A novel convolutional neural network based dysphonic voice detection algorithm using chromagram

Rumana Islam[1], Mohammed Tarique[2]

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, University of Windsor, Windsor, Canada
[2]Department of Electrical and Computer Engineering, College of Engineering and Technology, University of Science and Technology of Fujairah, Fujairah, United Arab Emirates

Article Info	ABSTRACT
	This paper presents a convolutional neural network (CNN) based non-invasive pathological voice detection algorithm using signal processing approach. The proposed algorithm extracts an acoustic feature, called chromagram, from voice samples and applies this feature to the input of a CNN for classification. The main advantage of chromagram is that it can mimic the way humans perceive pitch in sounds and hence can be considered useful to detect dysphonic voices, as the pitch in the generated sounds varies depending on the pathological conditions. The simulation results show that classification accuracy of 85% can be achieved with the chromagram. A comparison of the performances for the proposed algorithm with those of other related works is also presented.

Corresponding Author:

Rumana Islam
Department of Electrical and Computer Engineering, Faculty of Engineering, University of Windsor
401 Sunset Avenue, ON N9B 3P4, Canada
Email: islamq@uwindsor.ca

1. INTRODUCTION

Voice disability is a barrier to effective human speech communication. Primarily, voice disability occurs due to improper function of the components that constitute the human voice generation system [1]–[3]. Various voice disabilities have been reported in the literature. The American speech-language-hearing association (ASHA) has identified dysphonia [4] as the most common voice pathology. Dysphonia refers to abnormal voices [5], [6] that can develop suddenly (or gradually) over time. Commonly, dysphonia is caused by inflamed vocal folds that cannot vibrate properly to produce normal voice sounds [7], [8].

Both invasive and non-invasive methods are used for pathological voice detection. In invasive methods, physicians insert probe into the patient's mouth using endoscopic procedures, namely, laryngoscopy [9], stroboscopy [10], and laryngeal electromyography [11]. In non-invasive methods, audio signals are used for voice pathology detection. Audio signals including cough, breathing, and voice have been popularly used in many applications including telecommunication, robot control [12], domestic appliance control, data entry, voice recognition, disease diagnosis [13], [14], and speech-to-text conversion [15], [16]. Voice signals are mainly used to implement non-invasive voice pathology detection algorithms [17]. The two-fold objectives of these methods are i) to reduce the discomfort of a patient and ii) to assist the clinicians in preliminary diagnosis of voice pathology. In this method, voice samples are collected in a controlled environment and acoustic features are extracted from the voice samples. The next step is to classify the samples into two categories namely normal (i.e., healthy) and pathological using a classifier algorithm. Numerous classifier algorithms have been suggested in the literature for pathological voice detection. Recently, machine learning and deep learning algorithms have drawn considerable attention from

researchers [18]–[22]. Specifically, deep learning algorithms demonstrate promising results and provide high accuracy in voice pathology detection [19]–[22].

Various voice disability detection algorithms have been published in the literature. In [23], the authors have used a spectrogram to detect voice pathology. They have suggested minimizing the effects of jitter, shimmer, and harmonic-to-noise ratio (HNR) to improve the classification accuracy with the spectrogram. In [24], eight-voice pathologies have been investigated, and the results show that deep neural network (DNN) based classifier achieves a high accuracy (94.26% and 90.52% in male and female subjects, respectively).

Vocal disorders, namely neoplasm, phono-trauma, and vocal palsy have been investigated in [25]. The authors used a dense net recurrent neural network (DNRNN) in their work. The results show that the DNRNN algorithm achieves an accuracy of 71%. In another similar work [26], multiple neural networks have been used to detect voice pathology. The authors have used multilayer perceptron neural network (MLPNN), general regression neural network (GRNN), and probabilistic neural network (PNN) in this work, and they achieved the highest accuracy with the MLPNN. Some researchers suggested using multiple voice features to increase accuracy. For example, the researchers in [27] have used six-voice features, namely jitter, shimmer, harmonic-to-noise ratio (HNR), soft phonation index (SPI), amplitude perturbation quotient (APQ), and relative average perturbation (RAP). They have achieved the classification accuracy of 95.20% and 84.20% with generalized method of moments (GMM) and artificial neural network (ANN), respectively.

Support vector machine (SVM) and radial basis function neural network (RBFNN) have been used in [28] to detect voice pathology. In their work, the authors have also used several features, and they achieved an accuracy of 91% with RBFNN. On the other hand, SVM achieved an accuracy of 83%. Stuttering voice has been addressed in [29]. The authors developed a classifier that can detect stuttered speech in their work. The results presented in their work show that these algorithms can detect stuttered voices with an accuracy of 85% and 78% for males and females, respectively. Four voice attributes, namely roughness, breathiness, asthma, and strain have been considered in [30]. The proposed algorithm used these features for the classification by using a feed-forward neural network (FFNN) and the algorithm achieved an average F-measure of 87.25%.

Some researchers claim that spectrogram is the most suitable voice feature to detect voice pathology because it traces different frequencies and their occurrences in time. For example, the authors have used spectrogram to detect pathological voice disorder due to vocal cord paralysis in [31]. The spectrograms of pathological and normal speech samples are applied to the input of a convolutional deep belief network (CDBN). The authors have achieved 77% and 71% accuracy for CNN and CDBN, respectively.

Dysphonic voice detection using a pre-trained CNN has been presented in [32]. The results show that the proposed method can detect dysphonic voices with an accuracy of 95.41%. To detect dysphonic voice, a new marker called the dysphonic marker index (DMI) has been introduced in [33]. The marker consists of four acoustic parameters. The authors have employed a regression algorithm to relate this marker to discriminate pathological voices from healthy ones. A novel computer-aided pathological voice classification system is proposed in [34]. In this work, the authors have used a deep-connected ResNet for classification.

There are two significant limitations of the above-mentioned related works. One limitation is that none of the works considers the way human auditory system perceives the pitch contained in the sounds. Another limitation is that these works use classifiers that overwhelm the system with a substantial computational burden. The limitations mentioned above are overcome in this work by using the chromagram feature and a CNN. Chromagram has been commonly used to detect musical sounds, as far as our knowledge goes, we are the first to use the chromagram in pathological voice detection. The main contributions of this work are: i) developing a novel pathological voice detection algorithm based on signal processing and a deep learning approach, ii) extracting the chromagram feature from the voice samples and converting them into a suitable form for classification, iii) achieving a high classification accuracy without overwhelming the system with a vast computation burden, iv) providing a detailed performance analysis of the proposed system in terms of classification matrix, precision, sensitivity, specificity, and F1-score, and v) comparing the performances of the proposed algorithm with other related works to demonstrate its effectiveness. The rest of the paper is organized as follows: materials and methods are presented in section 2. The results are presented in section 3, and the paper is concluded with section 4.

2. MATERIALS AND METHODS

This investigation uses normal and dysphonic voice samples that are collected from the Saarbrucken voice database (SVD) [35]. The SVD database contains 687 normal (i.e., healthy) samples and 1,356 pathological samples. The samples contain the recordings of vowels and sentences. In this investigation, the

sustained phonation of the vowel '/a/' is used. The main reason is that a speaker can maintain a steady frequency and amplitude at a comfortable level during the voice generation of the vowel '/a/'. Moreover, it is free of articulatory and other linguistic confounds that often exist with other common speech components, including sentences and running speeches. Figure 1 shows the plots for a normal voice sample and a dysphonic voice sample that are randomly selected from the SVD database. It is depicted in the figure that the dysphonic voice samples demonstrate irregular distortion both in terms of magnitude and shape compared to that of the healthy sample.
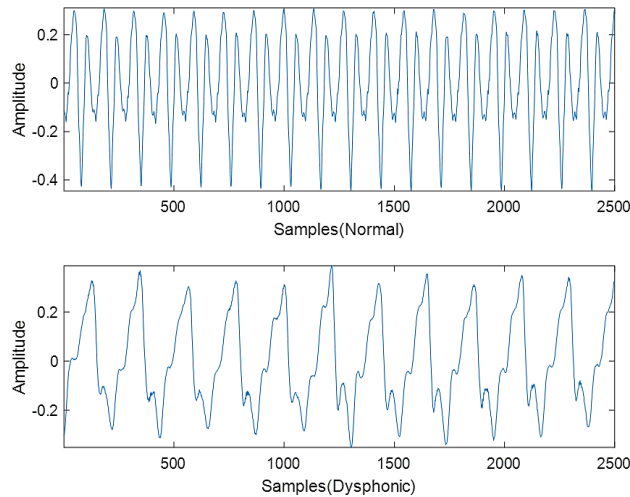

Figure 1. The normal and dysphonic voice samples

In this work, the chromagram audio feature has been used for the classification purpose. The chromagram is derived based on the principle of human's auditory perception of pitch in the sound. The pitch is represented as shown in Figure 2. Generally, humans perceive the pitch as extending along a scale from low to high on a circular scale traversing in the clockwise direction as shown in Figure 2(a). In addition, the pitch has a linear scale too. To accommodate both linear and circular dimensions, Shepard introduced the concept of chroma in [36]. He suggested that the human auditory system's pitch perception is better represented as a helix as shown in Figure 2(b). According to Shepard, the pitch, p perceived by a human, can be expressed in terms of chroma, c, and tone-height, h by (1).

$$p = 2^{h+c} \tag{1}$$

Patterson generalized Shepard's model in [37] to find the pitch by using (2).

$$f = 2^{h+c} \tag{2}$$

The chroma spectrum, $S(c)$ is defined as the measure of the strength of a signal with a given value of Chroma. This is analogous to the standard Fourier power spectrum of a signal. Then, the chroma spectrum is extended in the time domain to create a time-frequency distribution (TFD) $S(c, t)$. This distribution is called the chromagram and it represents a joint distribution of signal strength over the variables of chroma and time.

The chromagram is generated by two major steps in this work, namely frame segmentation and feature calculation. Rather than using a uniform frame size, a beat-synchronous frame segmentation is used in this work. This allows the frame sampling to track the rhythm of the voice. The second step of the algorithm is the feature calculation. A 12-element representation of the spectral energy has been used to calculate the chroma vector. Then, the chroma vector is computed by grouping the discrete Fourier transform (DFT) coefficients of a short-term window into 12 bins. Each bin represents 12 equal-tempered pitch classes of semitone spacing. Each bin produces the mean of log-magnitudes of the respective DFT coefficients defined by (3).

$$v_k = \sum_{n \in S_k} \frac{X_i(n)}{N_k}, \; k \in 0 \ldots 11 \tag{3}$$

where, S_k is a subset of the frequencies that correspond to the DFT coefficients and N_k is the cardinality of S_k. This computation results in a matrix, V with elements $V_{k,i}$, where indices k and i represent pitch-class and frame-number, respectively. The matrix, $V_{k,i}$ is represented in a suitable form to produce the chromagram. The detailed algorithm of computing the chroma vector is shown in Algorithm 1. The chromagram of normal and dysphonic voices is plotted in Figure 3. This figure shows the distinct differences between the chromagram of the normal and dysphonic voices. For example, the chromagram of the normal voice shows a few dominant coefficients, which are stable for a short period. However, the chromagram of dysphonic voice is noisier compared to that of normal voice.

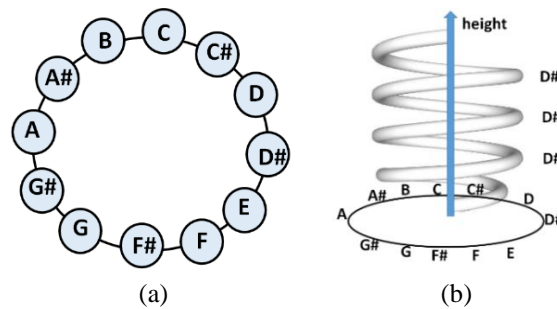(a)   (b)

Figure 2. The pitch representation (a) circular configuration of pitch and (b) Shepard's helix model [37]

Algorithm 1. Generating the chromagram of voice signals

```
    /*Set the initial value of the parameters */
Set the number of tone heights h = 12;
Set the number of bin B = 12;
Set the fundamental frequency f₀ = 55 Hz;
   Load the sound data into vector X;
Read the sampling rate from the soud file and store into variable Fs;
Normalize the data by X_norm = X/max(X);
Set window size W;
Calculate the number of frame N = length(X)/W;
    /* Set the frequency range to compute the spectrogram*/
Set minimum frequency f_min = 0;
Calculate the maximum frequency f_max = Fs/W ;
Calculate the chromatic scale f_i = f₀2^(i/h)
Compute the discrete-time Fourier transform of the signal vector X(k)
   /*Determine the log-magnitudes of the respective DFT co-efficients */
while i ≤ N
      while k ≤ B
             v_k = Σ_(n∈S_k) X_i(n)/N_k
      end while
      Convert v_k into V_(k,i)
end while
Generate chromagram C(m,t)
```

In this work, a CNN is employed as the binary classifier. The CNN model presented in [38] is used as the base to implement the classifier. The CNN model includes two networks namely feature extraction network and classifier network. The input data (i.e., chromagram) is applied to the feature extraction network. The extracted feature map is applied to the classification neural network. The feature extractor network consists of piles of convolutional layer and pooling layer pairs as shown in Figure 4. To avoid computational burden on the system, one convolutional layer is used as the feature extractor network in this work. The CNN model uses 20 convolutional filters of size 9×9. The feature map produced by the convolutional filters is processed by an activation function. The rectified linear unit (ReLU) function is used for this purpose. The output produced by the convolutional layer is then passed through the pooling layer. In this work, a 2×2 matrix for pooling is used to find the mean value from the input data. The hidden layer has 100 nodes that also use the ReLU activation function. The output layer of the CNN contains a single node as the decision made by the classifier is binary and the SoftMax function is used as the activation function at the output node.
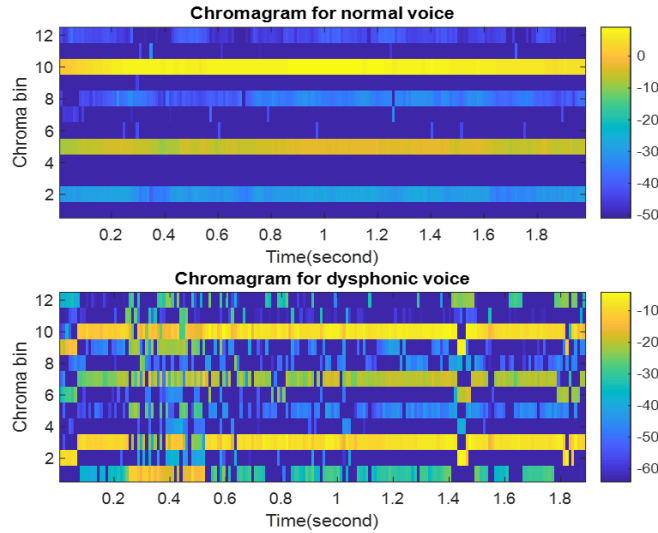
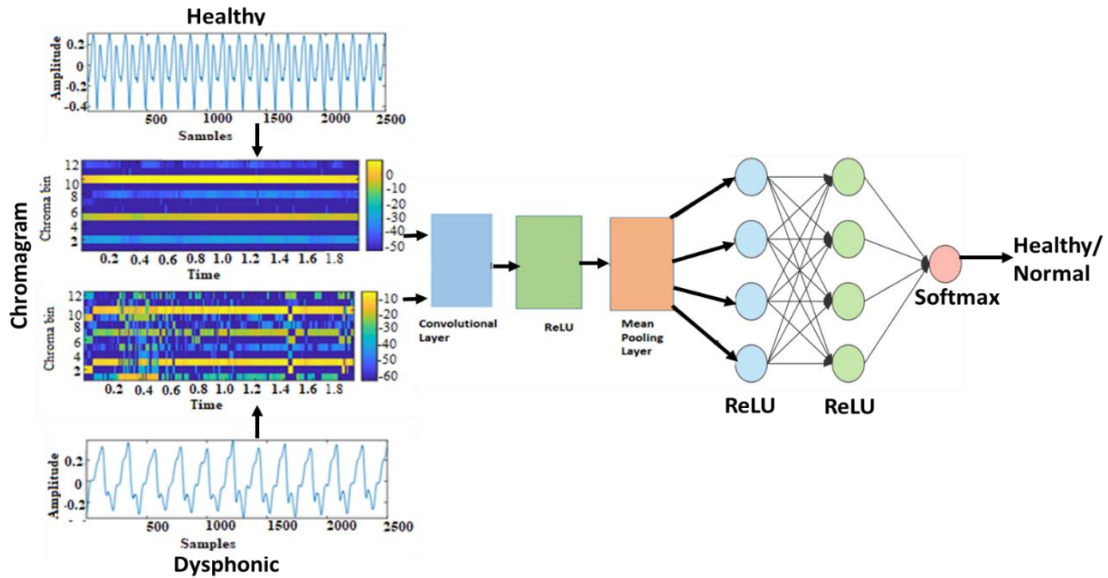Figure 3. The chromagram of normal voice and dysphonic voice

Figure 4. The architecture of the convolutional neural network

3.   RESULTS AND DISCUSSION

To measure the performances of the proposed algorithm, the four machine learning performance parameters, namely, accuracy, precision, recall, and F1 score [39], [40] are used. Ten simulations were conducted by using the chromagram of normal and pathological samples. First, the CNN is trained with the chromagram of 50 normal samples and 50 pathological samples. A five-fold cross-validation method is used to ensure the accuracy of the training. Once trained, the chromagram of 50 other normal and pathological samples are used to test the network's performance. The proposed algorithm's training, validation, and testing results are listed in Table 1. This table shows the proposed system achieves an average accuracy of 98.33%, 94.11%, and 84.99% for training, validation, and testing respectively.

The corresponding classification matrix is shown in Table 2. Based on the data presented in Table 2, it is observed that the proposed system can correctly detect pathological voices with an accuracy of 90.00%. On the other hand, the system can detect normal voices with an accuracy of 80.00%. This accuracy shows that the system performs almost equally in detecting normal and dysphonic voices. The classification matrices also show that the chromagram mistakenly classifies the pathological voices as a normal voice with a probability of 10.00% only. Also, the chromagram misclassifies the healthy samples as pathological samples with a probability of 20%. The performances of the proposed algorithm are listed in Table 3. This

table also shows that the proposed algorithm achieves accuracy, precision, recall, and F1 score of 85.00%, 81.80%, 90.00%, and 85.70%, respectively.

Finally, the performances of the proposed algorithm are compared with other related published works and the comparison has been presented in Table 4. As listed in the table, the spectrogram audio feature has been used in [25] and [31] and the authors have achieved an accuracy of 71% in both the works. Table 4 also shows that the proposed system achieved an accuracy of 85% with the chromagram and this accuracy is higher than that achieved in [25], [31]. Not only that, this accuracy is even close to those of some other related works presented in [24], [27], [28], [30]. The comparison listed in Table 4 also shows that the chromagram is a very useful audio feature to detect voice pathology provided a suitable classifier like the CNN is used. Comparing the results of the proposed system with those of other multiple feature-based systems, it can be concluded that a single feature like chromagram is sufficient to detect pathological voices with a high accuracy and hence multiple acoustic features can be avoided to reduce the computation burden of the pathological voice detection system.

Table 1. Training and testing accuracies for the chromagram

Simulation No.	Accuracy (%)		
	Training	Validation	Testing
1	95.83	92.00	79.16
2	95.83	90.00	83.33
3	100.00	93.45	87.50
4	100.00	92.33	87.50
5	100.00	96.67	87.50
6	95.83	92.33	79.16
7	100.00	98.00	87.50
8	100.00	96.00	83.33
9	95.83	92.33	87.50
10	100.00	98.00	87.50
Average	98.33	94.11	84.99

Table 2. The classification matrix for the chromagram

	Prediction (%)	
Actual	Normal	Pathology
Normal	**80.00**	20.00
Pathology	10.00	**90.00**

Table 3. The performance measures for the chromagram

Performance Measures	Chromagram (%)
Accuracy	85.00
Precision	81.80
Recall/Sensitivity	90.00
F1 Score	85.70

Table 4. The comparison of the performances

Research works	Phonemes	Features	Tools	Accuracy
Wu et al. [31]	Vowels, Speech	Spectrogram	CNN, CDBN	71%
Fang et al. [24]	Vowels	Mel frequency cepstral coefficients (MFCC)	SVM, GMM, DNN	94.26% (male) 90.52% (female)
Jun and Kim [25]	General voice samples	Mel spectrogram	DNRNN	71%
Wang and Cheolwoo [27]	Sustained vowel '/a/'	Feature vectors	Hidden Markov model (HMM), GMM, and SVM	95.2%
Sellam, and Jagadeesan [28]	Tamil phrases	Feature vectors	SVM, and RBFNN	91% (RBFNN) 83% (SVM)
Sassou [30]	Japanese Vowel	Higher-order local auto-correlation (HLAC)	FFNN, and autoregressive hidden Markov model (AR-HMM)	87.75%
Proposed Method	Sustained Vowel '/a/'	Chromagram	CNN	85.00%

4.   CONCLUSION

This paper presents a CNN-based non-invasive pathological voice detection algorithm using chromagram feature of voice samples. The simulation results suggest that the chromagram is a useful audio feature that can be used for pathological voice detection algorithms, although chromagram only attracted the

attention of the researchers in music detection. Dysphonia causes a change in the pitch of the sound and hence affects the chromagram feature more compared to other features available in the literature. Hence, a higher accuracy is achieved with the chromagram compared to those of other unique features. Other performance parameters including accuracy, precision, recall, and F1 score, also confirm the effectiveness of the proposed algorithm. The performances of the proposed algorithm have been compared with those of other related works available in the literature. This paper also shows that a pathological voice detection system implemented with the chromagram voice feature and the CNN can outperform some other existing systems available in the literature. The proposed algorithm discriminates only the dysphonic voices from the normal voices. However, determining the progression level of voice pathology is another challenging task and this issue has not been addressed in this work. Other popular machine learning algorithms such as SVM, KNN, and GMM also need to be investigated in the future to ensure the comparative effectiveness of the proposed algorithm.

REFERENCES

[1] "Voice disorders," *Johns Hopkins Medicine*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/voice-disorders (accessed Mar. 10, 2022).

[2] D. Taib, M. Tarique, and R. Islam, "Voice feature analysis for early detection of voice disability in children," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 2018, pp. 12–17, doi: 10.1109/ISSPIT.2018.8642783.

[3] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66749–66776, 2020, doi: 10.1109/ACCESS.2020.2985280.

[4] "Voice disorders," *American Speech-Language-Hearing Association (ASHA)*. https://www.asha.org/practice-portal/clinical-topics/voice-disorders/ (accessed Mar. 10, 2022).

[5] R. J. Stachler *et al.*, "Clinical practice guideline: hoarseness (dysphonia) (update)," *Otolaryngology-Head and Neck Surgery*, vol. 158, Mar. 2018, doi: 10.1177/0194599817751030.

[6] R. H. G. Martins, H. A. do Amaral, E. L. M. Tavares, M. G. Martins, T. M. Gonçalves, and N. H. Dias, "Voice disorders: etiology and diagnosis," *Journal of Voice*, vol. 30, no. 6, Nov. 2016, doi: 10.1016/j.jvoice.2015.09.017.

[7] A. C. Davis, "Vocal cord dysfunction," *Johns Hopkins Medicine*. Accessed: Mar. 10, 2022. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/vocal-cord-dysfunction

[8] M. M. Johns, R. T. Sataloff, A. L. Merati, and C. A. Rosen, "Article commentary: shortfalls of the American academy of otolaryngology-head and neck surgery's clinical practice guideline: hoarseness (dysphonia)," *Otolaryngology-Head and Neck Surgery*, vol. 143, no. 2, pp. 175–177, Aug. 2010, doi: 10.1016/j.otohns.2010.05.026.

[9] S. R. Collins, "Direct and indirect laryngoscopy: equipment and techniques," *Respiratory Care*, vol. 59, no. 6, pp. 850–864, Jun. 2014, doi: 10.4187/respcare.03033.

[10] D. D. Mehta and R. E. Hillman, "Current role of stroboscopy in laryngeal imaging," *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 20, no. 6, pp. 429–436, Dec. 2012, doi: 10.1097/MOO.0b013e3283585f04.

[11] Y. D. Heman-Ackah, S. Mandel, R. Manon-Espaillat, M. M. Abaza, and R. T. Sataloff, "Laryngeal electromyography," *Otolaryngologic Clinics of North America*, vol. 40, no. 5, pp. 1003–1023, Oct. 2007, doi: 10.1016/j.otc.2007.05.007.

[12] A. J. Moshayedi, A. M. Agda, and M. Arabzadeh, "Designing and implementation a simple algorithm considering the maximum audio frequency of persian vocabulary in order to robot speech control based on arduino," in *Lecture Notes in Electrical Engineering*, Springer Singapore, 2019, pp. 331–346.

[13] R. Islam, E. Abdel-Raheem, and M. Tarique, "Early detection of COVID-19 patients using chromagram features of cough sound recordings with machine learning algorithms," in *2021 International Conference on Microelectronics (ICM)*, Dec. 2021, pp. 82–85, doi: 10.1109/ICM52667.2021.9664931.

[14] R. Islam, E. Abdel-Raheem, and M. Tarique, "A study of using cough sounds and deep neural networks for the early detection of Covid-19," *Biomedical Engineering Advances*, vol. 3, 100025, Jun. 2022, doi: 10.1016/j.bea.2022.100025.

[15] A. J. Moshayedi, M. S. Hosseini, and F. Rezaei, "WiFi-based massager device with NodeMCU through Arduino interpreter," *Journal of Simulation and Analysis of Novel Technologies in Mechanical Engineering*, vol. 11, no. 1, pp. 73–79, 2018.

[16] O. T. Esfahani and A. J. Moshayedi, "Design and development of Arduino healthcare tracker system for Alzheimer patients," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 5, no. 4, pp. 12–16, 2015.

[17] A. Al-Nasheri *et al.*, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018, doi: 10.1109/ACCESS.2017.2696056.

[18] R. Islam and M. Tarique, "Classifier based early detection of pathological voice," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 2019, pp. 1–6, doi: 10.1109/ISSPIT47144.2019.9001836.

[19] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018, doi: 10.1109/ACCESS.2018.2856238.

[20] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020, doi: 10.1109/ACCESS.2020.2986171.

[21] R. Islam, E. Abdel-Raheem, and M. Tarique, "A novel pathological voice identification technique through simulated cochlear implant processing systems," *Applied Sciences*, vol. 12, no. 5, 2398, Feb. 2022, doi: 10.3390/app12052398.

[22] 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 *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, Jul. 2017, pp. 1–4, doi: 10.1109/IWOBI.2017.7985525.

[23] P. Murphy, "Development of acoustic analysis techniques for use in the diagnosis of vocal pathology," *Ph.D. Dissertation, School of Physical Sciences, Dublin City University, Dublin*, 1997, [Online]. Available: https://doras.dcu.ie/19122/1/Peter_Murphy_20130620152522.pdf

[24] S.-H. Fang *et al.*, "Detection of pathological voice using cepstrum vectors: a deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, Sep. 2019, doi: 10.1016/j.jvoice.2018.02.003.

[25] T. J. Jun and D. Kim, "Pathological voice disorders classification from acoustic waveforms." 2018. [Online]. Available: https://mac.kaist.ac.kr/~juhan/gct634/2018/finals/pathological_voice_disorders_classification_from_acoustic_waveforms_report.p

df

[26] V. Srinivasan, V. Ramalingam, and P. Arulmozhi, "Artificial neural network based pathological voice classification using MFCC features," *International Journal of Science, Environment*, vol. 3, no. 1, pp. 291–302, 2014.

[27] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition methods," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2007, pp. 3253–3256, doi: 10.1109/IEMBS.2007.4353023.

[28] V. Sellam and J. Jagadeesan, "Classification of normal and pathological voice using SVM and RBFNN," *Journal of Signal and Information Processing*, vol. 05, no. 01, pp. 1–7, 2014, doi: 10.4236/jsip.2014.51001.

[29] M. Chopra, K. Khieu, and T. Liu, "Classification and recognition of stuttered speech." Stanford University, [Online]. Available: https://web.stanford.edu/class/cs224s/project/reports_2017/Manu_Chopra.pdf.

[30] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2017, pp. 1243–1247, doi: 10.1109/APSIPA.2017.8282229.

[31] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Interspeech 2018*, Sep. 2018, pp. 446–450, doi: 10.21437/Interspeech.2018-1351.

[32] M. A. Mohammed *et al.*, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences*, vol. 10, no. 11, 3723, May 2020, doi: 10.3390/app10113723.

[33] L. Verde, G. De Pietro, M. Alrashoud, A. Ghoneim, K. N. Al-Mutib, and G. Sannino, "Dysphonia detection index (DDI): a new multi-parametric marker to evaluate voice quality," *IEEE Access*, vol. 7, pp. 55689–55697, 2019, doi: 10.1109/ACCESS.2019.2913444.

[34] H. Ding, Z. Gu, P. Dai, Z. Zhou, L. Wang, and X. Wu, "Deep connected attention (DCA) ResNet for robust voice pathology detection and classification," *Biomedical Signal Processing and Control*, vol. 70, 102973, Sep. 2021, doi: 10.1016/j.bspc.2021.102973.

[35] M. Putzer and W. J. Barry, "Saarbrücken voice database." Accessed: Mar. 10, 2022. [Online]. Available: http://stimmdb.coli.uni-saarland.de/index.php4#target.

[36] D.L. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE, 2006, pp.1-44, doi: 10.1109/9780470043387.ch1.

[37] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Elsevier, 1992, pp. 429–446.

[38] P. Kim, "Convolutional neural network," in *MATLAB Deep Learning*, Berkeley, CA: Apress, 2017, pp. 121–147.

[39] R. M. Rangayyan, "Pattern classification and diagnostic decision," in *Biomedical Signal Analysis*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, pp. 571–632.

[40] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, Dec. 2016, doi: 10.1007/s40484-016-0081-2.

BIOGRAPHIES OF AUTHORS

Rumana Islam (iD) (g) [SC] (P) completed her Master of Science (MSc) in Biomedical Engineering from Wayne State University, Michigan in 2004. Currently, she is pursuing her Doctor of Philosophy (Ph.D.) degree in the Department of Electrical and Computer Engineering, University of Windsor, Canada. She has professional experience in academy and industry as well. She served as a professional Engineer in SIEMENS and Bangladesh Telegraph and Telephone Board (BTTB) in Bangladesh. She served as an Adjunct Faculty member at East West University of Bangladesh. She worked as a Lecturer in the Department of Electrical and Electronic Engineering, American International University, Bangladesh (AIUB) from 2008 to 2011. Currently, she is working as an Adjunct Faculty member in the Department of Electrical and Computer Engineering at the University of Science and Technology of Fujairah (USTF), UAE. She has published several research papers in highly reputed journals and conference proceedings. Her research interests include biomedical engineering, biomedical signal processing, biomedical diagnostic tools using machine learning, digital signal processing, and biomedical image processing. She can be contacted at email: islamq@uwindsor.ca.

Mohammed Tarique (iD) (g) [SC] (P) completed his Doctor of Philosophy (Ph.D.) degree in Electrical Engineering from the University of Windsor, Canada in 2007. He had worked as an Assistant Professor in the Department of Electrical Engineering, American International University-Bangladesh from 2007-2011. He joined Ajman University in 2011. Currently, he is working as an Associate Professor of Electrical Engineering at the University of Science and Technology of Fujairah (USTF). His research interests are in wireless communication, mobile ad hoc networks, digital communication, digital signal processing, and biomedical signal processing. He has published 40 research papers in the top-ranked journals in the above-mentioned fields. He also published 15 research papers in the proceedings of highly reputed international conferences. He also had worked as a Senior Engineer in Beximco Pharmaceuticals Limited, Bangladesh from 1993 to 1999. He also completed Master of Business Administration (MBA) from the Institute of Business Administration (IBA), University of Dhaka, Bangladesh. He can be contacted at email: m.tarique@ustf.ac.ae.