IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE MULTIMEDIA FOR HEALTHCARE

Received June 2, 2018, accepted July 9, 2018, date of publication July 16, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2856238

Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework

MUSAED ALHUSSEIN AND GHULAM MUHAMMAD
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, King Saud University, Riyadh, Saudi Arabia, through the Research Group under Project RG-1436-016.

ABSTRACT The feasibility and popularity of mobile healthcare are currently increasing. The advancement of modern technologies, such as wireless communication, data processing, the Internet of Things, cloud, and edge computing, makes mobile healthcare simpler than before. In addition, the deep learning approach brings a revolution in the machine learning domain. In this paper, we investigate a voice pathology detection system using deep learning on the mobile healthcare framework. A mobile multimedia healthcare framework is also designed. In the voice pathology detection system, voices are captured using smart mobile devices. Voice signals are processed before being fed to a convolutional neural network (CNN). We use a transfer learning technique to use the existing robust CNN models. In particular, the VGG-16 and CaffeNet models are investigated in the paper. The Saarbrucken voice disorder database is used in the experiments. Experimental results show that the voice pathology detection accuracy reaches up to 97.5% using the transfer learning of CNN models.

INDEX TERMS Mobile multimedia healthcare, voice pathology detection, deep learning, Saarbrucken voice database.

I. INTRODUCTION

The healthcare industry is not only for making money but also for providing basic to high-end health services to people. Remote healthcare or tele-healthcare is needed nowadays. The need arises because of several reasons: (i) specialist doctors are scarce, (ii) commuting to remote areas is sometimes difficult, (iii) peak-hour traffic jam in the urban area may prohibit going to the hospital, (iv) patients are unwilling to visit the doctors for the follow-up, and (v) patients have busy schedule. Therefore, research on remote healthcare or mobile healthcare has increased in recent years.

A mobile healthcare framework needs several components, such as sensors that can collect data from patients, portable processing units, short- and long-range wireless communication, and edge and/or cloud servers. In addition to these components, registered doctors and caregivers along with dedicated vehicles are necessary to speed up the service. The immense development in wireless communication technologies and computing processing power has enabled mobile healthcare to provide fast, low-cost, comfortable, and hassle-free services. The mobile healthcare industry is projected to have revenue of over billions of US dollars in the next couple of years [1]. However, mobile healthcare is yet to obtain wide patient acceptance due to trust, privacy, and security issues. Nevertheless, sophisticated and outstanding developments in technologies related to accuracy, privacy, and security increase its patient acceptance [2], [3].

Many telemedicine or mobile healthcare frameworks have been proposed in literature. Some of them target the infrastructure improvement of the framework [4], [5], others focus on edge and cloud computing [6], [7], and the rest focuses on computing and accuracy [8]–[10]. The infrastructure is mainly related to the arrangement of the Internet of Things (IoT), software defined networks, and the use of radio and wide access networks. Cloud and edge computing deal with access to servers, authentication and security, data processing, seamless distribution of jobs, coordination between clients and stakeholders, and processing acceleration by arranging edge caches before transmitting signals to cloud servers.

This study aims to improve the accuracy of the healthcare system by using machine learning techniques. Accurate processing and classification of signals are important tasks. Conventionally, many patients do not trust the outcome of an automated system unless it is interpreted by a specialist physician, which we also agree.

FIGURE 1. Illustration of a smart healthcare system [26].

Medical assessment has two types: manual and automatic. Each of these assessments has pros and cons. The manual assessment needs (i) the physical presence of the doctor, (ii) invasive sensors, such as endoscopy, and (iii) costly medical devices. The assessment outcome varies depending on the judgment and experience of the doctor. The automatic assessment needs (i) a good algorithm that can correctly identify the nature of the signal, (ii) high-speed processing, and (iii) a small number of transmission errors. The outcome of the assessment must be screened by an experienced doctor for the final judgment. A total of 85% of aged US citizens prefer to have treatment at home if adequate facility is available [25]. Smart cities are being developed these days to facilitate the healthcare needs of citizens and residents. Figure 1 shows an infrastructure of a smart city [26].

Among many healthcare facilities, we choose voice pathology detection in this study. Around 7.6% of adult people in the United States suffer from voice pathology [11]. Spasmodic dysphonia, which is caused by an involuntary movement of muscles in the larynx, is a common voice pathology. This type of voice pathology is common among people who use their voice excessively, such as singers, teachers, and lawyers [12]. Voice pathology is more common in women than in men [11].

The treatment of voice pathology needs to start as soon as possible after its detection; otherwise; the voice problem will be permanent. Automatic voice pathology detection is a non-invasive method, in which the detection is done using the voice signals of the patients only. The voice signals can be captured by a microphone, a smart phone, or any voice recorder. A sustained voice of the sound /a/ is normally used to evaluate voice pathology because it has high amplitude and is comfortable for the patients to utter [13].

First, a mobile healthcare framework is designed in this study. Then, a voice pathology detection system is introduced in the framework. The system uses the conventional neural network (CNN) model as the state-of-the-art technique. Several CNN models are investigated using transfer learning. This study contributes to literature by (i) integrating the CNN models in the voice detection system, (ii) using the system in a mobile healthcare framework, and (iii) utilizing cepstrum derivatives before the CNN models.

The rest of the paper is organized as follows. Section II provides a brief literature review on voice pathology detection system. Section III describes the framework and the proposed system. Section IV discusses the experimental results. Finally, Section V elaborates the conclusions of the study.

II. SELECTED WORK ON VOICE PATHOLOGY DETECTION

Voice pathology detection has been rarely investigated. In the late 1960s, voice quality was measured by shimmer, jitter, and harmonic to noise ratio [14]. These measurements were initially derived to assess voice quality during transmission. The features of voice pathology can be classified into three types: imported from the speech recognition applications [15], [16], solely dedicated for the voice pathology detection [17], [18], and a combination between the previous two [19], [20].

The speech features that are normally used in voice pathology detection include Mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and pitch frequency. The MFCC can simulate the hearing mechanism of humans, whereas the LPCC can simulate the voice/speech production mechanism of humans.

The specific features that are dedicated to voice pathology detection include shimmer, jitter, harmonic-to-noise ratio, glottal noise ratio, and vocal tract tube fluctuation [21]. Some of these measurements need a good estimation of the

pitch period; however, finding the pitch period is a challenging task. These features alone also cannot detect the voice pathology in mild condition [20].

Some features are chosen from other domains, such as multimedia indexing and image processing. These features include MPEG-7 audio features [20], interlaced derivative patterns (IDP) [13], and co-occurrence matrix. The MPEG-7 features were initially developed for multimedia indexing, and they were then used in other applications, such as environment recognition and voice pathology detection. The IDP features were first proposed for face recognition applications, and they were then incorporated in facial emotion recognition, speech recognition, and voice pathology detection. The co-occurrence matrix is a popular image texture classification matrix, which was later used in many different applications.

A research community needs a good database for analysis. Among several voice pathology databases, the Massachusetts Eye and Ear Infirmary (MEEI) database [22] is the most common; however, it is a commercial database and suffers from some limitations [16]. The limitations include different recording environments of normal and pathological voices, various sampling frequencies of the voice signals, and an imbalanced number of samples between normal and pathological voices. The second database is the Saarbrucken voice disorder (SVD) database, which is publicly available via the Internet [23]. The database contains not only voice samples but also electroglottographic (EGG) signals. The signals contain the information of the glottis movement during voice phonation. Another database is the Arabic voice pathology database (AVPD), which was recently developed at King Saud University, Riyadh [24]. The database contains samples of sustained vowels, words, and paragraphs. All the speakers were native to Arabic language.

An IDP-based voice pathology detection system was proposed in [13]. The system achieved 99.4%, 93.2%, and 91.5% accuracies in the MEEI, SVD, and AVPD databases, respectively. The system was also evaluated in cross-database cases, in which accuracies between 78% and 88% were obtained using different combination of databases in the training and testing.

A comparison of voice pathology detection accuracies using sustained vowels and running speech was made in [15]. The signal features were obtained by analyzing the spectrum of the signal. The results showed that the accuracy using the sustained vowel was better than that using the running speech. Notably, the running speech is more commonly used than the sustained vowel.

Godino-Llorente *et al*. used Gaussian mixture models and MFCC features to detect voice pathology [16]. Their system achieved 0.988 area under curve using the MEEI database. Glottal noise measures were used to detect voice pathology in [17]. These measures were the improvement of the typical harmonic-to-noise ratio and suitable for the voice assessment.

The co-occurrence matrix-based system was proposed in [18]. In the system, the inputs were voice and EGG signals. GMMs were used as the classifier. The classifier probabilities from the two types of inputs were fused by a Bayesian sum rule. Using the SVD database, the system achieved 99.87% accuracy. Different fusion strategies were investigated for voice pathology detection in [28]. Specifically, score- and audio-level fusions were compared. The audio-level fusion produced better result than the score-level fusion. In the SVD database, the accuracy was around 81%. A personalized frequency estimation-based voice pathology classification was proposed in [34].

A voice pathology detection system based on a vocal tract irregularity measure was proposed in [21]. A support vector machine (SVM)-based classifier was used in the system. The system gained 99.22% and 94.7% accuracies with the MEEI and SVD databases, respectively. A preliminary study of voice pathology detection using a deep learning model was conducted in [29]. A CNN model combined with a long short-term memory network was used. Using the SVD database, the authors obtained 68.08% accuracy. Correlation features between different frequency regions were investigated in [31]. The results showed that the accuracy using several band-limited signals was better than the accuracy using the full-band signal. Another deep learning-based system was developed in [33], in which the accuracy was 99.32% using the MEEI database. In this system, the input was the cepstrum vector of the voice signal.

The voice pathology detection system in a telehealthcare framework was used in some works. In [30], Muhammad *et al.* integrated the IoT and cloud computing for a voice pathology detection framework. A local binary pattern for feature extraction and an extreme learning machine for classification were used in the study. An accuracy of 98.1% was achieved using the SVD database. Local features together with GMM-based classifiers were used in a smart city paradigm in [32]. Voice and EGG signals were the inputs. An accuracy of 94.2% was achieved using the SVD database. Another disease prediction system based on the IoT and cloud with fuzzy logic was proposed in [37]; however, this study was solely for voice pathology detection.

III. MATERIAL AND METHOD

A. FRAMEWORK

Figure 2 illustrates the proposed mobile multimedia healthcare framework. The framework consists of several components. Mobile smart sensors are present to capture signals from the patients. These sensors are accessible via short-range wireless networks or Bluetooth. The captured signals can be transmitted to the local machine, which then transmit these signals to the cloud. The cloud has a cloud manager, which manages the data flow in and out of the cloud; and an authentication and privacy manager, which authenticates the signals that are coming from the registered users and those that are going to the registered doctors and caregivers. Servers are also present in the cloud for deep learning processing. A storage for storing all the medical records is present as well. The communication between the local machine and the cloud is done by a radio access network. An optimal scheduling

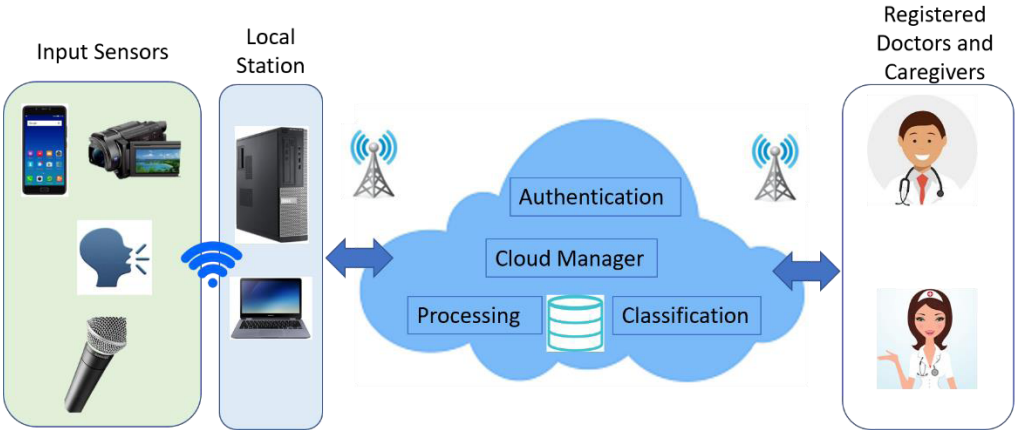

FIGURE 2. Mobile healthcare framework.

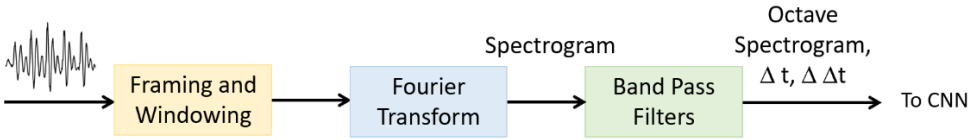

FIGURE 3. Voice signal processing in the proposed system.

can be embedded to assign the servers appropriately and seamlessly.

Once the detection is performed in the cloud server, the cloud manager sends the decision and some samples of the voice signal to a registered doctor. The doctor double checks the decision by analyzing the voice signals and provides his or her feedback to the cloud. The cloud manager then notifies the patient about the decision and informs a registered caregiver for the care of the patient if needed. The entire framework is non-invasive in nature and is thus comfortable for the patient.

B. PROPOSED SYSTEM

Figure 3 shows the proposed voice pathology detection system using the deep learning approach. The input to the system is the voice signal from the patient, and the output is a decision whether the patient has normal or pathological voice.

The voice signal is 1 s long. If the input is more than 1 s, then it is cut from the middle to make it a signal of 1 s long. The signal is divided into overlapping frames of 40 ms, in which the overlapping is 20 ms. The frame length of 40 ms is a good balance between capturing the pitch periods and smoothing out the voice breaks. If the length is very long, then the voice breaks or the noises that cause irregular opening and closing of the vocal folds fade away. If the frame length is short, then the sustained effect and the pitch period are lost.

A fast Fourier transform is applied to the framed signal to convert the signal into a frequency-domain signal. After concatenating all the frequency-domain representations of the frames, we obtain a spectrogram. The spectrogram can be treated as an image. A total of 20 bandpass filters are applied to the spectrogram. The filters are octave-scaled centered. The octave scale usually performs better than the Mel-scale in voice pathology detection [31]. First- and second-order time derivatives are applied to the output of the octave spectrogram. After this operation, we obtain three image-like patterns, namely, octave spectrogram and its first- and second-order derivatives. The three image-like patterns are the input of the CNN models.

In the proposed system, we investigated two CNN models: the VGG16 Net [36], [38] and the CaffeNet [39].

The VGG16 Net is a very deep CNN model, which contains five blocks of convolution (see Figure 4). In the first block, there are two convolutional layers of size 3×3 and the number of filters per layer is 64. The second block has also two layers (128 filters per layer). The next three blocks have three layers each of size 3×3. The number of filters per layer are 256, 512, and 512, respectively. There are pooling layers after each convolution block. The last block is followed by two layers of fully-connected neural networks, and a softmax layer.

The architecture of the CaffeNet is shown in Figure 5. There are five convolutional layers and three max pooling layers. The first two and the fifth convolutional layers are followed by the max pooling layers. In the figure, the size of the convolution filters and the number of filters are shown. For example, the first convolution layer filters are of size 11×11, and the number of filters is 96. After the last max pooling layer, there are two fully-connected neural networks and a softmax layer.

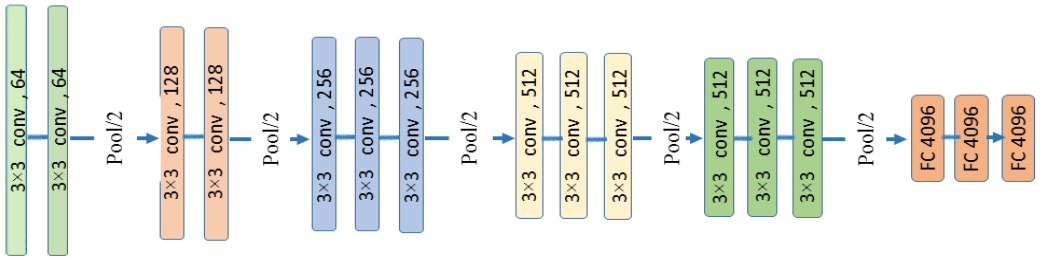

FIGURE 4. Architecture of VGG16 Net.

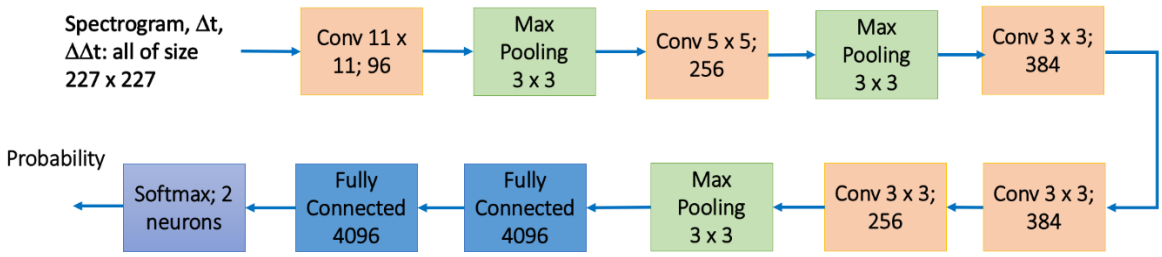

FIGURE 5. Architecture of CaffeNet.

In the proposed voice pathology detection system, the octave spectrogram and its first- and second-order time derivatives are the input to the VGG16 Net or the CaffeNet. The inputs are resized to fit the corresponding models. In this case, we resized the inputs to 227×227. Both the VGG16 Net and the CaffeNet are trained with a large number of images, and hence these models are robust to many applications. The number of samples in the voice pathology databases is very small, and, therefore, we cannot use these models for the training from scratch. Rather, we use the transfer learning and the fine-tune approaches to get the benefit of these robust models [40], [41]. Many applications, especially where the number of samples is limited, have successfully adopted the transfer learning and the fine tuning.

In the proposed system, we replace the final softmax layer of the original CNN models by another softmax layer of two neurons because we have two classes: normal voice and pathological voice. The initial parameters of the models are unchanged except the softmax layer, where random weights are assigned. After we replace the softmax layer, we fine-tune the model parameters using the training samples of the voice pathology database.

The learning rate of the new softmax layer is set to a higher value than that of other layers. This is because the weights of other layers are already pre-trained and the model changes slowly with the new training data. On the other hand, the new softmax layer needs to learn fast because the weights are assigned randomly. The step size to change the learning rate is set to a small value because the weights are already optimized for a large number of data and needs only a small change with the new data.

The weights of the softmax layer were initialized using random numbers with zero mean Gaussian distribution and the standard deviation of 0.01. The training parameters of the model were set as follows: learning rate $= 0.001$, momentum $= 0.9$, weight decay $= 0.0005$. The weights were optimized by using a stochastic gradient decent algorithm with a batch size of 40 samples. 50% dropout was applied in the fully-connected layers.

Once the CNN model was fine-tuned, we removed the softmax layer. The last fully-connected layer before the softmax layer was fed into an SVM classifier. The SVM is a powerful binary classifier and is proved to be efficient in many applications such as image classification, object recognition, speech recognition, speaker recognition, and environment recognition [42]. In the SVM classification, a kernel function projects the input data space into a high dimensional space so that a hyperplane can separate the samples of two classes. The objective of the SVM is to find an optimal hyperplane that can maximize the separation between support vectors of two classes. There are two parameters in the SVM: kernel function parameter (in our case, we used the radial basis function for its good generalization capability) and the optimization parameter. We tried different values of these parameters using an extensive grid search, and finally settled to kernel parameter $= 0.1$ and optimization parameter $= 0.09$ because they gave the best results.

C. DATABASE

There are several databases to do research on voice pathology detection. In our experiments, we used the SVD database [23] and the MEEI database [22]. The SVD database is a publicly

available database and it contains voice samples of three sustained vowels /a/, /i/, /u/ at four different pitch intonations: high, mid, low, and normal. We used /a/ uttered at normal pitch condition. The MEEI database contains sustained /a/ voice samples. There are two different sampling frequencies; we downsampled 50 kHz to 25 kHz to have all the samples the same sample frequency. We chose three voice pathologies which are common to these databases. The pathologies are vocal fold cyst, vocal fold polyp, and vocal fold (unilateral) paralysis. In addition to these, we also used the normal voice samples. The number of samples in each database is shown in Table 1.

TABLE 1. Number of samples in the databases. (M = male, F = female).

Database	Cyst		Polyp		Paralysis		Normal	
	M	F	M	F	M	F	M	F
MEEI	6	4	8	7	38	32	21	32
SVD	5	1	19	25	121	73	137	125

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Several experiments were conducted using the SVD database and the MEEI database. The experiment cases were as follows:

(i) Training with the MEEI database and testing with the SVD database.

(ii) Training with SVD database and testing with the MEEI database.

(iii) Training and testing with the SVD database. In this case, only files with sustained vowel /a/ produced at normal and people older than 15 are used. 1616 (743 males and 873 females) samples belong to pathological speakers and 686 (259 males and 427 females) to normal speakers. We divided the dataset into two subsets. In one experiment, we used one subset for the training and the other for the testing. In another experiment, the training and the testing subsets were interchanged. Finally, we averaged the results of these two experiments.

TABLE 2. Performance of the proposed system using the VGG16 net.

Experiment case	% Accuracy	% Sensitivity	% Specificity
(i) Training: MEEI; Testing: SVD	73.3	75.4	71.7
(i) Training: SVD; Testing: MEEI	94.5	95.4	93.2
(i) Training: SVD; Testing: SVD	93.5	94.8	92.4

Table 2 shows the accuracy, the sensitivity, and the specificity of the proposed system using the VGG16 Net and the SVM. In the case (i) experiment (training with the MEEI and the testing with the SVD), the system achieved 73.3% accuracy, 75.4% sensitivity, and 71.7% specificity. In the case (ii) experiment (training with the SVD and the testing with the MEEI), the system obtained 94.5% accuracy, 95.4% sensitivity, and 93.2% specificity. Finally, in the case (iii) experiment (training and testing with the SVD), the system got 93.5% accuracy, 94.8% sensitivity, and 92.4% specificity. From the experimental results we observe that the system did not achieve good accuracy in the case (i) experiment. As the MEEI database has several issues (mentioned earlier), the system was not trained well with the MEEI database.

TABLE 3. Performance of the proposed system using the CaffeNet.

Experiment case	% Accuracy	% Sensitivity	% Specificity
(i) Training: MEEI; Testing: SVD	75.1	76.4	78.7
(i) Training: SVD; Testing: MEEI	94.1	95.7	93.8
(i) Training: SVD; Testing: SVD	93.9	94.8	92.8

Table 3 shows the accuracy, the sensitivity, and the specificity of the proposed system using the CaffeNet and the SVM. Similar trend as in Table 2 is observed in the results of Table 3. The accuracies in the case (i), the case (ii), and the case (iii) experiments were 75.1%, 94.1%, and 93.9%, respectively. If we compare the results between Table 2 and Table 3, we find that the system achieved slightly better results with the CaffeNet than with the VGG16 Net. As the CaffeNet is shallower than the VGG16 Net and the voice samples are not too large, the CaffeNet fits better than the VGG16 Net into the system.

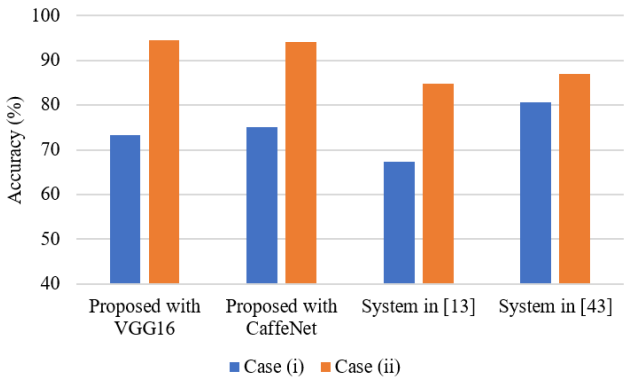

FIGURE 6. Comparison of accuracies obtained by different systems in case (i) and case (ii) experiments.

Figure 6 shows a comparison of accuracies between different available systems in the case (i) and the case (ii) experiments. Two systems in [13] and [43] were considered because these two systems used the same voice samples as with the proposed system. The results of these two systems were taken from the corresponding papers. From the results we can see

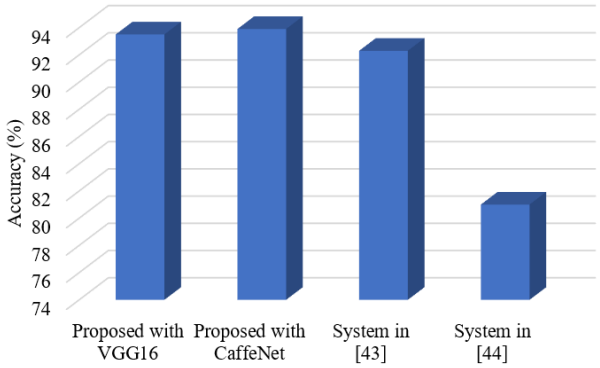

FIGURE 7. Comparison of accuracies obtained by different systems in case (iii) experiment.

that the proposed system either with the VGG16 Net or with the CaffeNet outperformed the other two systems in both the cases.

Figure 7 shows an accuracy comparison between the systems in the case (iii) experiment. From the figure we find the proposed system achieved better results than other systems. The accuracies of the other systems were obtained from the respective papers. It can be noted that the systems in [13], [43], and [44] did not use the CNN models in their systems. Therefore, we can state that the use of the CNN model can significantly enhance the performance of the voice pathology detection system. We did not perform any voice pathology classification because the number of samples per pathology class were too small to fine-tune the CNN models.

We used an Intel quad-core machine having 16 GB RAM and NVIDIA 8 GB GPU. The fine-tuning of the CNN models took approximately one hour, and the detection of a voice sample during testing took an average of 1.5 seconds. This timing is acceptable for a real-time voice pathology detection in a mobile multimedia healthcare framework.

V. CONCLUSION

A mobile multimedia healthcare framework was designed. The proposed voice pathology detection system was embedded to the framework to constantly assess the voice condition of a patient. A deep learning in the form of the CNN models was used in the system. Several popular CNN models were investigated. In the experiments on the SVD database, the system achieved 98.77% accuracy with the CaffeNet CNN model followed by the SVM classifier. The result is promising because it outperforms some of the previous results reported in literature. We will perform cross-database experiments in a future study.

The possible directions of future research are as follows. The voice signal can be divided into several band-limited signals, and parallel CNN models can be applied to these band-limited signals. Then, a fusion strategy can be utilized to fuse the deep-learned features from the CNN models. Another direction is the use of different types of inputs, such as voice and EGG signals, combined by deep fusion strategy.

REFERENCES

[1] Orbis Research, Reuters. *mHealth Market Worth $23 Billion in 2017 and Estimated to Grow at a CAGR of more than 35% Over the Next Three Years*. Accessed: May 10, 2018. [Online]. Available: https://www.reuters.com/brandfeatures/venture-capital/article?id=4640
[2] Z. Ali *et al.*, "Edge-centric multimodal authentication system using encrypted biometric templates," *Future Gener. Comput. Syst.*, vol. 85, pp. 76–87, Aug. 2018.
[3] Z. Ali, M. S. Hossain, G. Muhammad, and M. Aslam, "New zero-watermarking algorithm using hurst exponent for protection of privacy in telemedicine," *IEEE Access*, vol. 6, pp. 7930–7940, Dec. 2018.
[4] M. Chen, J. Yang, Y. Hao, S. Mao, and K. Hwang, "A 5G cognitive system for healthcare," *Big Data Cogn. Comput.*, vol. 1, no. 1, pp. 1–5, 2017.
[5] V. Foteinos, D. Kelaidonis, G. Poulios, P. Vlacheas, V. Stavroulaki, and P. Demestichas, "Cognitive management for the Internet of Things: A framework for enabling autonomous applications," *IEEE Veh. Technol. Mag.*, vol. 8, no. 4, pp. 90–99, Dec. 2013.
[6] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 60–65, Apr. 2018.
[7] L. Hu, M. Qiu, J. Song, M. S. Hossain, and A. Ghoneim, "Software defined healthcare networks," *IEEE Wireless Commun.*, vol. 22, no. 6, pp. 67–75, Dec. 2015.
[8] P. T. Kim and R. A. Falcone, Jr., "The use of telemedicine in the care of the pediatric trauma patient," *Seminars Pediatric Surg.*, vol. 26, no. 1, pp. 47–53, 2017.
[9] M. S. Hossain, "Cloud-supported cyber–physical localization framework for patients monitoring," *IEEE Syst. J.*, vol. 11, no. 1, pp. 118–127, Mar. 2017, doi: 10.1109/JSYST.2015.2470644.
[10] M. S. Hossain and G. Muhammad, "An emotion recognition system for mobile applications," *IEEE Access*, vol. 5, pp. 2281–2287, 2017.
[11] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, 2014.
[12] J. M. Schweinfurth, M. Billante, and M. S. Courey, "Risk factors and demographics in patients with spasmodic dysphonia," *Laryngoscope*, vol. 112, no. 2, pp. 220–223, 2012.
[13] G. Muhammad *et al.*, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomed. Signal Process. Control*, vol. 31, pp. 156–164, Jan. 2017.
[14] M. H. L. Hecker and E. J. Kreul, "Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1275–1282, Apr. 1971.
[15] R. Fraile, J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, and J. M. Gutierrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech," *J. Voice*, vol. 27, no. 1, pp. 11–23, 2013.
[16] J. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.
[17] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Lang., Hearing Res.*, vol. 43, no. 2, pp. 469–485, 2000.
[18] G. Muhammad, M. F. Alhamid, M. S. Hossain, A. S. Almogren, and A. Vasilakos, "Enhanced living by assessing voice pathology using a co-occurrence matrix," *Sensors*, vol. 17, no. 2, p. 267, Jan. 2017.
[19] 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, p. 20, 2016.
[20] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomed. Signal Process. Controls*, vol. 11, pp. 1–9, May 2014.
[21] G. Muhammad *et al.*, "Automatic voice pathology detection and classification using vocal tract area irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016.
[22] *Massachusetts Eye and Ear Infirmary, Elemetrics Disordered Voice Database (Version 1.03)*, Voice Speech Lab., Boston, MA, USA, 1994. [Online]. Available: http://www.kayelemetrics.com/
[23] W. J. Barry and M. Putzer. *Saarbrucken Voice Database*. Accessed: May 10, 2018. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/
[24] T. Mesallam *et al.*, "Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *J. Healthcare Eng.*, vol. 2017, Oct. 2017, Art. no. 8783751.

[25] V. McKelvey. (2010). *Spending More on in-Home Care*. Accessed: Mar. 1, 2018. [Online]. Available: http://www.aarp.org/relationships/caregiving/info-01-2010/spendingmore-on-in-home-care.html
[26] *Realtime Technology and The Healthcare Internet of Things*. Accessed: May 2018. [Online]. Available: http://www.pinsdaddy.com/
[27] M. S. Hossain and G. Muhammad, "Healthcare big data voice pathology assessment framework," *IEEE Access*, vol. 4, pp. 7806–7815, Dec. 2016.
[28] D. Martinez, E. Lleida, A. Ortega, and A. Miguel, "Score level versus audio level fusion for voice pathology detection on the Saarbrücken voice database," in *Proc. IberSPEECH Conf.*, Madrid, Spain, Nov. 2012, pp. 110–120.
[29] 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.
[30] G. Muhammad, S. K. 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.
[31] 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, 2017.
[32] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Syst.*, pp. 1–11, Jul. 2018, doi: 10.1007/s00530-017-0561-x.
[33] S.-H. Fang *et al.*, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, pp. 1–8, 2018, doi: 10.1016/j.jvoice.2018.02.003.
[34] L. Verde, G. De Pietro, and G. Sannino, "A methodology for voice classification based on the personalized fundamental frequency estimation," *Biomed. Signal Process. Control*, vol. 42, pp. 134–144, Apr. 2018.
[35] M. S. Hossain and G. Muhammad, "Audio-visual emotion recognition using multi-directional regression and Ridgelet transform," *J. Multimodal User Interfaces*, vol. 10, no. 4, pp. 325–333, 2016.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[37] P. M. Kumar, S. Lokesh, R. Varatharajan, G. C. Babu, and P. Parthasarathy, "Cloud and IoT based disease prediction and diagnosis system for healthcare using fuzzy neural classifier," *Future Gener. Comput. Syst.*, vol. 86, pp. 527–534, Sep. 2018.
[38] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[39] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Conf. Multimedia (MM)*, 2014, pp. 675–678.
[40] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Giro-i-Nieto, "Cultural event recognition with visual convnets and temporal models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–44.
[41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
[42] S. Abe, *Support Vector Machines for Pattern Classification*. Berlin, Germany: Springer-Verlag, 2005.
[43] 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, Dec. 2018.
[44] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit," in *Advances in Speech and Language Technologies for Iberian Languages*. Berlin, Germany: Springer, 2012, pp. 99–109.

MUSAED ALHUSSEIN 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 on the Faculty of the Computer Engineering Department, College of Computer and Information Science, King Saud University. He is currently the Founder and Director of the Embedded Computing and Signal Processing Research Laboratory. His research activity is focused on typical topics of 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, eHealthcare, and body area networks.

GHULAM MUHAMMAD received the B.S. degree in computer science and engineering from the Bangladesh University of Engineering and Technology in 1997, and the M.S. degree and the Ph.D. degree in electrical and computer 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 has authored or co-authored over 200 publications including IEEE/ACM/Springer/Elsevier journals, and flagship conference papers. He has a U.S. patent on audio processing. He supervised over 10 Ph.D. and Master Theses. He is involved in many research projects as a principal investigator and a co-principal investigator. He was a recipient of the Japan Society for Promotion and Science Fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan. His research interests include image and speech processing, cloud and multimedia for healthcare, serious games, resource provisioning for big data processing on media clouds and biologically inspired approach for multimedia and software system. He received the Best Faculty Award of Computer Engineering Department, KSU, from 2014 to 2015.