Article

Voice Pathology Detection and Classification Using Convolutional Neural Network Model

Mazin Abed Mohammed [1],*, Karrar Hameed Abdulkareem [2], Salama A. Mostafa [3],
Mohd Khanapi Abd Ghani [4], Mashael S. Maashi [5], Begonya Garcia-Zapirain [6], Ibon Oleagordia [6],
Hosam Alhakami [7] and Fahad Taha AL-Dhief [8]

[1] College of Computer Science and Information Technology, University of Anbar, 11, Ramadi, 55, Anbar, Iraq
[2] College of Agriculture, Al-Muthanna University, Samawah 66001, Iraq; khak9784@mu.edu.iq
[3] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia,
   Batu Pahat 86400, Johor, Malaysia; salama@uthm.edu.my
[4] Biomedical Computing and Engineering Technologies (BIOCORE) Applied Research Group,
   Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
   Durian Tunggal 76100, Melaka, Malaysia; khanapi@utem.edu.my
[5] Software Engineering Department, College of Computer and Information Sciences, King Saud University,
   Riyadh 11451, Saudi Arabia; mmaashi@ksu.edu.sa
[6] eVIDA Lab. University of Deusto, Avda/Universidades 24, 48007 Bilbao, Spain;
   mbgarciazapi@deusto.es (B.G.-Z.); ibruiz@deusto.es (I.O.)
[7] Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University,
   Makkah, Saudi Arabia; hhhakam@uqu.edu.sa;
[8] Faculty of Engineering, School of Electrical Engineering, Universiti Teknologi Malaysia (UTM),
   Johor Bahru 81310, Malaysia; taha-1989@graduate.utm.my
* Correspondence: mazinalshujeary@uoanbar.edu.iq

Abstract: Voice pathology disorders can be effectively detected using computer-aided voice pathology classification tools. These tools can diagnose voice pathologies at an early stage and offering appropriate treatment. This study aims to develop a powerful feature extraction voice pathology detection tool based on Deep Learning. In this paper, a pre-trained Convolutional Neural Network (CNN) was applied to a dataset of voice pathology to maximize the classification accuracy. This study also proposes a distinguished training method combined with various training strategies in order to generalize the application of the proposed system on a wide range of problems related to voice disorders. The proposed system has tested using a voice database, namely the Saarbrücken voice database (SVD). The experimental results show the proposed CNN method for speech pathology detection achieves accuracy up to 95.41%. It also obtains 94.22% and 96.13% for F1-Score and Recall. The proposed system shows a high capability of the real-clinical application that offering a fast-automatic diagnosis and treatment solutions within 3 s to achieve the classification accuracy.

1. Introduction

There are many factors that can caused by voice pathologies. Some of include infections of voice tissue, tiredness, environmental changes, muscular dystrophy, face soreness and others [1]. The voice pathology has a negative impact on vibration regularity and voice functionality, which leads to an increase in vocal noise. The normal voice turned to be tense, weak and hoarse [2] that affects the quality of voice [3]. To date, the current vocal pathology detection methods have a biased evaluation

based on subjective matters [4]. An example of the subjective evaluation is auditory-perceptual assessment in hospitals, which is widely applied by visual laryngostroboscopy assessment [5]. Several clinical examinations are applied for auditory-perceptual parameters to scale the rate of severity diagnosis [6]. However, those evaluation methods are subject to parameter sensitivity and are also time consuming and laborious [7]. In addition, these methods require a physical patient examination in the clinic which could be difficult for patients with severe conditions. An example of objective evaluation is using a computer-aided tool to identify and analyse vocal signals without any surgical intervention. The automatic detection can also recognize inaudible sounds [1].

These evaluation methods are not subjective as they do not depend on a human decision. Besides, they are easy to apply since the voice recordings can be made available remotely via different internet recoding applications. Therefore, some studies such as [8] have developed vocal processing methods to determine the vocal pathology aspects can be effectively combined with a machine learning method to detect the voice pathology automatically in one framework to accurately distinguish healthy people from people with voice pathologies. In the literature, various voice pathology databases have been widely applied for the objective evaluation of voice pathology. The most common voice pathology databases are the Saarbruecken Voice Database (SVD) [9], Arabic Voice Pathology Database (AVPD) [10] and the Massachusetts Eye and Ear Infirmary Database (MEEI) [11]. The vocalization of the vowel /a/. [1,11] is available in many language databases [2], therefore, it is commonly analysed by researchers. Other combinations of vowels are also analyzed by researchers [1,12]. Notably, the majority of the researchers in the voice pathologies community have limited the datasets to specific pathologies sets [12].

Usually, the clinical interpretation of vocal features is conducted before the process of pathology detection [13]. Examples of vocal features are glottal-to-noise excitation ratio (GNE) [14], Mel frequency cepstral coefficients (MFCC) [15], multidimensional voice program parameters (MDVP) [16] and many others. See [17] for more details about the settings of speech pathologies. Once the vocal features are extracted, many conventional classification methods are applied for voice pathology detection. For detection purposes, most studies have used Random Forests (RF), Artificial Neural Networks (ANN) [18], Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and other classifiers [19,20]. It is observed that the study results show notable differences. Because of the different set selection, a sample of voice pathology, a vocal feature and classification method are applied in the experiments. This drives us to the following conclusions: most works analyse a single speech task, mainly the sustained phonation of the vowel /a/ (language-independent speech task)

- The majority of studies focus only on analyzing a single voice segment. In particular the vocalization of the vowel /a/ in the independent language speech task.
- The most analysis is conducted on limited acoustic pathologies from the SVD, AVPD and MEEI databases.
- The conventional dysphonic feature is the feature most extracted to determine the voice aspect for a particular voice pathology.
- Artificial Neural Network (ANN), Random Forest (RF) and Support Vector Machine (SVM) are the most conventional machine learning methods employed for vocal-pathologies detection.

In this paper, we attempt to make a comparative analysis with published results on the speech recordings of the vocalization of the vowel /a/. In spite of other studies, the comparison analysis will cover a bigger segment of SVD [21]. In order to widen the problem scope and maximize the generality of application of the proposed method, we will not limit the vocal pathologies in the database to the popular subtest that is commonly used in the literature. Therefore, a big number of voice pathologies with minimum voice recordings will be included in our dataset for this study. As far as we are concerned, no study presented in the literature [22], is based on a deep learning method for detection of vocal pathologies. In this study, a conventional voice pathology detection method is used and combined with a vocal feature selection method. We also will employ a gradient boosting method as a classifier. An investigation of anomaly detection methods usage is also conducted in this study to manage the wide distribution of vocal pathologies that are associated with a limited number of voice phytology recordings. In this study, we propose automatic rapid voice pathology detection based on

a deep learning classifier, namely a DNN system for voice pathology detection. Our proposed methods are applied through four primary phases, including preparation of dataset following by learning process phase then a training and validation phase and finally an inference processing phase.

This paper is organized as follows: An overview of currents studies and some related works are presented in Section 2. In Section 3, we describe our proposed voice pathology detection based on deep learning. We present the experimental results in Section 4. Finally, we present our conclusions and directions of future research in Section 5.

2. Related Work

The utilization of machine learning (ML) can be useful in many applications such as medical diagnosis [23], cancer detection [24], smart building applications [25], and others [1,11,26]. Machine learning methods are valuable for discriminatory detection and classification tasks [27,28]. These methods have been used in diverse speech identification uses, where one of these uses is pathological voice investigation [29]. The identification and the classification of voice pathology techniques are still one of the difficult domains within the investigation of speech detection. Besides, these basic techniques are expensive and need more time and many sorts of gear [30]. Many researchers focus on the Saarbrücken voice database (SVD) in their studies. The researchers that utilized SVD extracted different features from voice records prior to pathology identification. The features that are frequently extracted are entropy, energy, time, contained Mel-frequency cepstral coefficients (MFCC), cepstral domains, frequency, harmonics-to-noise ratio, short-term cepstral parameters, normalized noise energy, and others [2,31–33]. After this stage the classification task will begin. Many binary and multi-classification methods have been used such K-means clustering, Support Vector Machine, and so on. To our best knowledge, our study is the first study to present voice pathology detection and classification using a convolutional neural network (CNN).

The outcomes of the published studies vary greatly because of the variances among the datasets used in the experimental results. According to Martınez et al. [34], the accuracy achieved utilizing 200 records of sustained vowel /a/ represent a high value and it's very close to our study. Other studies utilized the combination of vowels /a/, /i/ and /u/ to get high accuracy and do not focus on the pathology causes. In the studies by Souissi et al. in [35] they achieved high accuracy of 87.82% utilizing subset involving four kinds of voice pathologies that include 71 types. Also, Al-Nasheri et al. [16,36] achieved an accuracy of 99.68% due to their use of a subset involving a few of the pathologies to conduct a test on information that was moreover displayed in other accessible datasets, such as Arabic Voice Pathology Database (AVPD), and Massachusetts Eye and Ear Infirmary Database (MEEI). Another study conducted by Muhammad et al. [13] utilized a subset involving three kinds of voice pathologies that achieved an accuracy of 93.20%. In addition, they utilized a combination of voice records as an electroglottograph signal to increase the accuracy to 99.98%. However, in another study conducted by Hemmerling et al. [37] they achieved a high accuracy of 100% in the detection issue by their method to separate male and female speakers.

The study of Hammami et al. [38] assessed the execution of the proposed high order statistic feature highlights extricated from wavelet space to segregate between normal voices and pathological ones. Traditional features such as Cruel Wavelet Esteem, Cruel Wavelet Vitality and Cruel Wavelet Entropy were used in the experiments. These highlights, combined with a SVM classifier, reach the most elevated correctness of 99.26% within the location step and 100% when classifying the information. In order to include concrete logical included values a clinical evaluation was performed on information collected from subjects from a healing center in Tunez. The results were acceptable and the precisions were 94.82% and 94.44% for the location and classification, respectively. Fonseca et al. [39] worked on the discovery of co-existent laryngeal issues for which the major phonic side effect is the same, creating features with noteworthy inter-class coverage. Based on the combination of SE, ZCR and SH, all utilized for extraction, related with DPM, particularly received for classification, the proposed approach was effectively concluded, productively dealing within definitions and inconsistencies with an estimated precision of 95%. The ongoing challenge of

dysphonia voice research is the small size of the database produced by Rueda and Krishnan [40]. It is very complicated to use more advanced deep learning methods without underfitting or overfitting. They proposed an adaptive method utilized to break down a signal into its components employing a Fourier-based sychrosqueezing change (FSST) for information enlargement and change. The 2D TF representation output becomes the input to CNN.

It is clear that each voice disorder produces distinctive frequencies depending on the sort of voice disorder and its area on the vocal folds, as we observed. Thus, monitoring the frequency groups is exceptionally vital to evaluate which one contributes more to the discovery and classification of voice afflictions. For example, Pouchoulin et al. [41] stated that lower frequencies (3000 Hz) are more reasonable for recognizing dysphonic voices than higher frequencies. Furthermore, Fraile et al. [42] demonstrated that the control of dysphonic voice flags is altogether less steady within the recurrence area between 2000 and 6400 Hz than the other recurrence regions. To discuss the outcomes of a comparative literature review, they analysed voice records of maintained phonation of the vowel /a/ as well. However, in contrast to past studies, we analyze a bigger database collected from SVD [1]. Moreover, to propose systems capable of powerful voice pathology detection and classification, we do not confine the database as if it were a subset of popular voice pathologies. In this study, the database includes an expansive number of pathologies with small recordings. As we observed in the related works, in spite of past work [1], no other studies have utilized deep learning methods for voice pathology identification. In the following sections we utilize a robust voice pathology identification model based on the acoustic feature extraction strategy. We use voice pathology detection and identification utilizing a CNN approach. We utilize the transfer learning method for using the current powerful CNN models. Particularly, the ResNet34 models were used. To handle the issue of inadequate distribution of an assortment of voice pathologies with few recordings in the datasets, we also explore the utilization of abnormality detection methods.

3. Proposed Methodology

3.1. Dataset Used

As already stated, we have opted to use continuous vowel /a/ phonation as the base for our experiments. A speaker is asked to maintain vowel phoning during this specific speech task, to maintain the amplitude and frequency at a realistic rate [21]. The benefit of this speech task is that it is free of articulative and other linguistic confusions compared with other language standard tasks such as reading or speaking activities. This uniqueness makes it an ideal alternative for this mission for building the large database required for supervised deep learning models [43]. Thus, the only speech task used in this process is sustained /a/ vowel phoning.

The Saarbruecken Voice Database (SVD) is built based on 2000 speakers [10] and voice and electroglottography (EGG) signal sets are included in this dataset. It comprises records of 687 healthy individuals (259 men and428 women) and 1356 individuals (629 men and 727 women) with different pathologies. The recording procedure involves: (a) vowels /i, a, u/ formed in normal speech, (b) high and low pitches; vowels /i, a, u/ with rising-falling pitch; and (c) the German sentence "Guten Morgen, wiegeht es Ihnen?" ("Good morning, how are you?"). Each recorded SVD voice was sampled with a resolution of 16-bit at 50 kHz. This dataset is fairly recent and has therefore been used by very few studies in the field of voice pathology. Following the three diseases criteria, we downloaded files from the website listed in [44] and selected only the continuous vowel /a/ samples generated at normal pitch.

3.2. Proposed Method

The key aim of the study is to extract features that enhance the accuracy for detection and classification of voice pathology and to investigate the impact on the detection and classification processes of different frequency regions (bands). Before feeding to a convolutionary neural network (CNN), the voice signals are processed. To use existing stable CNN models, we use a transfer-

learning platform. The paper explores, in particular, the ResNet34 models. The block diagram of the proposed solution is shown in Figure 1.
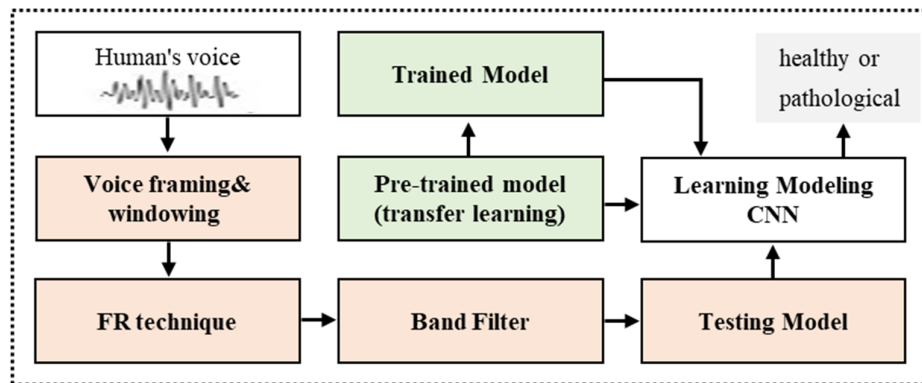

Figure 1. Voice signal processing in the proposed system.

Figure 1 demonstrates the proposed deep learning approach to the detection of voice pathology. The system is fed a patient's voice, and the output determines whether the patient's voice is normal or pathological. The signal of the voice is 1 s. If an input reaches 1 s, then a signal of 1 s is cut from the centre. The signal is split in 40 ms frames, of which the gap is 20 ms. The 40 ms frame duration is a well-equilibrated pitch capture and voice breaks smoothing option. If this is very long, then the voice breaks or some sounds cause the vocal folds to be irregularly opened and closed. The continuation effect and pitch duration are lost if the frame length is short. The framed signal is transformed by a fast Fourier transform to a frequency-domain signal. We get a spectrogram after concatenating all frequency-domains of the frames. The spectrogram could be viewed as an image. The spectrogram includes a minimum of 20 filters for the band pass. The filters are based on the octave. In the area of voice pathology detection, the octave scale typically functions better than the Mel scale [34]. Time derivatives of the first and second order for the octave spectrum output are used. After this method, we get three image-like patterns: the octave and its derivatives of first and second order. The input of the CNN models is made up of three image patterns. We tested ResNet34 in the proposed method.

In this paper, for several reasons the transfer learning strategy for the CNN training is applied: (i) to overcome the lack of adequate voice pathologic attributes, in particular voice diseases derived from patients with reported infections, (ii) reducing the learning duration needed to acquire the final learned typical, and (iii) increasing the classification precision of the voice pathology identification. The technique of transfer learning is intended to boost neural network output in realistic applications bypassing learning from another task [45]. For example, the training of a CNN to classify the case into two groups (e.g., pathological or healthy) may help to classify cases of different disease types. In this case, we used an effective ResNet34 pertained model. Residual Network (ResNet) is one of the highest-profile CNNs and the recipient of the 2015 ILSVRC ImageNet classification award [46]. ResNet is much like the other CNNs, which are sequentially packed with convolutionary, pooling, activation maps and fully interconnected layers. The only big difference between ResNet and other CNNs is the connection identity from the input layer to the end of the residual block (as shown in Figure 2b). The architecture of ResNet34 begins with a convolutionary operation and max-pooling of the use of size kernels (5*5) pixels and (2 *2) pixels, respectively. Thereafter, four stages with a different number with residual blocks are introduced, using size kernels (2 *2) pixels to perform the convolutionary operation. When one passes from one point to the next, the depth of the channel is doubled, and the size of the input sample is halved. In this study the ResNet34 has an average pooling layer with two neurons (for example, positive pathological and normal case as healthy) followed by a completely connected layer. Table 1 illustrates the main details of the ResNet34 architecture. Following the proposed training methodology, experiments were performed using k-folds cross validation samples of SVD voice pathology as a ResNet34 training sample, although the reset samples

were used for the testing. In the course of the training, 20 percent of the training set was chosen randomly and used as a validator set to test the model's general capacity and store the configuration of weights that gave the validation set the minimum error rate. The best model (based on hyper-parameters) used in the proposed Voice Pathology Detection System can be found in Table 2. To summarize, the principal steps in the proposed approach for training are as follows:

1) Divide the SVD dataset into three separate sets: training, test and validation set.
2) Select initial hyper parameter values (e.g., learning rate, dynamic, and so on).
3) Train the ResNet34 utilizing the training group and the hyper-parameters group in step 2.
4) The validation set is used to evaluate ResNet34 performance during the training phase.
5) Reiterate phases 3 to 4 for 12 epochs.
6) Choose the best candidate typical with minimum validation error rate.
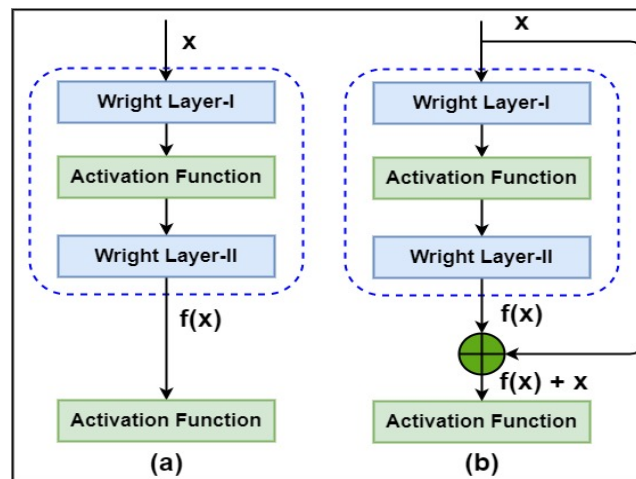7) Through utilizing the test group, the best-trained model is identified.

Figure 2. The difference in a building block: (**a**) A regular block, and (**b**) A residual block.

Table 1. The ResNet34 architecture factors.

Layer Name	ResNet34
Conv-1	5 × 5, 128 stride 2
Pool-1	2 × 2 Max-Pooling stride 2
Step-1	$\begin{bmatrix} 2 \times 2, & 32 \\ 2 \times 2, & 32 \end{bmatrix} \times 2$
Step -2	$\begin{bmatrix} 2 \times 2, & 64 \\ 2 \times 2, & 64 \end{bmatrix} \times 5$
Step -3	$\begin{bmatrix} 2 \times 2, & 128 \\ 2 \times 2, & 128 \end{bmatrix} \times 8$
Step -4	$\begin{bmatrix} 2 \times 2, & 256 \\ 2 \times 2, & 256 \end{bmatrix} \times 2$
Pool-2	Average pooling, stride 1
Fc	2-d, Softmax

Table 2. The best-trained model parameters utilized in Voice Pathology Detection DeepNet System.

Hyper-Parameters	Values
Optimization Method	Adam
Momentum	0.93
Weight-Decay	0.0006
Dropout	0.3
Batch Size	90
Learning Rate	0.02
Total No. of Epochs	10

3.3. Training and Validation Stage

To train and validate the deep learning model, the dataset have been split into different sets as mentioned before. Subsequently, 10-fold cross-validation indices were generated for each set in the training and validation phases so that for each experiment we can use the same data sets. For the final evaluation of the models, the test set was left. Next, the testing and validation sets are stratified to the age and gender classes, by medical status (Healthy—H, Pathology—P). The long recordings were divided into several chunks which were necessary to prevent leakages into the test or validation set. These chunks have been carefully removed from the set. The other chunks were included in the training set. In each point of the validation confusion matrix, we used specific number samples that were taken from 150 healthy—H, pathological—P samples. To detect a pathology using the CNN model, we used 874 pathological and 200 healthy samples for testing the confusion matrix. We separated all the dataset into training, validation, and testing sets and made sure that the number of healthy and pathological samples was equivalent in each of the training and validation sets. The remainder has been added to the test set. In sum, 960 (480 healthy, 480 pathological) samples have been used for the training, 300 (150 healthy, 150 pathologic) samples have been applied for validation, and 874 (200 healthy and 674 pathological) samples have been used for testing. In the training phase, the distribution of the samples is unequal. We responded to this by adjusting the weights of samples, which are used for the minority groups during training to compensate. A 3-part weight product is the weight of the final sample. The number of subgroups in the group chosen (e.g., ratio of normal as well as pathological) is quantified by increasing partial weight. To this end, we presented a class weight α, gender weight β, as well as a group of gender-age weight γ that led to a final sample weight ω that is calculated as $\omega = \alpha \cdot \beta \cdot \gamma$. Furthermore, weights can be determined for a given sample in subgroup αi in group α, βi in group β, and γi in group γ. We have chosen the best hyperparameters for the cross-validation configuration as an output measurement. After tuning the hyperparameters, we have retrofitted and then evaluated the deep learning algorithms with the unsurpassed hyper-parameters over the whole set of training. In terms of a classification report (CR) and a confusion matrix (CM), the final results are presented. Formulation 1, 2 and 3 define how the CR tables calculate the recall, precision, and F1 score (weighting the average accuracy and recall). These three measurements are determined as follows:

- The *Precision* metric is used for measuring the proportion of the subjects that are of great importance. With this metric, the classifier's ability to reject unimportant subjects is measured. The following is an expression of the metric:

$$\text{Precision} = \frac{(TP)}{(TP + FP)} \tag{1}$$

- The *F1 score* is described as the weighted average of the precision and recall, the best value of F1 score is reached at 1, while the worst score is at 0. The precision and recall make an equal relative contribution to the F1 score. The F1 formula is given as follows:

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{2}$$

- The *Recall* metric is used for the evaluation of the proportion of important subjects that are identified. With this metric, the classifier's ability to provide all subjects that are of importance are measured. The following is an expression of the recall metric

$$\text{Recall} = \frac{(TP)}{(TP + FN)} \tag{3}$$

4. Experimental Results

In this work, we propose a deep learning Convolutional Neural Network (CNN) model to perform pathology detection based on numerical analysis of voice signals. It is implemented in a Voice Pathology Detection DeepNet system using Python programming language. It is trained and

tested by using a Google Colaboratory server. The testing computer has a 69 K GPU graphics card and 8 GB of RAM. It runs the Windows 10.1 operating system. The model architecture includes four fully connected layers. The CNN model of the Voice Pathology Detection DeepNet system is tested using the Saarbruecken Voice Database (SVD) dataset [1]. The dataset contains recordings of 71 types of voice signals in which the signals were split into 64 ms long Hamming windowed segments with 30 ms overlap. It is arranged as a sequence of time-based vectors. Each vector as labelled as healthy or pathological class. Table 3 presents the training confusion matrix of the SVD dataset. The table shows that the 300 training samples are divided equally to 150 healthy and 150 pathological classes to ensure a balanced training process.

Table 3. The confusion matrix of the training phase.

Class	True P	True H	No. of Samples
Pred: pathological	139	11	150
Pred: healthy	9	141	150

The CNN model is able to achieve an average prediction score of 93.72% accuracy, 94.11% sensitivity or recall and 95.41% specificity. Table 4 shows the results of precision, f1-score and recall for healthy and pathological classes in the training phase.

Table 4. The classification report of the training phase.

Class	Precision	f1-Score	Recall
Pathological	94.54	93.98	95
Healthy	93.46	92.02	94

Table 5 presents the testing confusion matrix to the SVD dataset. The table shows that the 1074 testing instances are divided into 200 healthy and 874 pathological classes to insure robust testing process.

Table 5. The confusion matrix of the testing phase.

Class	True P	True H	No. of Samples
Pred: pathological	839	35	874
Pred: healthy	11	189	200

In the testing phase, the CNN model is able to achieve an average prediction score of 96.11% accuracy, 95.38% sensitivity or recall and 95.97% specificity. Table 6 shows the results of precision, f1-score and recall for healthy and pathological classes in the testing phase.

**Table 6.** The classification report of the testing phase.

Class	Precision	F1-Score	Recall
Pathological	95.41	94.22	96.13
Healthy	94.59	93.78	95.87

Based on the training methodology that is mentioned early, the setting of the prediction model includes 0.02 learning rate, 2 batch size, 10 number of epochs, and 0.93 momentum as presented in Table 2. Table 7 shows the training Train Loss and testing Valid Loss for each epoch along with the accuracy and time per epoch results. From the observation of the results, during the progress of the Epoch, the Train Loss and testing Valid Loss are slightly increased, the accuracy result is slightly increased and the time per epoch is almost constant at 02:54 s. This result can be attributed to the ResNet34 training algorithm's ability to rapidly adapt to the discriminative features of the SVD dataset and provide significant generalization.

Table 7. The test results of the Voice Pathology Detection system.

Epoch	Train Loss	Valid Loss	Accuracy	Time (s) Per Epoch
0	0.44897	0.21589	0.9514	02:54
1	0.42963	0.198756	0.9510	02:55
2	0.30987	0.15890	0.9631	02:54
3	0.24964	0.136987	0.9601	02:54
4	0.243253	0.129874	0.9628	02:54
5	0.19250	0.112589	0.9536	02:55
6	0.20147	0.023698	0.9574	02:55
7	0.12369	0.029647	0.9611	02:55
8	0.024898	0.002369	0.9594	02:54
9	0.025369	0.004782	0.95.47	02:54

To further analyze the behaviour of the model, we investigated the relationship between the loss and learning rate. The analysis result shows that the best order of magnetite falls in the middle of the learning rate when the loss have values between 0.46 and 0.48 (i.e., between log scales le-04 and le-03) as shown in Figure 3. Hence, the centroid of this area is selected to set the learning rate of the ResNet34 algorithm.
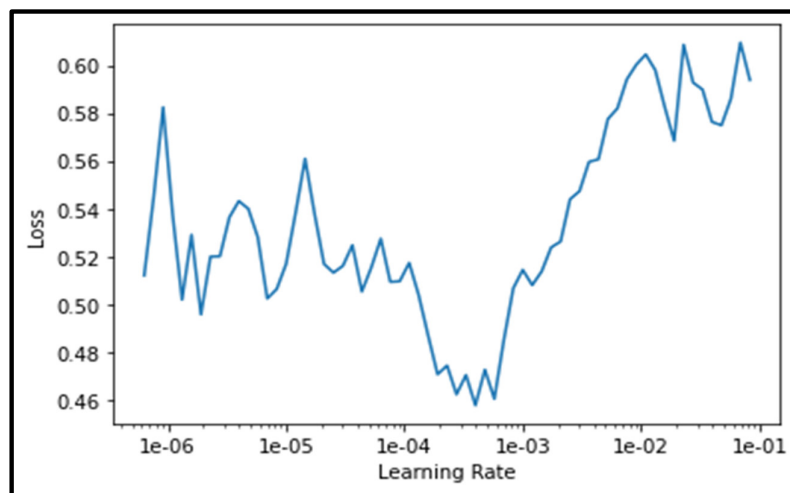
Figure 3. The curve of the log scale of the learning rate against the loss.

Some of the SVD voice pathology features can clinically identify pathological cases without the need for advanced analytical systems. However, voice pathology features changes with the changes in the stages of the disease development which affects the accuracy of the voice pathology detection results. In the early stages of the disease, the voice of the patient will not be significantly affected while when the disease becomes severe there will be a high irregularity in the voice of the patient. Hence, this work proposes a Voice Pathology Detection system that integrates a CNN model to perform voice pathology detection. The deep learning CNN of the Voice Pathology Detection system achieves high prediction accuracy results of up to 96.28% in distinguishing healthy from pathological cases. The main contribution of this work lays on modelling and tuning a deep learning CNN with ResNet34 layers to provide accurate voice pathology detection. Additionally, most of the related works use conventional dysphonic voice pattern analysis features which are easy to predict and even clinically interpretable while this dataset considered complex and challenging [1,12]. On the other hand, the limitations of this work are mainly due to the SVD testing dataset used and can be summarized as: (i) the small size of the tested cases, (ii) the absence of gender separation in the cases and (iii) ignoring the severity of the pathology in the features.

5. Conclusions

This work investigates the possibility of improving the accuracy of voice pathology detection in a search for robust solutions. The main problem hindering progress in this research field is the limited availability of reliable testing samples. Most of the related studies use conventional dysphonic voice features which are easy to predict and clinically interpretable. The Saarbruecken Voice Database (SVD) dataset was selected for this work because it has complex and challenging dysphonic voice pattern analysis features. Subsequently, this paper introduces a novel and real-time system for voice pathology detection using a deep learning Convolutional Neural Network (CNN) model. The model has been implemented in a Voice Pathology Detection system. The development methodology of the Voice Pathology Detection system comprises dataset preparation, learning process, training and validation, and inference process stages. Initially, we apply the SVD dataset in a pre-trained CNN model to set the relative prediction accuracy of the proposed model. This paper aimed to carry out a preliminary study which would clarify whether the use of the deep learning CNN in voice pathology detection, would prove worthy of further exploration. The main contribution of this work is modelling and tuning a deep learning CNN and ResNet34 layers to provide accurate voice pathology detection results. The deep learning CNN of the Voice Pathology Detection system achieves high prediction accuracy results of up to 94.54% accuracy on training data and 95.41% accuracy on testing data. The limited number of samples in general, the limited number of healthy persons compared with the pathology patient and the availability of unique cases in the SVD are the main reasons preventing further improvement of the results. Future work should consider extracting enhanced dataset dimensions and features quality including a new combination of vowels and separating genders. Furthermore, testing different types of CNN and training models might further improve the voice pathology detection approach. Future work could include the application of this method to esophageal voices [46–48] as patients that had a larynx cancer often have a low intelligibility voice and this reduces a lot their social communication skills. Any contribution to this topic of the esophageal speech will be very helpful for patients with a laryngectomy. Finally, utilization of the proposed system in the real-clinical application is promising through providing a fast-automatic diagnosis and treatment solutions within 3 s to achieve the classification accuracy.

Author Contributions: Formal analysis, Mashael S. Maashi and Maria Begona Garcia-Zapirain Soto; Funding acquisition, Ibon Oleagordia Ruiz; Investigation, Hosam Alhakami; Methodology, Karrar Hameed Abdulkareem; Supervision, Mohd Khanapi Abd Ghani; Validation, Salama A. Mostafa; Visualization, Fahad AL-Dhief; Writing – review & editing, Mazin Abed Mohammed. All authors have read and agreed to the published version of the manuscript.

References

1. FAl-Dhief, F.T.; Latiff, N.M.A.A.; Malik, N.N.N.A.; Salim, N.S.; Baki, M.M.; Albadr, M.A.A.; Mohammed, M.A.A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access* **2020**, *8*, 64514–64533.
2. Titze, I.R.; Martin, D.W. Principles of Voice Production; the Journal of the Acoustical Society of America. The Acoustical Society of America. 104, 1148 (1998); doi: 10.1121/1.424266.
3. Teager, H. Some observations on oral air flow during phonation. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 599–601.
4. Hillenbrand, J.; Houde, R.A. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *J. Speech Lang. Hear. Res.* **1996**, *39*, 311–321.
5. Saenz-Lechon, N.; Godino-Llorente, J.I.; Osma-Ruiz, V.; Gómez-Vilda, P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomed. Signal Process. Control* **2006**, *1*, 120–128.

6. Markaki, M.; Stylianou, Y. September. Using modulation spectra for voice pathology detection and classification. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; pp. 2514–2517.

7. Hossain, M.S.; Muhammad, G.; Alamri, A. Smart healthcare monitoring: A voice pathology detection paradigm for smart cities. *Multimed. Syst.* **2019**, *25*, 565–575.

8. Mehta, D.D.; Hillman, R.E. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Curr. Opin. Otolaryngol. Head Neck Surg.* **2008**, *16*, 211.

9. Al-Nasheri, A.; Ali, Z.; Muhammad, G.; Alsulaiman, M. Voice pathology detection using auto-correlation of different filters bank. In Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), Doha, Qatar, 10–13 November 2014; pp. 50–55.

10. Muhammad, G.; Alsulaiman, M.; Ali, Z.; Mesallam, T.A.; Farahat, M.; Malki, K.H.; Al-nasheri, A.; Bencherif, M.A. Voice pathology detection using interlaced derivative pattern on glottal source excitation. *Biomed. Signal Process. Control* **2017**, *31*, 156–164.

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

12. Amami, R.; Smiti, A. An incremental method combining density clustering and support vector machines for voice pathology detection. *Comput. Electr. Eng.* **2017**, *57*, 257–265.

13. Muhammad, G.; Alhamid, M.F.; Hossain, M.S.; Almogren, A.S.; Vasilakos, A.V. Enhanced living by assessing voice pathology using a co-occurrence matrix. *Sensors* **2017**, *17*, 267.

14. Michaelis, D.; Gramss, T.; Strube, H.W. Glottal-to-noise excitation ratio–a new measure for describing pathological voices. *Acta Acust. United Acust.* **1997**, *83*, 700–706.

15. Saldanha, J.C.; Ananthakrishna, T.; Pinto, R. Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. *J. Med Imaging Health Inform.* **2014**, *4*, 168–173.

16. Al-Nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z.; Mesallam, T.A.; Farahat, M.; Malki, K.H.; Bencherif, M.A. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *J. Voice* **2017**, *31*, 113-e9.

17. Mekyska, J.; Janousova, E.; Gomez-Vilda, P.; Smekal, Z.; Rektorova, I.; Eliasova, I.; Kostalova, M.; Mrackova, M.; Alonso-Hernandez, J.B.; Faundez-Zanuy, M.; et al. Robust and complex approach of pathological speech signal analysis. *Neurocomputing* **2015**, *167*, 94–111.

18. Abd Ghani, M.K.; Mohammed, M.A.; Arunkumar, N.; Mostafa, S.A.; Ibrahim, D.A.; Abdullah, M.K.; Jaber, M.M.; Abdulhay, E.; Ramirez-Gonzalez, G.; Burhanuddin, M.A. Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques. *Neural Comput. Appl.* **2020**, *32*, 625–638.

19. Abdulkareem, K.H.; Mohammed, M.A.; Gunasekaran, S.S.; Al-Mhiqani, M.N.; Mutlag, A.A.; Mostafa, S.A.; Ali, N.S.; Ibrahim, D.A. A Review of Fog Computing and Machine Learning: Concepts, Applications, Challenges, and Open Issues. *IEEE Access* **2019**, *7*, 153123–153140.

20. Mohammed, M.A.; Al-Khateeb, B.; Rashid, A.N.; Ibrahim, D.A.; Ghani, M.K.A.; Mostafa, S.A. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. *Comput. Electr. Eng.* **2018**, *70*, 871–882.

21. J. Barry, P ützer, M., Saarbrucken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/

22. Harar, P.; Alonso-Hernandezy, J.B.; Mekyska, J.; Galaz, Z.; Burget, R.; Smekal, Z. Voice pathology detection using deep learning: A preliminary study. In Proceedings of the 2017 international conference and workshop on bioinspired intelligence (IWOBI), Funchal, Portugal, 10–12 July 2017; pp. 1–4.

23. Mohammed, M.A.; Ghani, M.K.A.; Hamed, R.I.; Ibrahim, D.A.; Abdullah, M.K. Artificial neural networks for automatic segmentation and identification of nasopharyngeal carcinoma. *J. Comput. Sci.* **2017**, *21*, 263–274.

24. Mohammed, M.A.; Ghani, M.K.A.; Arunkumar, N.A.; Hamed, R.I.; Abdullah, M.K.; Burhanuddin, M.A. A real time computer aided object detection of nasopharyngeal carcinoma using genetic algorithm and artificial neural network based on Haar feature fear. *Future Gener. Comput. Syst.* **2018**, *89*, 539–547.

25. Djenouri, D.; Laidi, R.; Djenouri, Y.; Balasingham, I. Machine learning for smart building applications: Review and taxonomy. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–36.

26. Alhussein, M.; Muhammad, G. Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access* **2018**, *6*, 41034-41041.

27. Mostafa, S.A.; Mustapha, A.; Mohammed, M.A.; Hamed, R.I.; Arunkumar, N.; Ghani, M.K.A.; Jaber, M.M.; Khaleefah, S.H. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. *Cogn. Syst. Res.* **2019**, *54*, 90–99.

28. Obaid, O.I., Mohammed, M.A., Ghani, M.K.A., Mostafa, A. and Taha, F., 2018. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology, 7(4.36), pp.160-166.

29. Mohammed, M.A.; Ghani, M.K.A.; Arunkumar, N.A.; Mostafa, S.A.; Abdullah, M.K.; Burhanuddin, M.A. Trainable model for segmenting and identifying Nasopharyngeal carcinoma. *Comput. Electr. Eng.* **2018**, *71*, 372–387.

30. Kukharchik, P.; Martynov, D.; Kheidorov, I.; Kotov, O. September. Vocal fold pathology detection using modified wavelet-like features and support vector machines. In Proceedings of the 2007 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007; pp. 2214–2218.

31. Dubuisson, T.; Dutoit, T.; Gosselin, B.; Remacle, M. On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination. *EURASIP J. Adv. Signal Process.* **2009**, *2009*, 173967.

32. Fredouille, C., Pouchoulin, G., Bonastre, J.F., Azzarello, M., Giovanni, A. and Ghio, A., 2005. Application of automatic speaker recognition techniques to pathological voice assessment. In proceeding of international conference on acoustic speech and signal processing (ICASSP 2005).

33. Wang, J.; Jo, C. Performance of gaussian mixture models as a classifier for pathological voice. In Proceedings of the 11th Australian International Conference on Speech Science and Technology, Australia, 2006; Volume 107, pp. 122–131.

34. Martínez, D.; Lleida, E.; Ortega, A.; Miguel, A.; Villalba, J. 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*; Springer: Berlin/Heidelberg, Germay, 2012; pp. 99–109.

35. Souissi, N.; Cherif, A. December. Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In Proceedings of the 2015 7th international conference on modelling, identification and control (ICMIC), Sousse, Tunisia, 18–20 December 2015; pp. 1–6.

36. Al-nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z. Investigation of voice pathology detection and classification on different frequency regions using correlation functions. *J. Voice* **2017**, *31*, 3–15.

37. Hemmerling, D.; Skalski, A.; Gajda, J. Voice data mining for laryngeal pathology assessment. *Comput. Biol. Med.* **2016**, *69*, 270–276.

38. Hammami, I.; Salhi, L.; Labidi, S. Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features. IRBM 2020, 41, 161-171.

39. Fonseca, E.S.; Guido, R.C.; Junior, S.B.; Dezani, H.; Gati, R.R.; Pereira, D.C.M. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). *Biomed. Signal Process. nd Control* **2020**, *55*, 101615.

40. Rueda, A.; Krishnan, S. Augmenting Dysphonia Voice Using Fourier-based Synchrosqueezing Transform for a CNN Classifier. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6415–6419.

41. Pouchoulin, G.; Fredouille, C.; Bonastre, J.F.; Ghio, A.; Révis, J. Characterization of the Pathological Voices (Dysphonia) in the Frequency Space; nternational Congress ofPhonetic Sciences (ICPhS), Aug 2007, Saarbrücken, Germany. pp.1993-1996.

42. Fraile, R.; Godino-Llorente, J.I.; Sáenz-Lechón, N.; Osma-Ruiz, V.; Gutiérrez-Arriola, J.M. Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech. *J. Voice* **2013**, *27*, 11–23.

43. Arunkumar, N., Mohammed, M.A., Ghani, M.K.A., Ibrahim, D.A., Abdulhay, E., Ramirez-Gonzalez, G. and de Albuquerque, V.H.C., 2019. K-means clustering and neural network for object detecting and identifying abnormality of brain tumor. Soft Computing, 23(19), pp.9083-9096.

44. Little, M., McSharry, P., Hunter, E. et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nat Prec (2008). https://doi.org/10.1038/npre.2008.2298.1

45. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.

46. Ishaq, R.; Zapirain, B.G.; Shahid, M.; Lovstrom, B. Subband Modulator Kalman filtering for Single Channel Speech Enhancement. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7442–7446.

47. Garcia, B.; Vicente, J.; Alonso, A.; Loyo, E. Esophageal voices: Glottal flow restoration. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 23-23 March 2005; pp. 141–144.

48. Garcia, B.; Ruiz, I.; Vicente, J.; Alonso, A. Formants measurement for esophageal speech using wavelet with band and resoultion adjustment. In Proceedings of the 2006 IEEE International Symposium on Signal Processing and Information Technology, Vancouver, BC, Canada, 27–30 August 2006; pp. 320–325.