model. The Saarbruecken voice database (SVD) is used in the experiments [16].

As the primary contributions of this research, we list the following:

(i) A multi-modal VPD system, which utilizes voice and EGG signals, and is therefore able to diagnose patients with dysarthria with increased accuracy and provide a better foundation for pathological voice identification. This system helps unite the two modalities and provides confidence in the outcomes.

(ii) Several pre-trained CNN models such as ResNet50, MobileNet v2, and XceptionNet are investigated as the backbone of the VPD system to minimize training time and provide more accurate results.

(iii) Using an LSTM network, features that were merged before were used to enrich the sound features, filter out redundant information, and increase classification accuracy.

The rest of this paper is organized as follows. Section II describes the types of voice pathologies. Section III provides

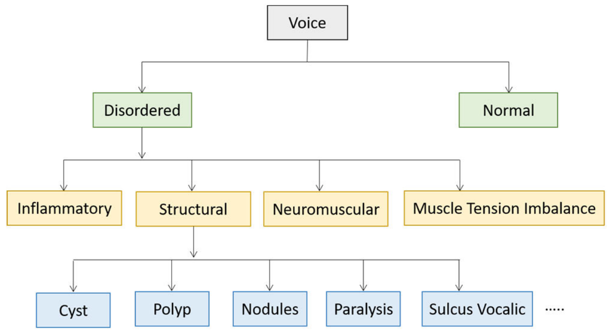


FIGURE 1. Voice pathology classification.

a brief literature review. Section IV presents the proposed VPD system. Section V presents experimental results and discussion. Finally, Section VI presents the conclusion and gives some directions for future work.

II. TYPES OF VOICE PATHOLOGIES

There are several types of voice pathologies. As mentioned earlier, an abnormal growth in the vocal fold(s) causes a voice pathology. Fig. 1 shows a voice classification strategy. The voice is mainly classified as either normal or disordered (pathological). A disorder can be caused by inflammation, abnormality in structural or neuromuscular design, or muscle tension imbalance. The structural abnormality can be generated by vocal fold(s) cysts, polyps, nodules, paralysis, and sulcus vocalis (some example images are given in Fig. 2). The focal point of this research is structural voice pathology.

A. CYST

A cyst is a development that can be found under the vocal fold mucosa's surface layer. If a void forms between the two vocal folds, the vocal folds cannot vibrate normally. Additionally, a cyst can stiffen the vocal fold mucosa, which may render the folds unable to vibrate normally. A cyst can also influence voice strength and/or affect the development of the voice.

On rare occasions, cysts will appear on only one vocal fold. The voice's tone can include natural speech, breathy speech, and harsh, raspy speech. Many of these patients suffer from cysts, and their complaints include fatigue after a long conversation.

B. VOCAL FOLD POLYP

Polyps and cysts both grow from the vocal fold mucosa. Strong or fluid packed, they may expand to quite large sizes. The size and position of the vocal folds determines the intensity of their vibrations.

The voice may range from only barely intelligible to seriously dysphonic (extremely poor voice quality). People who are suffering from cysts will generally express their complaints with the following: fatigue after a long conversation, and intolerable irritation in the throat.

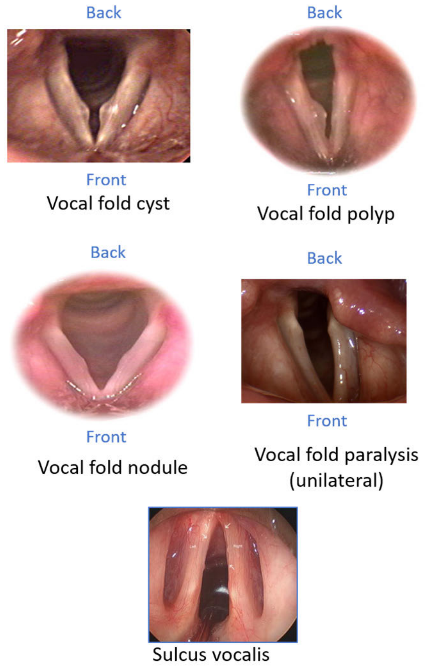


FIGURE 2. Different types of voice pathologies.

C. NODULES

A symmetric prose is known as a vocal fold nodule when it is located on both sides of the vocal folds in the center of the voicebox. When vocal fold nodules stop the folds from closing completely, dysphonia is the most prominent vocal symptom. Young males and females who serve in occupations that rely on speech, including teachers, attorneys, and musicians, are more likely to have vocal fold nodules.

Nodules in the vocal folds may make a voice sound natural, breathy, or extremely raspy. People who are severely deaf or hard of hearing may become unable to talk softly and will find it difficult to produce gentle noises. The sound does not start immediately as a person attempts to talk higher and softer. There is a pause when an audible air escape occurs, and then the sound suddenly begins. Many of these patients suffer from cysts, and their complaints include using their voices for an extended period.

D. PARALYSIS

Paralysis originates from vocal fold immobility. A significant distance between the two vocal folds is due to one or both vocal folds becoming immobile. This distance permits air to escape and interferes with natural movement. Such forms of paralysis are caused by the position of the vocal folds, while others arise because of the contraction of the vocal folds. Another example is where no action is present in the vocal folds; this situation is classified as paralysis. Paresis, which means "weakness," may refer to movement.

Because of paralysis, the voice may be weak, raspy, harsh, two-toned (sounding like two different notes at the same time), or nothing more than a whisper. Many of these patients suffer from cysts, and their complaints include difficulty projecting their speech effectively due to being tired.

E. SULCUS

The sulcus is a linear depression that is located on the mucosal surface of the vocal folds parallel to the free border, and it is characterized by a variable depth and bilateral symmetry. As a consequence of the sulcus vocalis, the vocal folds are inhibited from closing, and, therefore, a mild to a serious degree of dysphonia is generated.

III. LITERATURE REVIEW

There are research works on VPD in the literature. We briefly describe some of them below.

Measurement of voice quality is important for both testing and voice assessment. Auditory perceptual appraisal is used by professional voice therapists who practice in facilities that have access to advanced acoustic, aerodynamic, and vocal fold imaging instrumentation. Furthermore, the APA (American Psychological Association) provides baseline knowledge on the degree of dysphonia, which may be used to evaluate the improvement in a patients' condition over time. The usefulness of auditory-perceptual evaluations of voice dysfunction may be due, at least in part, to the following: low expense, little time required, and patient convenience. Additionally, auditory-perceptual evaluation defines speech quality and vocal intensity by examining particular auditory parameters that are present in sound. It has many problems when it comes to the issue of objectivity: (i) judges consistently disagree with each other, (ii) there are no quantitative metrics, and (iii) a common scale of perceptual measurement is lacking. The subjects assert that scales based on the senses may also have an effect on errors and variability for the following reasons: Scales used in clinical and research settings are, at times, not the best choice for measuring voice quality attributes; when evaluating persons with laryngopharyngeal symptoms, however, scales should not be used solely as a diagnostic instrument, given that low positive and negative predictive values for visual and auditory tests are seen when combined with the laryngoscopic review [17]; several experiments have shown that there are only modest correlations between instrumental tests (i.e., machine measurements of voice quality) and perceptual ratings of voice quality (i.e., the listener's impression of voice quality) [18]. When judges evaluate voice content, they use several different kinds of voice stimuli, including sustained vowels and running speech. Some claim that running speech gives a more accurate and realistic representation of natural speech than do continuous vowels [19]. External rating systems with predefined parameters are used in order to decide whether or not a psychiatric speech condition is present. A common scale, used by both Hirano [20] and Cummings [21], is the GRBAS, according to Hammarberg [22]. This phonetic scales and vocabulary sets were created and adopted in 1969 by the Committee for Tests of Phonatory Functions of the Japan Society of Logopedics and Phoniatrics. G, grade of dysphonia, R, roughness, B, breathiness, A, asthenia, and S, strain make up the acronym GRBAS. For the evaluation of parameters, values are graded into four categories: 0, no deviations; 1, minor variations; 2, mild variations; and 3, significant variations.

Various risk factors may trigger organic and functional larynx disorders that contribute to persistent voice loss (harmful chemicals, neglect of hygienic requirements in jobs with high voice tension, stress, etc.). Disabling the vocal folds to vibrate correctly is one of many modifications that may be carried out in the larynx. Some speech conditions like partial paralysis or total paralysis of the laryngeal muscles, as well as tumors, are causes of these improvements. Hoarseness of the voice or speech is the consequence of voice pathology. Changes in the voice may be caused by any one of the following: A decrease in vocal capacity, an increase in vocal tone, an increase in noise or wind in the original voice, a widening of the vocal spectrum (or an increase in low-frequency level), etc. Any of these indications (symptoms) can be present depending on the form of the voice condition. There are several testing facilities worldwide whose aim is to create diagnostic support approaches that are focused on acoustic voice analysis [23].

Using complex neuronal representations, Hadjitodirov *et al.* [24] studied larynx pathology identification. Their new solution is expected to lead to a significant improvement in the diagnosis of laryngeal disease and, thus, to the elimination of the most common mistake in classifying patients with laryngeal disorders as regular speakers. Probability density functions (PDFs) for the input vectors for regular and pathological topics are used as the foundation of their method, which is called probability distribution map (PDM). Using a template PDM, the PDF of uncertain regular or abnormal topics was also modeled. It was done using a formula derived from unique comparisons, instead of by simply comparing a threshold with any sort of distance/similarity. In their studies, they enhanced the precision of the classification and provided a tool for screening laryngeal pathologies in their paper. Speech analysis methods using nonlinear processing have been proposed and thoroughly tested. The strategies for obtaining voice parameters for speech analysis have been committed. This research demonstrates that the extraction of the first formant AM (amplitude modulation) characteristics are possible during pre and post clinical voice therapy, and analyzed using this algorithm. In this study, the authors state that the extracted function can be linked to safe and pathological patterns of vocal fold vibratory movement.

Many methods have been used in the area of automated voice pathology identification and classification, and we found that similar nonlinear methods were often employed. Due to its failure to work with nonlinearities, this process does not accurately depict machine nonlinearities. These authors [23] analyzed some parameters and the effects of these parameters on function regularity estimation. Additional methods that were used to estimate parameters included the baseline value of 1, which is used in several methods that use the auto mutual information criterion;

and using a parameter that was calculated from the embedding window. Other experiments have shown that nonlinear dynamic regression can be used to examine conventional evidence from both healthy and dysphonic pediatric populations. Additionally, it was discovered that perturbation methods such as jitter analysis can be used to determine the number dysphonic populations, such as infants. An alternate approach to studying pseudo periodic signals using a dynamic network methodology that featured a network-based transformation was shown. Novel methods that could be useful for pathological evaluation were suggested to aid in distinguishing mild and pathological topics. A new methodology to remove features was introduced in [25], in which dynamic networks were referred to as new concepts. This approach demonstrates a graphical way of visually distinguishing healthy individuals from those who are experiencing obesity.

In reference [26], voice samples in compressed MP3 format and other various binary rates (160, 96, 64, 48, 24, and 8 kb/s) were used. This study attempted to define the spoken signal using the Gaussian mixture model (GMM) and support vector machine (SVM) as classifiers. According to Wang *et al.* [27], mel-frequency cepstral coefficients (MFCCs) have an influence on classification, and other approaches (e.g., GMM) may help to boost classification accuracy. The study compared the GMM classifier and the SVM with the GMM classifier to see whether they could identify speech abnormalities. The phonation of sustained vowels was tested, and they discovered that a result of 96.1% was possible. In their research, Jang *et al.* [28] compared many different designed pitch detection algorithms (PDAs) on 99 patients with vocal fold polyps, cysts, and nodules. The research authors then determined that PDAs were adequate. They noticed that when the vocal subject had a higher level of chaotic and aperiodic voice, they saw an increase in the number of pitch mistakes. Another study conducted by Gomez-Velda *et al.* [29] showed how they employed biomechanical measurements as characteristics. This feature was obtained from a sample of vocal sounds and could estimate vocal fold displacement noninvasively. Their research subjects were 52 patients who suffered from polyps, nodules, persistent laryngitis, and Reinke's edema as a result of vocal fold polyp surgery.

In a recent study [30], Vasilakis and Stylianou studied how tiny temporal gaps (or jitter) could be utilized to diagnose vocal disease. Eadie and Doyle [10] used both acoustic and auditory perception measurements in classifying dysphonic voices. Overall voice severity was 48%, whereas dysphonic voice speakers were considered perceptually to have a rating of 40%. The authors used a logistic regression analysis to assess the categorization performance of both auditory-perceptual measurements and acoustic measurements. With these steps, they were able to achieve the correct categorization of objects with respect to their acoustic properties, while also considering how the items were seen. When both acoustic and auditory-perceptual data were merged, the accuracy of categorization was enhanced to 100 percent.

Many patients with dysphonia come to a physician's office. Various pathological situations ranging from functional issues to malignancy may induce vocal abnormalities, as defined by the category of vocal disorders that includes dysphonia. More common among those who use their voices professionally [31]. A detrimental effect on the quality of life, and hence on economic output, may be caused by this condition. Benign lesions of the vocal folds are a significant contributor to dysphonia, and their source is widely characterized according to the tissue layer they originate in and their anatomic location. While benign vocal fold lesions include vocal fold polyps, nodules, and cysts, certain lesions are classed as benign vocal fold lesions as well. Males are more likely to have vocal fold polyps, and vocal fold polyps arise almost exclusively on one side of the vocal fold mucosa. Most of the time, they happen because of some kind of vocal abuse. Polyps result in excessive air stimulation via the process of phonation, and this happens because the condition mostly involves early vocal fatigue, such as frequent voice breaks in singers, and severe dysphonia.

Detecting and classifying speech pathology has the potential to expedite the treatment process and link the medical and IT domains [32], [33]. Voice pathology evaluation cannot be trusted because it varies according to experience and competence. Either an automated system for speech pathology detection and classification may be suboptimal if it is not properly built, or suboptimal design may lead to VPD and classification failure [34].

Dysphonia, that is, disordered speech, has been included in the examination and therapy of the human voice because it is an energetic component in clinical voice evaluations and treatments. The method of perceptual and acoustic measurement is combined with an endoscopic examination of the larynx and voice folds in order to comprehensively evaluate dysphonia. The exam incorporates several evaluation scales, such as a CAPE-V and the GRBAS, which evaluates the overall level of dysphonia, the degree of roughness, breathability, asthenia, and strain. Clinical applications of these procedures, however, may be limited because of the subjectivity of the assessment. Such restrictions may include the clinician's expertise, the degree of dysphonia in the patient, the kind of auditory perception scale and the incentive or speaking function. Researchers and physicians have created a new way to assess the amount of dysphonia a patient has by evaluating their voice using an acoustic analysis. Acoustic analysis yields a numerical number that characterizes the severity of the disease, informs patients about their conditions, and offers therapy and follow-up. In an acoustic analysis, many voice physicians employ sustained vowel samples instead of continuous speech samples for their patients [35].

Various kinds of characteristics, such as long-term and short-term signal analysis, may be used to produce automatic speech pathology identification and categorization. Acoustic analysis may be used to obtain long-term parameters. The short-term parameters are separated into two groups: parametric features and non-parametric characteristics [which are also referred to as parametric features and non-parametric features, respectively]. LPC (linear predictive coding)-based cepstrum (LPCC) and LPC-based parametric characteristics depict the resonant structure of the human vocal cords. Mel-frequency cepstral analysis (MFCC) [36] creates nonparametric characteristics that are similar to the human auditory system. The HMM (hidden Markov model), GMM, vector quantization (VQ), SVM, MLP (multilayer perceptron), NN (neural networks), KNNs (k-means nearest neighbors), LDA (linear discriminant analysis), and LVQ (learning VQ) are used to detect and classify voice disorders. There are several methods in which one may report different parameters. The parameters may be reported as correctly accepted, which means that a pathology has been found, correctly rejected (ii), which means that no pathology has been found, falsely accepted (iii), which means that pathology has not been found, and falsely rejected (iv), which means that pathology has really been found. The results of the automated pathology identification system were previously published in [37]. There were many early studies, and all of them found that long-term acoustic characteristics could detect vocal diseases. However, calculating the fundamental frequency was quite difficult for abnormal voices. This technique is useful for differentiating voice pathologies and distinguishing them from the normal state because of several long-term acoustic features, namely, pitch, shimmer, jitter, APQ (amplitude perturbation quotient), PPQ (pitch perturbation quotient), HNR (harmonic to noise ratio), NNE (normalized noise energy), VTI (voice turbulence index), SPI (soft phonation index), FATR (frequency amplitude tremor), and the glottal to noise excitation ratio (GNE), which are often used in published studies [38]. The vocal fold vibration characteristics, namely jitter, and shimmer, may be used to diagnose both diseased and healthy persons. Both parameters are used in clinical and scientific studies. The aforementioned seven acoustic characteristics, including shimmer and jitter, were recovered using an iterative residual signal estimator developed by Rosa *et al.* [39], and jitter supplied over half of the total pathology diagnosis accuracy (54.8 percent) when used for the detection of 21 diseases.

Classifiers KNN and SVM were employed in [40] to identify two kinds of characteristics, LPC and MFCC, and the classifiers determined that these characteristics fell into three categories. The included samples were representative of the private database's sustained vowel /a/. The spoken samples were categorized as healthy, nodular, and diffuse in this database, which was generated at the Department of Medicine in Lithuania. Accuracy was 67.31% for LPC and 73.08% for MFCC in this study. In [41], a speech pathology detection system was built in which several different tests were done on a sample of speech data from the Massachusetts Eye & Ear Infirmary (MEEI) database. Two distinct kinds of characteristics, MDVP (multidimensional voice parameters) and MFCC, which represented input into various modeling algorithms, were derived from the MEEI database by employing the sustained vowel /a/. Using only the prolonged vowel /a/ to learn about dysphonic and normal voices, the researchers identified MFCC parameters. Different classifiers, such as HMMs, GMMs, SVMs, and ANNs, were all used in this research to identify abnormal voices. Recorded sounds for several vocal illnesses were included in this database, including polyps, palsy, laryngitis, glottis cancer, nodules, and cysts. Using the GMM, the accuracy attained was 95.2%. Extracted from the MEEI database, the LPC characteristics in [42] included the sustained vowel /a/. These examples illustrated several types of voice diseases and normal voices, such as those mentioned above: (i) samples from vocal fold edema, (ii) samples from vocal fold paralysis, and (iii) samples from normal cases. This study tested two classifiers: the KNN and the NN.

Some methodological concerns that surrounded the system's implementation are discussed in [43]. The relevance of having a consistent database for comparing systems or characteristics was noted in [44] by Campbell and Reynolds. There was a significant improvement in voice recognition because of the use of standard speech corpora for testing and development. By comparing speech corpora, researchers can identify which approaches are more accurate and efficient to use. A study of the literature found that voice pathologies are created to determine whether or not the patient is normal or abnormal by the patient's ongoing vowels, namely the vowel /a/. This is usually the case, with most research using a single set of characteristics, and just a few systems using a wide range of characteristics.

Recently, many VPD systems that use deep learning have been developed. The main advantage of using deep learning is that it does not require the extraction of hand-crafted features. In [13], Pavol *et al.* suggested the use of a deep neural network (DNN) to identify voice pathology. A bidirectional long and short-term memory recurrent neural network (BLSTM) was trained on the glottal pulse waveform features. To overcome the challenges of dysarthria detection and speech reconstruction, Daniel *et al.* [46] applied a multi-task learning technology training model. While the annotation data that is presently available is minimal, simulated data were used together with actual data in order to increase the model's resilience.

The design of many state-of-the-art procedures is built upon classifying sound signals, which are affected by things like age, sex, and emotional state. Because dysarthria has a harsh, low-pitched sound, it is particularly difficult to classify. Instead of only mentioning speech phonological features, a more precise explanation should be presented from multiple aspects. Cases where lesions arise on the vocal cords may show two characteristics: the voice may become hoarse, and the vocal cords may not vibrate regularly. The voice's status is reported via the sound signal. This is critical for locating the vocal cord vibration information because the process of EGG signal production represents the change in the contact surface during vocal cord movement.

To arrive at their findings, the study by Wu *et al.* [47] classified voice pathology identification as an image classification issue and used frequency domain transformations on time-domain sound data. This model was based on a short-time Fourier transform (STFT) approach and a CNN network, which consisted of ten convolutional layers with a filter size of 8 × 8.

This does not mean that no data is available, but it does mean that there are presently a limited number of data sets available, and the adoption of ML algorithms may easily lead to overfitting or underfitting. One group of academics is looking at transfer learning mechanisms as a possible solution. In addition, Mohammed *et al.* [48] used transfer learning to create a ResNet34 network that analyzed various sound waves that were processed using STFT and a band-pass filter bank, and they used this data as the input to a CNN network. The resulting accuracy is 95.41%. Using STFT and Mel filters, Guedes *et al.* [49] produced Mel-spectrograms from audio sources. Feature extraction was then done using a VGG network, which was previously utilized for transfer learning. The LSTM ultimately distinguished between pathological and nonpathological vocals, whereas the one-dimensional CNN recognized pathological vocals.

There are some recent deep learning-based pathology detection methods in the literature. In [50] and [51], cloud computing and 5G communications were used to detect pathology. 1D and 2D convolutions were used to extract features from electroencephalogram (EEG) signals for pathology detection in [52].

IV. PROPOSED VPD SYSTEM

In this paper, we propose a VPD system within a smart healthcare framework. The smart healthcare framework consists of several elements: IoTs, deep learning, edge and cloud computing, and 5G communications. Fig. 3 shows such a smart healthcare framework, which is used for the VPD. In the framework, IoTs such as microphones and EGG devices are used to capture the intended signals from the person. These signals are transferred to edge computing for preprocessing, such as for extracting spectrograms. Then, the spectrograms are transferred to cloud computing, where there are AI/ML/deep learning servers and storage. The decision is then conveyed to the stakeholders and the client via 5G.

A. VPD SYSTEM

Fig. 4 shows the proposed VPD system using two modalities: voice signals and EGG signals. The signals are processed separately and are fused at a later stage (after the CNN). The microphone captures the voice signal, while an EGG device captures the EGG signal. EGG device is put around the vocal folds. Spectrograms are obtained from these signals using the following consecutive modules: bias removal, short-time framing (30 ms frames with 10 ms overlapping), hamming windowing, and STFT. For the investigation, we also use a Mel-spectrogram, which is obtained by applying Mel scale-spaced band-pass filtering (36 Mel filters). High-order harmonic distortion in the spectrogram is reduced by pre-processing the voice sample before STFT.

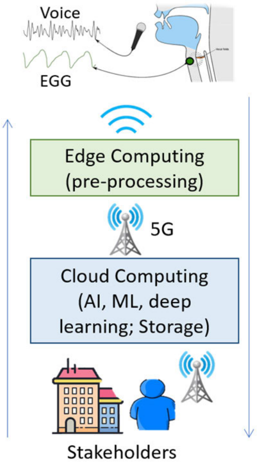

Voice
EGG
Edge Computing (pre-processing)
5G
Cloud Computing (AI, ML, deep learning; Storage)
Stakeholders

FIGURE 3. Smart healthcare framework for the VPD system.

Feature mapping data is reduced when sampling is performed at a frequency of 16 kHz, as doing so decreases the quantity of data for training, making the process faster. In order to boost the high-frequency resolution of the speech, pre-emphasis is applied to the frame.

The spectrograms are fed into a pre-trained CNN model. In the experiments, we used ResNet50 [53], Xception [54], and MobileNet [55]. The pre-trained models are used to train the system quickly because we do not have a large number of samples. Table 1 shows the general information for these three CNN models. MobileNet has much fewer parameters than ResNet50 and ResNet50, and can, therefore, be used in real-time applications. However, due to the rapid increase in processing speed, a CNN with a large number of parameters can also be used for fast processing. ResNet50 and MobileNet have an input size of 224 × 224, while Xception has an input size of 299 × 299. Based on the input size, the spectrograms and the Mel-spectrograms are resized accordingly.

Fig. 5 shows examples of (top row) a voice signal and an EGG signal of a healthy person, (middle row) corresponding spectrograms, and (bottom row) corresponding Mel-spectrograms.

LSTM units [39] make up the LSTM model. Input, forget, and output are the three gates that regulate the LSTM unit. Current time data and concealed historical time data flow through the LSTM gates. Three completely linked layers using the sigmoid function compute the values of the input, forget, and output gates. An LSTM layer may be created using stacked LSTM units. Either bidirectional or unidirectional LSTM may be formed from these LSTM layers.

Two layers operate in the forward and backward time directions about each other in a bidirectional LSTM (BiLSTM). Time-dependent relationships may be learned using these successive layers. Each BiLSTM layer has 256 stacked LSTM blocks. Softmax is used in the final BiLSTM layer to classify the embedded patterns. We first train the CNN model to extract the features, and then we freeze it to preserve the

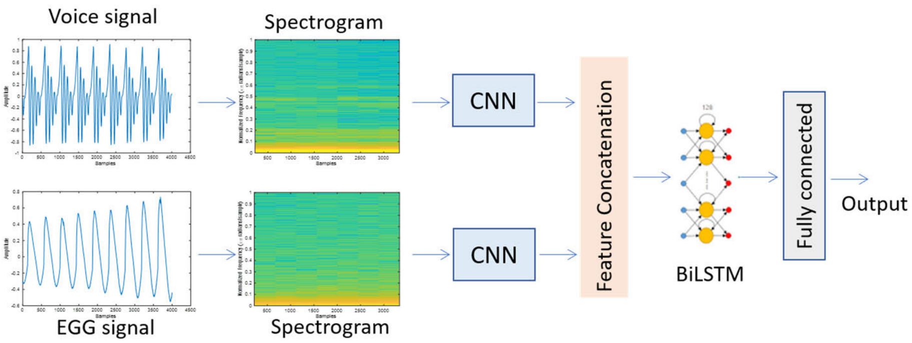


FIGURE 4. Proposed system architecture: a) a diagram of the proposed microphone and environment classification system from audio; b) the deep neural network architecture using CRNN.

TABLE 1. Pre-trained CNN model information.

CNN model	*Xception*	*ResNet50*	*MobileNet*
Input size	299×299	224×224	224×224
Parameters	22.9 Million	25.6 Million	4.2 Million
Size (MB)	88	96	16
Depth	126	—	88
Layers number	71	50	28

retrieved features and feed them to the BiLSTM model for temporal feature extraction.

In the proposed system, a dropout of 50% is applied before the fully connected layer. Cross-entropy is used as the loss function.

B. DATABASE

The SVD database [16] is extensively used in speech pathology detection research, which comprises voice recordings of more than 2000 persons and 71 different voice pathologies. To train, test, and validate the system, the sustained vowel /a/ voice signal and EGG signals were employed. For each training sample, there were 842 groups; there were 791 pathological groups, which is comparable to 60% of the overall sample; for each verification sample, there were 281 healthy groups, which was equal to 20% of the whole sample. In the experiments, we used samples of speakers in the age group from 15 years to 60 years.

V. EXPERIMENTS

Several experiments were carried out to validate the proposed VPD system. The proposed system was compared with other related systems in the literature. The four performance metrics accuracy, recall, precision, and F1-score were used to measure the effectiveness of various systems.

Accuracy is the percentage of all samples that were correctly predicted to have been included in the set. Precision rate is a measurement that reflects the ability of the model to accurately detect negative samples. Recall rate is the percentage of positive test results that are expected to be positive. What this means is that greater recall shows that the model is better at recognizing positive samples, which is extremely important when conducting an experiment. A higher F1 score suggests stronger categorization abilities.

To optimize the parameters of the model, an Adam optimizer was used. The learning rate was 10^{-4}, the batch size was 32, and the number of training epochs was 200. The optimization method used an Adam optimizer, where the learning rate was 10^{-4}, the batch size was 32, and the number of training epochs was 100.

A. ACCURACY AND LOSS CURVES OF THE MODEL

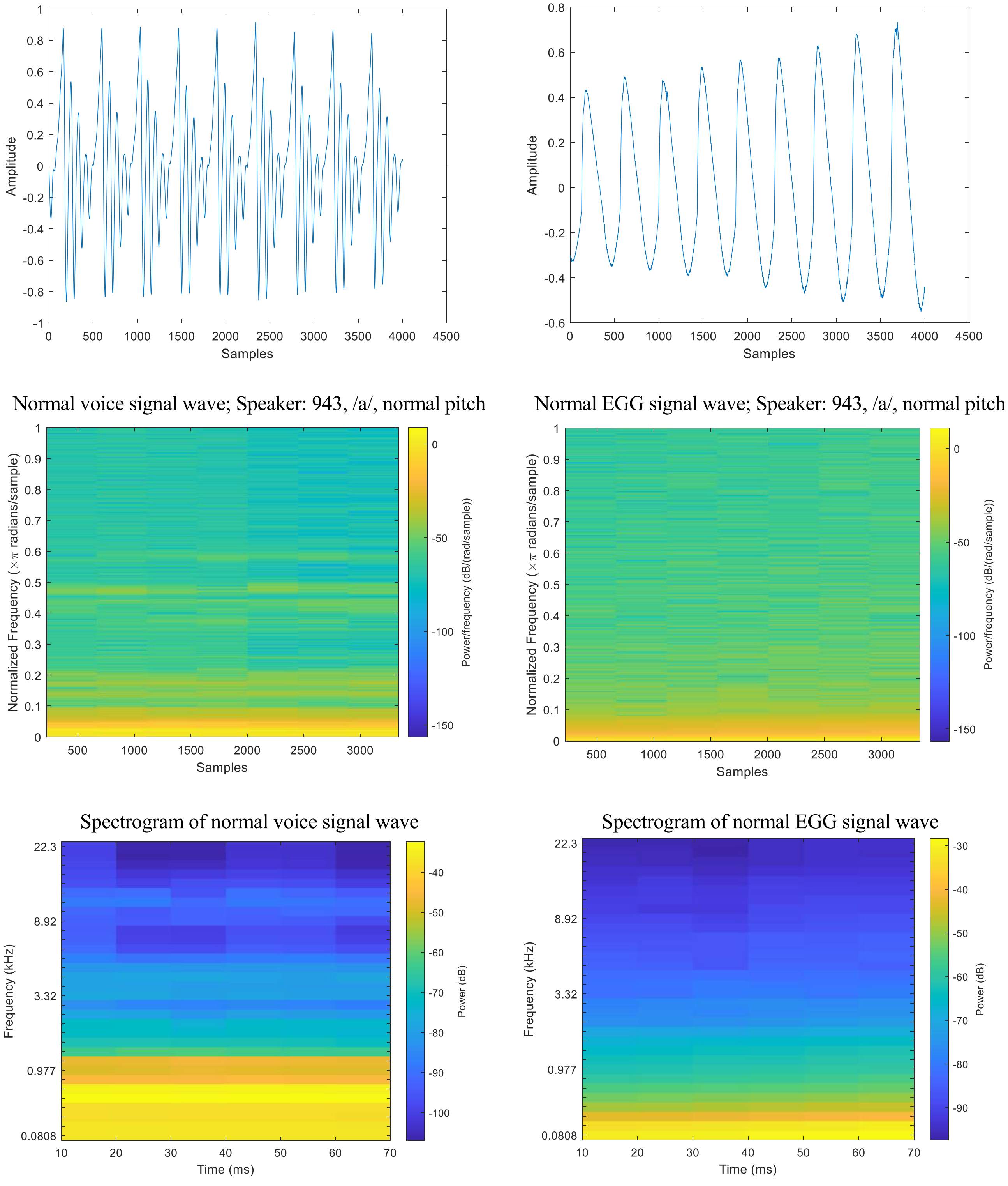
Fig. 6 and Fig. 7 show the accuracy curve and the loss curve, respectively, of the model in the proposed system. In this case, Xception was used as the CNN model.

B. AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE (AUC)

Fig. 8 shows the ROC curve of the proposed system using the Xception and BiLSTM models. The AUC was 0.998. The 95% confidence interval was [0.987 0.998], showing the implications of the data in the two classes (normal and pathological).

C. PERFORMANCE OF THE PROPOSED SYSTEM

Table 2 shows the values of the performance metrics for the proposed system. We compared performance between the single modality with voice, the single modality with EGG,



Normal voice signal wave; Speaker: 943, /a/, normal pitch

Normal EGG signal wave; Speaker: 943, /a/, normal pitch

Spectrogram of normal voice signal wave

Spectrogram of normal EGG signal wave

Mel-spectrogram of normal voice signal wave

Mel-spectrogram of normal EGG signal wave

FIGURE 5. Examples of voice signal and EGG signal, and their corresponding spectrograms and Mel-spectrograms.

and the bi-modality (the proposed system). From the table, we see that the proposed system performed better than the single modality. Therefore, the fusion of voice and EGG signals improved the performance of the VPD system. The performance reported in Table 2 was obtained using the Xception model.

TABLE 2. Performance of the proposed system. The numbers are percentages.

Modality	Accuracy	Precision	Recall	F1-score
Voice	93.94	95.08	94.87	94.93
EEG	93.71	94.86	94.77	94.25
Fusion (proposed)	95.65	95.56	95.78	95.64

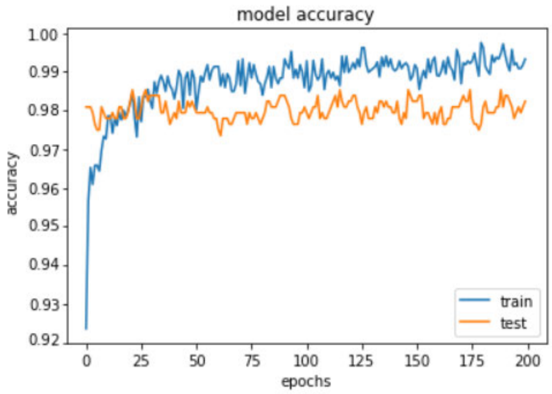


FIGURE 6. Accuracy curve of the model in the proposed system.

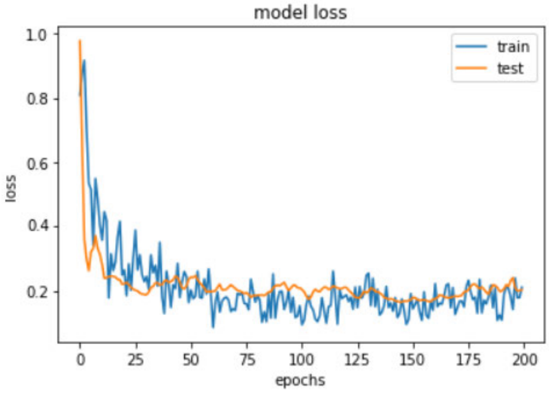


FIGURE 7. Loss curve of the model in the proposed system.

We investigated three different pre-trained CNN models in the proposed system. Fig. 9 shows the accuracy of the system using these CNN models. From the figure, we see that Xception performed better than the other two. It can be noted that though MobileNet used fewer parameters than the other two, it did not perform poorly.

We compared the proposed system with other related systems using the same database. It can be noted that we seldom found the fusion of voice and EGG signals in the literature. Table 3 shows the accuracy comparison of various systems. From the table, we see that the proposed system outperformed all other compared systems.

There are many other AI and IoT-based smart healthcare systems [57]; however, the application of VPD in smart healthcare is limited in the literature. As teachers are greatly affected by voice pathology, a smart class environment could integrate a VPD system for sustainable teaching [58].

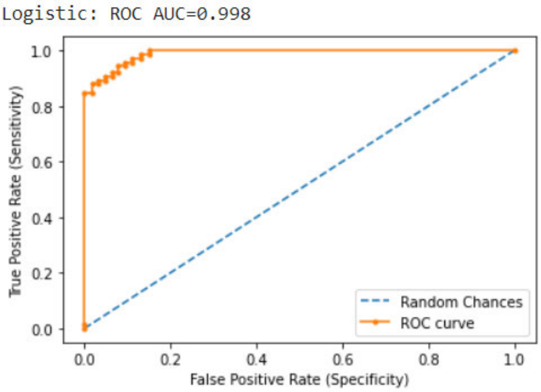


FIGURE 8. ROC curve of the proposed system.

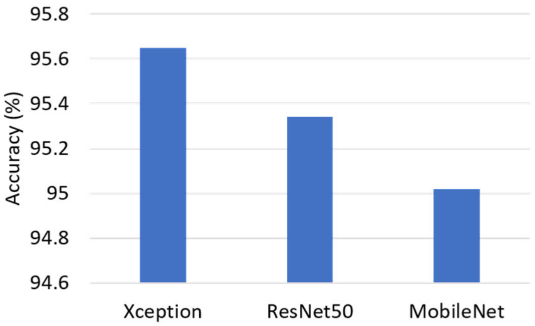


FIGURE 9. Accuracy of the system using three different pre-trained CNN models.

TABLE 3. Comparison of accuracies (%) of different systems.

System [41]: voice	System [8]: voice	System [56]: voice	System [12]: bimodal	Proposed: bimodal
91.0	92.8	93.9	94.2	95.6

VI. CONCLUSION

This work proposed a pathological voice detection method based on bimodal input. The method takes the voice signal and EGG signal as inputs. The proposed VPD system extracts spectrograms from the signals and feeds them to a CNN model. Later, the extracted features from the two modalities are fused and fed to the BiLSTM model. The experimental results showed that the proposed system achieved greater than 95% accuracy, precision, and recall. The system also outperformed other related systems. It was demonstrated that bimodal inputs were better than single inputs.

As future work, we will investigate the effect of signal transmission over a network in the VPD system. In addition, we may use attention mechanism in deep learning to improve the performance of the system.

REFERENCES

[1] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *J. Speech, Lang., Hearing Res.*, vol. 47, no. 2, pp. 281–293, Apr. 2004.

[2] A. Behrman, "Common practices of voice therapists in the evaluation of patients," *J. Voice*, vol. 19, pp. 454–469, Sep. 2005.

[3] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," *Biomed. Eng. OnLine*, vol. 10, no. 1, p. 41, 2011.

[4] N. P. Solomon, L. B. Helou, and A. Stojadinovic, "Clinical versus laboratory ratings of voice using the CAPE-V," *J. Voice*, vol. 25, no. 1, pp. e7–e14, Jan. 2011.

[5] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *Amer. J. Speech-Lang. Pathol.*, vol. 18, no. 2, pp. 124–132, May 2009.

[6] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, and A. Alamri, "Smart health solution integrating IoT and cloud: A case study of voice pathology monitoring," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 69–73, Jan. 2017.

[7] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model," *J. Voice*, vol. 30, no. 6, pp. 757.e7–757.e19, 2016.

[8] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, Dec. 2018.

[9] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *J. Acoust. Soc. Amer.*, vol. 87, no. 5, pp. 2218–2224, May 1990.

[10] T. L. Eadie and P. C. Doyle, "Classification of dysphonic voice: Acoustic and auditory-perceptual measures," *J. Voice*, vol. 19, no. 1, pp. 1–14, Mar. 2005.

[11] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Detection of voice pathology using fractal dimension in a multiresolution analysis of normal and disordered speech signals," *J. Med. Syst.*, vol. 40, no. 1, Jan. 2016, p. 10.

[12] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Syst.*, vol. 25, no. 5, pp. 565–675, 2019.

[13] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, and M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, pp. 10745–10753, Dec. 2019.

[14] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio–visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[15] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[16] *Saarbruecken Voice Database*. Accessed: Apr. 20, 2021. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4

[17] C. A. Eckley, W. Anelli, and A. D. C. Duprat, "Auditory voice-perception analysis sensitivity and specificity in the screening of laryngeal disorders," *Brazilian J. Otorhinolaryngol.*, vol. 74, no. 2, pp. 168–171, Mar. 2008.

[18] J. Oates, "Auditory-perceptual evaluation of disordered voice quality: Pros, cons and future directions," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[19] I. V. Bele, "Reliability in perceptual analysis of voice quality," *J. Voice*, vol. 19, pp. 555–573, Dec. 2005.

[20] M. Hirano, *Clinical Examination of Voice*. New York, NY, USA: Springer-Verlag, 1981.

[21] L. Cummings, *Clinical Linguistics*. Edinburgh, U.K.: Edinburgh Univ. Press, 2008.

[22] B. Hammarberg, "Voice research and clinical needs," *Folia Phoniatrica et Logopaedica*, vol. 52, nos. 1–3, pp. 93–102, Jan./Jun. 2000.

[23] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in smart cities," *IEEE Access*, vol. 5, pp. 3900–3908, 2017.

[24] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Trans. Inf. Technol. Biomed.*, vol. 4, no. 1, pp. 68–73, Mar. 2000.

[25] A. Alamri, M. M. Hassan, M. A. Hossain, M. Al-Qurishi, Y. Aldukhayyil, and M. S. Hossain, "Evaluating the impact of a cloud-based serious game on obese people," *Comput. Hum. Behav.*, vol. 30, pp. 468–475, Jan. 2014.

[26] N. Saenz-Lechon, V. Osma-Ruiz, J. I. Godino-Llorente, M. Blanco-Velasco, F. Cruz-Roldan, and J. D. Arias-Londono, "Effects of audio compression in automatic detection of voice pathologies," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 12, pp. 2831–2835, Dec. 2008.

[27] X. Wang, J. Zhang, and Y. Yan, "Discrimination between pathological and normal voices using GMM-SVM approach," *J. Voice*, vol. 25, no. 1, pp. 38–43, Jan. 2011.

[28] S.-J. Jang, S.-H. Choi, H.-M. Kim, H.-S. Choi, and Y.-R. Yoon, "Evaluation of performance of several established pitch detection algorithms in pathological voices," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Aug. 2007, pp. 620–623.

[29] P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, F. Díaz, F. J. Fernández-Camacho, V. Rodellar, A. Álvarez, and R. Martínez, "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters," *J. Voice*, vol. 21, no. 4, pp. 450–476, Jul. 2007.

[30] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009.

[31] M. Chen, J. Yang, L. Hu, M. S. Hossain, and G. Muhammad, "Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 14–20, Nov. 2018.

[32] M. Masud, M. S. Hossain, and A. Alamri, "Data interoperability and multimedia content management in e-Health systems," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1015–1023, Nov. 2012.

[33] M. S. Hossain, "Cloud-supported cyber–physical localization framework for patients monitoring," *IEEE Syst. J.*, vol. 11, no. 1, pp. 118–127, Mar. 2017.

[34] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and A. Al-Nasheri, "Automatic voice pathology detection and classification using vocal tract area irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016.

[35] Z. Ali, M. S. Hossain, G. Muhammad, and A. K. Sangaiah, "An intelligent healthcare system for detection and classification to discriminate vocal fold disorders," *Future Gener. Comput. Syst.*, vol. 85, pp. 19–28, Aug. 2018.

[36] B.-H. Juang and L. Rabiner, *Fundamentals of Speech Recognition* (Signal Processing Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[37] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1s, pp. 1–17, Feb. 2019.

[38] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[39] M. D. O. Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan. 2000.

[40] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 36–47, Jul. 2008.

[41] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *J. Voice*, vol. 31, no. 1, pp. 3–15, Jan. 2017.

[42] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition methods," in *Proc. 29th Annu. Int. Conf. IEEE EMBS*, Lyon, France, Aug. 2007, pp. 3253–3256.

[43] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, Apr. 2006.

[44] J. P. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, USA, Mar. 1999, pp. 829–832.

[45] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *Proc. Int. Conf. Workshop Bioinspired Intell. (IWOBI)*, Funchal, Portugal, Jul. 2017, pp. 1–4.

[46] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, ''Interpretable deep learning model for the detection and reconstruction of dysarthric speech,'' in *Proc. Int. Speech Commun. Assoc. (Interspeech)*, Graz, Austria, Sep. 2019, pp. 1–6.

[47] H. Wu, J. Soraghan, A. Lowit, and G. D. Caterina, ''Convolutional neural networks for pathological voice detection,'' in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Honolulu, HI, USA, Jul. 2018, pp. 1–4.

[48] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. K. A. Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. Al-Dhief, ''Voice pathology detection and classification using convolutional neural network model,'' *Appl. Sci.*, vol. 10, no. 11, p. 3723, May 2020.

[49] V. Guedes, F. Teixeira, A. Oliveira, J. Fernandes, L. Silva, A. Junior, and J. P. Teixeira, ''Transfer learning with audioset to voice pathologies identification in continuous speech,'' in *Proc. Int. Conf. Enterprise Inf. Syst. (CENTERIS)*, Sousse, Tunisia, 2019, pp. 662–669.

[50] M. S. Hossain and G. Muhammad, ''Deep learning based pathology detection for smart connected healthcare,'' *IEEE Netw.*, vol. 34, no. 6, pp. 120–125, Nov./Dec. 2020.

[51] G. Muhammad, M. F. Alhamid, and X. Long, ''Computing and processing on the edge: Smart pathology detection for connected healthcare,'' *IEEE Netw.*, vol. 33, no. 6, pp. 44–49, Nov. 2019.

[52] G. Muhammad, M. S. Hossain, and N. Kumar, ''EEG-based pathology detection for home health monitoring,'' *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 603–610, Feb. 2021.

[53] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[54] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[55] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[56] M. Alhussein and G. Muhammad, ''Voice pathology detection using deep learning on mobile healthcare framework,'' *IEEE Access*, vol. 6, pp. 41034–41041, Dec. 2018.

[57] F. Alshehri and G. Muhammad, ''A comprehensive survey of the Internet of Things (IoT) and AI-based smart healthcare,'' *IEEE Access*, vol. 9, pp. 3660–3678, Jan. 2021.

[58] A. Alelaiwi, A. Alghamdi, M. Shorfuzzaman, M. Rawashdeh, M. S. Hossain, and G. Muhammad, ''Enhanced engineering education using smart class environment,'' *Comput. Hum. Behav.*, vol. 51, pp. 852–856, Oct. 2015.

GHULAM MUHAMMAD (Senior Member, IEEE) received the B.S. degree in computer science and engineering from the Bangladesh University of Engineering and Technology, in 1997, and the M.S. and Ph.D. degrees in electronic and information engineering from the Toyohashi University of Technology, Japan, in 2003 and 2006, respectively. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He is a principal investigator and a co-principal investigator in many research projects. He has authored or coauthored more than 250 publications, including for IEEE/ACM/Springer/Elsevier journals, and flagship conference papers, and he holds two U.S. patents. His research interests include signal processing, machine learning, the IoT, medical signal and image analysis, AI, and biometrics. He was a recipient of the Japan Society for Promotion and Science (JSPS) Fellowship from the Ministry of Education, Culture, Sports, Science, and Technology, Japan. He received the Best Faculty Award of the Computer Engineering Department, KSU, from 2014 to 2015. He has supervised more than 15 Ph.D. and master's theses.

MUSAED ALHUSSEIN (Member, IEEE) was born in Riyadh, Saudi Arabia. He received the B.S. degree in computer engineering from King Saud University, Riyadh, in 1988, and the M.S. and Ph.D. degrees in computer science and engineering from the University of South Florida, Tampa, Florida, in 1992 and 1997, respectively. Since 1997, he has been a faculty of the Computer Engineering Department, College of Computer and Information Science, King Saud University, where he is currently the Founder and the Director of the Embedded Computing and Signal Processing Research Laboratory. His research interests include computer architecture and signal processing, and in particular on VLSI testing and verification, embedded and pervasive computing, cyber-physical systems, mobile cloud computing, big data, e-healthcare, and body area networks.