Article

Deep Neural Network-Based Respiratory Pathology Classification Using Cough Sounds

B T Balamurali [1,*], Hwan Ing Hee [2,3], Saumitra Kapoor [1], Oon Hoe Teoh [4], Sung Shin Teng [5], Khai Pin Lee [5], Dorien Herremans [6] and Jer Ming Chen [1]

1   Science, Mathematics and Technology, Singapore University of Technology and Design,
    Singapore 487372, Singapore; saumitrakapoor@gmail.com (S.K.); jerming_chen@sutd.edu.sg (J.M.C.)
2   Department of Paediatric Anaesthesia, KK Women's and Children's Hospital, Singapore 229899, Singapore;
    hee.hwan.ing@singhealth.com.sg
3   Anaesthesiology and Perioperative Sciences, Duke-NUS Medical School, 8 College Road,
    Singapore 169857, Singapore
4   Respiratory Medicine Service, Department of Paediatrics, KK Women's and Children's Hospital,
    Singapore 229899, Singapore; teoh.oon.hoe@singhealth.com.sg
5   Department of Emergency Medicine, KK Women's and Children's Hospital, Singapore 229899, Singapore;
    teng.sung.shin@singhealth.com.sg (S.S.T.); lee.khai.pin@singhealth.com.sg (K.P.L.)
6   Information Systems, Technology, and Design, Singapore University of Technology and Design,
    Singapore 487372, Singapore; dorien_herremans@sutd.edu.sg
*   Correspondence: balamurali_bt@sutd.edu.sg

Abstract: Intelligent systems are transforming the world, as well as our healthcare system. We propose a deep learning-based cough sound classification model that can distinguish between children with healthy versus pathological coughs such as asthma, upper respiratory tract infection (URTI), and lower respiratory tract infection (LRTI). To train a deep neural network model, we collected a new dataset of cough sounds, labelled with a clinician's diagnosis. The chosen model is a bidirectional long–short-term memory network (BiLSTM) based on Mel-Frequency Cepstral Coefficients (MFCCs) features. The resulting trained model when trained for classifying two classes of coughs—healthy or pathology (in general or belonging to a specific respiratory pathology)— reaches accuracy exceeding 84% when classifying the cough to the label provided by the physicians' diagnosis. To classify the subject's respiratory pathology condition, results of multiple cough epochs per subject were combined. The resulting prediction accuracy exceeds 91% for all three respiratory pathologies. However, when the model is trained to classify and discriminate among four classes of coughs, overall accuracy dropped: one class of pathological coughs is often misclassified as the other. However, if one considers the healthy cough classified as healthy and pathological cough classified to have some kind of pathology, then the overall accuracy of the four-class model is above 84%. A longitudinal study of MFCC feature space when comparing pathological and recovered coughs collected from the same subjects revealed the fact that pathological coughs, irrespective of the underlying conditions, occupy the same feature space making it harder to differentiate only using MFCC features.

Keywords: LRTI; URTI; asthma; cough classification; respiratory pathology classification; MFCCs; BiLSTM; deep neural networks

1. Introduction

Cough is a prevalent clinical presentation in many childhood respiratory pathologies including asthma, upper and lower respiratory tract infection (URTI and LRTI), atopy, rhinosinusitis and post-infectious cough [1–3]. Because of its wide range of aetiologies, the cause of cough can be misdiagnosed and inappropriately treated [1]. Clinical differentiation for pathological respiratory conditions takes into consideration the history of the presenting respiratory symptoms as well as clinical signs such as pyrexia (i.e., raised

body temperature), respiratory rate, shortness of breath and chest auscultation of pathognomonic breath sounds. In some cases, additional investigations such as chest radiographs, laboratory blood tests, bronchoscopy and spirometry are required to reach a definitive diagnosis. These investigations often require hospital visits and place demands on healthcare resources. Moreover, such visits may create a negative social economic impact on the ill child and on his/her family (such as time away from work and childcare arrangements). Furthermore, some of these investigations such as chest radiographs, and blood tests can result in more harm than benefit, if performed indiscriminately.

There is a growing interest in characterizing acoustic features to allow objective classification of cough sounds originating from different respiratory conditions. Previous studies have looked at medical screenings based on cough sounds [4–8]. Abaza et al. [4] analysed the characteristics of airflow and the sound of a healthy cough to train a classifier that distinguishes between healthy subjects and those with some kind of lung disease. Their model incorporates a reconstruction algorithm that uses principal component analysis. It obtained an accuracy of 94% and 97% to identify abnormal lung physiology in female and male subjects, respectively. Murata et al. [5] used time expanded wave forms combined with spectrograms to differentiate between productive (i.e., coughs producing phlegm) and non-productive coughs (i.e., dry coughs). Cough sound analysis has also been used to diagnose pneumonia [6] and Swarnkar et al. [7] used it to assess the severity of acute asthma. The latter reported that their model can predict between children suffering from breathing difficulties involving acute asthma and can characterize the severity of airway constriction. In [9], tuberculosis (TB) screening was investigated using short-term spectral information extracted from cough sounds. They reported an accuracy of 78% when distinguishing between coughs of TB positive patients and healthy control group. Furthermore, it was noted that the TB screening accuracy increased to 82% when clinical measurements were included along with features extracted from cough audio. The cough sounds used in some of the aforementioned investigations were carefully recorded in studio environments, whereas the database used in this investigation is collected using a smartphone in a real hospital setting (see Section 2). This type of ecological data collection (or unconstrained audio collection) is of more practical use to physicians, and may also help in developing a mobile phone app in the future that will be more robust when performing early diagnosis of respiratory tract infections in a real-life setting.

There are some studies that use a realistic cough sound database: A Gabor filterbank (GFB) [8] was used to classify coughs sounds as being 'dry' or 'productive'. They reported an accuracy of more than 80% when incorporating acoustic cough data collected through a public telephone hotline. Another study reported a similar accuracy in classifying wet and dry cough sounds, though the data were collected using a smartphone [10]. Recently, this strategy of collecting cough sounds has become popular [11–13]. Such audio-based strategy has profound implication when examining symptomatic cough sounds associated with COVID-19 whereby cough is a primary symptom, alongside fever and fatigue. Convolution Neural Network (CNN)-based systems were trained to detect cough and screen for COVID-19, and reported accuracy exceeding 90% in [14–16] and while another study had reported 75% accuracy [17]. Features were extracted (both handcrafted and transfer learned) from a crowd-sourced database containing breathing and cough sounds [18] and were used to train a support vector machine and ensemble classifiers to screen COVID-19 individuals from healthy controls. They reported an accuracy around 80%.

There is another line of research inquiry which mainly focuses on cough event detection (i.e., to identify the presence of cough events) in audio recordings [19–24]; however, in this investigation, we manually segment the cough epochs, and thus review of such studies is outside the scope of this report. Having said that, with the advent of deep learning, there is good progress made in the cough event detection from smartphone recordings, and incorporating such techniques at the preprocessing stage in the cough screening system could bypass the tedious manual segmentation process altogether [25–27].

This study aims to determine if a predictive machine learning model, trained using acoustic features extracted from cough sounds, could be a useful classifier to differentiate common pathological cough sounds from healthy-voluntary coughs (i.e., cough sounds collected from healthy volunteers). The knowledge gained through such methods could support with the early recognition and triage of medical care, as well as assist physicians with the clinical management which includes making a differential screening and monitoring of the health status in response to medical interventions.

In the authors' earlier work, audio-based cough classification using machine learning has shown to be a potentially useful technique to assist in differentiating asthmatic cough sounds from healthy-voluntary cough sounds in children [28,29]. The current paper builds upon this previous work (the earlier one used a simple Gaussian Mixture Model—Universal Background Model (GMM-UBM) [28,29]) and uses the collected cough sound dataset to train a deep neural network (DNN) model that can differentiate between pathological and healthy-voluntary subjects. The proposed deep neural network model is trained using acoustic features extracted from the cough sounds. Three different pathological conditions were considered in this investigation: asthma, upper respiratory tract infection (URTI) and lower respiratory tract infection (LRTI). The accuracy of the proposed trained model is evaluated by comparing their predictions against the clinician's diagnosis.

2. Data Collection

2.1. Subject Recruitment

Subjects in this study were divided into 2 cohorts: Healthy cohort (without respiratory conditions) and the pathological cohort (with respiratory conditions which included LRTI, URTI and asthma; LRTI included a spectrum of respiratory diseases such as bronchiolitis, bronchitis, bronchopneumonia, pneumonia, lower respiratory tract infection). Participants were recruited from KK Children's Hospital, Singapore. Inclusion criteria in the pathological cohort was the presence of concomitant symptom of cough, while inclusion criteria for the healthy cohort was the absence of active cough and active respiratory conditions. Pathological cohorts were recruited from the Children's Emergency Department, Respiratory Ward, and Respiratory Clinic. The cough sounds were recorded during their initial presentation at the hospital. The healthy cohorts were recruited from the Children Surgical Unit. These healthy children were first screened by the anaesthetic team and recruited for the study.

2.2. Cough Dataset

A smartphone was used to record cough sounds from both pathological and healthy children (i.e., without respiratory conditions). For both groups, the subjects were instructed to cough actively. This often resulted in multiple cough epochs per participant (on average 10 to 12). Recordings were collected at a sampling rate of 44.1 kHz in an unconstrained clinic setting, i.e., a hospital ambience with background noise such as talking in background, beeping sounds from monitoring devices, alarm sounds, ambulance siren, etc. The collected cough audio files were manually segmented into individual coughs (such that non-cough signal portions are negligible) to form different entries in the dataset. Characteristics of the resulting dataset are shown in Table 1. The working diagnosis for the aetiology of the cough was determined by the clinician based on the clinical history, physical examination, and for some cases investigations such as laboratory tests and chest X-rays were also used.

Table 1. Characteristics of the collected cough dataset.

	Healthy Cohort	Asthma Cohort	LRTI Cohort	URTI Cohort
Number of Subjects	89	89	160	78
Number of Coughs	1149	1192	2344	1240
Age in Years (SD)	9.07 (2.88)	8.51 (3.02)	6.77 (2.65)	7.21 (2.96)
Gender—(Male:Female)	80:9	60:29	94:66	35:43
Race—Chinese	38	24	73	33
Race—Malay	43	44	54	34
Race—Indian	6	14	22	7
Race—Others	2	7	11	4
Duration of history of cough at presentation; day (SD)	NA *	3.87 (4.23)	6.63 (5.93)	5.22 (2.72)

NA *—Not Applicable.

3. Trained Models

Using the dataset described above, five different classification models based on deep neural networks were built.

3.1. Healthy vs. Pathology (2-Class) Model

The first model (*Healthy vs. Pathology (2-class) Model*) was trained to classify whether each cough segmented is a healthy-voluntary cough or pathological. Here, we consider all pathological coughs as one class, known as 'pathological cough'.

3.2. Healthy vs. LRTI Model, Healthy vs. URTI Model, Healthy vs. Asthma Model

The second set of models (three in total) were trained to classify between healthy-voluntary coughs and a particular respiratory pathology (i.e., one respiratory pathology at a time). *Healthy vs. LRTI Model*—was trained to predict whether a cough is healthy-voluntary or from a subject diagnosed with LRTI; *Healthy vs. URTI Model*—was trained to predict whether a cough is healthy-voluntary or from a subject diagnosed with URTI; and finally *Healthy vs. Asthma Model*—was trained to predict whether a cough is healthy or from a subject diagnosed with Asthma.

3.3. Healthy vs. Pathology (4-Class) Model

The final classification model was trained to predict all the four chosen classes. Thus, *Healthy vs. Pathology (4-class) Model*—classifies whether a cough is healthy-voluntary or associated with any of the three pathological conditions of LRTI, URTI, or asthma.

4. Classification Model

4.1. Long–Short-Term Memory (LSTM)

An LSTM-based network was chosen as the classification model in this investigation. LSTM networks take sequence data as the input, and makes predictions based on their sequence dynamic characteristics by learning long-term dependencies between time steps of sequence data. They are known to work well for their ability to handle sequence data due to their memory mechanism [30]. Our choice for LSTM is motivated by the sequential nature of audio data and its ability to handle input audio features that vary in length [30,31], as is the case with the features extracted from the collected cough sounds (see Section 5.3).

In this investigation, we used a four-layer neural network with two deep layers of bidirectional LSTMs (BiLSTMs) (see Figure 1). Each BiLSTM layer learns bidirectional long-term dependencies from sequence data. These dependencies will help the network to understand the long-term dynamics present in the features and thus learning the complete time series [32,33]. We have investigated different deep neural network types such as fully connected deep neural networks, LSTMs, BiLSTMs, to identify the best classification model

for our cough screening problem. In the end, BiLSTMs were chosen, as they were found to produce better results for the chosen feature sets (These network comparison results are not shown as they are outside the scope of this paper; a similar outcome preferring BiLSTM was reported in [33]).
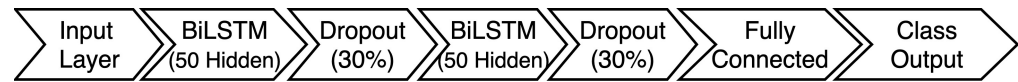

Figure 1. Deep Neural Network Architecture using BiLSTM layers.

4.2. BiLSTM Architecture

The first layer (input layer) has a dimension of 42 to match the size of the MFCC feature vectors corresponding to every audio frame (see Section 5.3). The second layer is a BiLSTM layer with 50 hidden units. This is followed by a dropout layer which in turn is followed by another BiLSTM and a dropout layer. The second BiLSTM layer also has 50 hidden units. A 30% dropout was chosen for both dropout layers. Finally, depending on the classification objective, we used either two fully connected layers (for the 2-class classification problem) or four fully connected layers (for the 4-class classification problem). The networks were optimized to minimize cross-entropy loss with sigmoid activation. This particular architecture was selected after multiple hyper-parameter optimization steps. We used grid search to find the optimal number of hidden units, the number of hidden layers, as well as the dropout rate. The resulting combination reported in this paper was able to reach the lowest training loss (or in other words maximum training accuracy; precluding overfitting of the classifier) when trained for multiple cough classification hypotheses.

5. Experimental Methodology

5.1. Dataset Split

The collected dataset was randomly split (70–30%) into two non-overlapping parts: training and test set. The resulting split sizes are shown in Table 2. We made sure that cough sounds belonging to the same person were either in the test or in the training set, but not in both. Since the test data have not yet been seen by the model during the training phase, one could expect that the resulting performance of this model offers a good approximation for what can be expected in a real scenario (i.e., when the model is asked to make a prediction for an unseen cough).

Table 2. Number of instances of the cough sounds in the training and test set.

Class	Number of Children		Number of Coughs	
	Training	Test	Training	Test
URTI	54	24	849	391
LRTI	113	47	1,679	665
Asthma	65	24	726	466
Healthy	51	38	645	504

5.2. Methodology

The general experimental methodology followed in this investigation is shown in Figure 2. We first trained our deep neural network models using features extracted from data from our training set, and then proceeded to evaluate the models using a separate test set. The trained model is used to predict which class a cough sound belongs to. This cough prediction was subsequently used to screen whether a subject is healthy or having some respiratory conditions. This screening is done based on the most frequent (mode) prediction outcome of all the cough sounds belonging to a particular subject. In what follows, we discuss how the data have been pre-processed, which audio features were chosen for this investigation, and how the model was built.
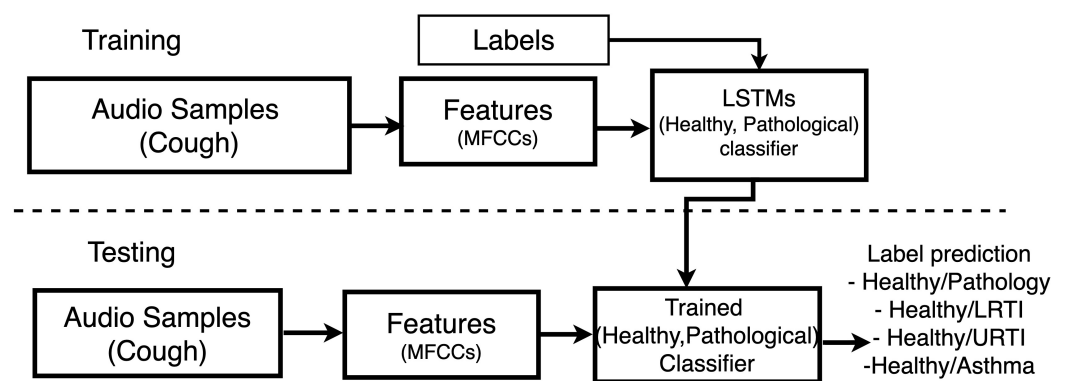
Figure 2. Experimental Methodology followed in our investigation.

5.3. Cough Sound Processing and Audio Feature Extraction

The segmented cough sounds were detrended to remove any linear trends, baseline shifts, or slow drifts, then normalized (to have a maximum sample value of one), and finally downsampled (downsampled to 11.025 kHz from the original sampling rate of 44.1 kHz).

The pre-processed audio signals were first segmented into frames of 100 ms, after which a Hamming window was applied, followed by the extraction of audio features. Mel-Frequency Cepstral Coefficients (MFCCs) were chosen for this investigation owing to their effectiveness when it comes to audio classification problems [34,35]. MFCCs are a set of features that focus on the perceptually relevant aspects of the audio spectrum, additionally the coefficients could contain information about the vocal tract characteristics [36,37]. In this investigation we used 14 MFCCs with their deltas and delta-deltas, thus resulting in a total of 42 coefficients (14 MFCCs, 14 deltas and 14 delta-deltas) for every audio frame. The result obtained using MFCCs thus serves as a baseline against which future investigations can be compared.

5.4. Measuring Performance

The performance of DNN models is measured by calculating the classification accuracy and is further analysed using the receiver operating characteristic (ROC) [38] and confusion matrix [39].

5.4.1. Accuracy

The classification accuracy is calculated by comparing the predicted outputs with the actual outputs.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (1)$$

5.4.2. Receiver Operating Characteristic (ROC)

The ROC is created by plotting the true positive rates (i.e., sensitivity (or recall): the ratio of true positives over the sum of true positives and false negatives) against the false positive rates (i.e., (100—specificity); specificity is the ratio of true negatives over the sum of false positives and true negatives) for various decision thresholds. A perfect model results in a ROC curve which passes close to the upper left corner, indicating a higher overall accuracy. This would thus result in a ROC of which the area underneath (AROC) equals 1.

5.4.3. Confusion Matrix

The performance of a classifier was further analysed using confusion matrices, whereby the true and false positives and negatives are displayed for each class. For a good classifier, the resulting confusion matrix will have large numbers along the diagonal (i.e., values closer to 100%). The percentage of misclassified data is reflected in the off-diagonal elements.

6. Results

6.1. Power Spectrum Comparison

From the original cough sounds, the power spectrum (i.e., the distribution of energy contained within the signal over various frequencies) was estimated. These frequencies were then grouped into five equal bins between 0 to $f_s/2$ (whereby f_s is the sampling frequency) and the corresponding spectral power present in each of these bins was calculated.

The distribution of the power spectrum for 500 randomly chosen cough samples (of different respiratory conditions) is shown using a boxplot (see Figure 3). The median is shown using a red line. The bottom and top edges of each of the boxes indicate the 25th and 75th percentile, respectively. The likely range of variation (i.e., inter-quartile range (IQR)) is given by distances between the tops and bottoms [40].
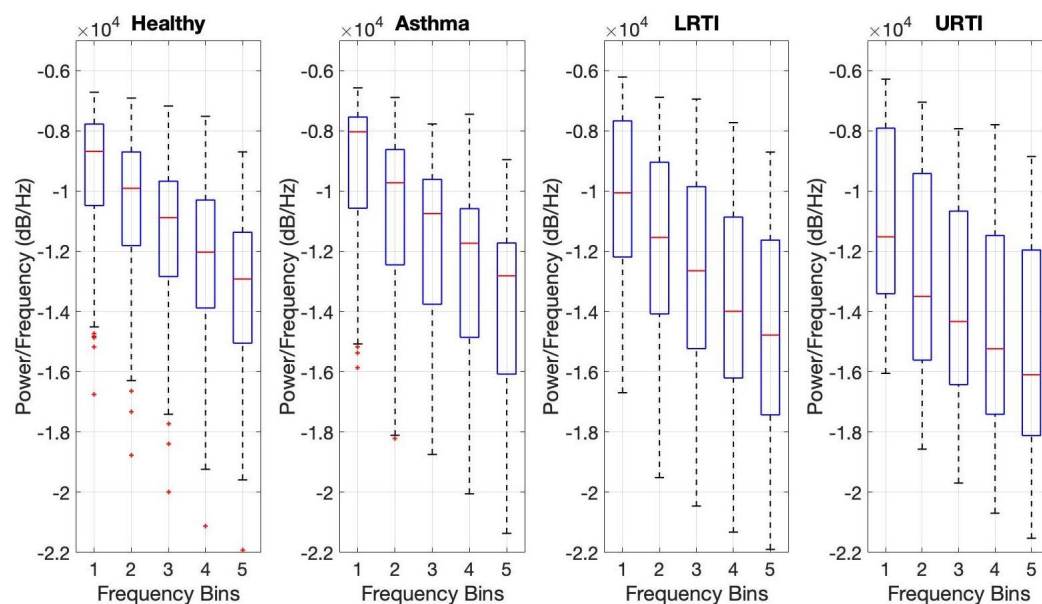

Figure 3. Boxplot showing the distribution of the power spectrum across various frequency bins: Bin 1—0 to 1.1 kHz, Bin 2—1.1 to 2.2 kHz, Bin 3—2.2 to 3.3 kHz, Bin 4—3.3 to 4.4 kHz, Bin 5—4.4 to 5.5 kHz

The median line corresponding to every bin (for both the healthy and pathological coughs) does not appear to be centred inside the box (i.e., the possible mean of each bin), thus indicating that the power distribution is slightly skewed for each bin. IQR is found to be slightly larger in spectral power bins of pathological cough when compared to the healthy spectral bin. Overall, there are no clear trends between the median value of the spectral bin for healthy and pathological coughs. The asthmatic spectral bins tend to have a slightly higher median value compared to the spectral bins of healthy coughs. The opposite trend was found when comparing spectral bins of LRTI and URTI against that of healthy. We speculate that this may be due to the fact that both these conditions (LRTI and URTI) include inflamed airway tissues, which may increase acoustic damping (especially at high frequency). This postulate requires further investigation. In addition, the difference observed maybe attributed to variability in subject characteristics between the groups such as age, gender between groups (see Table 1).

6.2. Feature Analysis—MFCCs—Extracted for Investigation

The objective of this feature analysis is to understand if cough sounds contain any subtle cues to distinguish between healthy and pathological subjects. The higher-dimensional MFCC features extracted from various respiratory pathological coughs were compared against the healthy coughs after transforming them to a lower dimension using Principal Component Analysis (PCA) [41]. Such dimensionality reduction techniques often give some insight into the feature space of the chosen classes. The resulting visualisation of the

first three PCA components (the first three principal components correspond to the largest three eigen values and capture more than 95% of the variance (information) in this dataset) is shown in Figure 4. MFCCs extracted from 5000 audio frames from each of the categories were used for this visualisation. All these audio frames were part of the training set used for training the BiLSTM network.

No clear clusters are visible in the feature space (see Figure 4). This is true for all the four investigated cases: features of healthy versus pathological cough sound signal, and features of healthy coughs when compared to features from each individual respiratory pathologies (see Figure 4a–d). This reflects anecdotal observations that clinicians themselves find it hard to distinguish these pathologies based on cough sound alone.
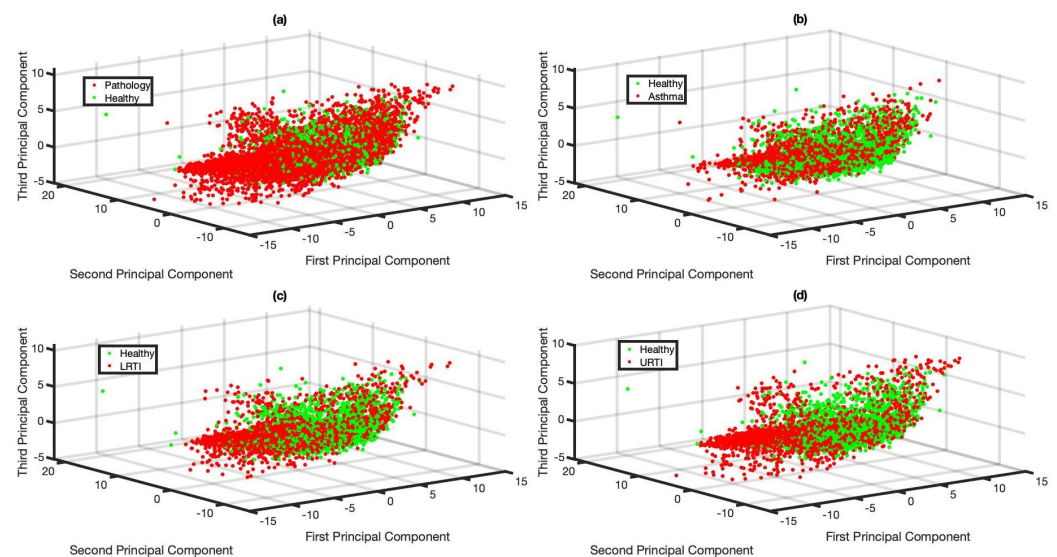

Figure 4. MFCC feature visualisation after transforming the original 42 dimensions to 3 dimensions using PCA: (**a**) Healthy vs. Pathology, (**b**) Healthy vs. Asthma, (**c**) Healthy vs. LRTI, (**d**) Healthy vs. URTI.

6.3. Feature Analysis—MFCCs—Longitudinal Study

The objective of this longitudinal study is to understand the evolution of the feature space of MFCCs over time for the different classes of respiratory conditions. For this study, the cough sounds were collected and organised in a two-stage process. In the first stage, 51 subjects recruited from the hospital were asked to make multiple voluntary cough sounds (on average 10 to 12 coughs). There were 24 subjects with asthma, seven with URTI and 20 with LRTI. In the second stage, these 51 subjects were followed up upon recovery after hospital discharge (approximately two weeks after hospital discharge) and voluntary cough sounds (on average 10 to 12) were again collected. It is important to note here that Stage 1 coughs were a part of the cough dataset used for training the BiLSTM model; however, Stage 2 coughs were not used in any training process. The cough sounds were recorded as described in Section 2.2.

MFCCs were extracted from the coughs collected from these 51 subjects as described in Section 5.3. There was a total of 3810 frames analysed as part of this longitudinal study: 1675—recovered, 746—LRTI, 399—URTI and 990—Asthmatic. The extracted MFCCs' dimensionalities were then reduced using PCA for visualisation purposes (see Figure 5). Stage 1 coughs can be considered to be pathological whereas Stage 2 coughs (i.e., recovered) can be considered to represent healthy-voluntary coughs. The evolution of the MFCC feature space is explored here, since the coughs were collected from the same subject over a period of time. As in Figure 4, no clear clusters are visible when analysing evolution of the extracted features (see Figure 5). Additionally, it can be seen that MFCC features extracted from Stage 1 coughs occupy relatively the same feature space irrespective of the underlying respiratory conditions (see Figure 5b).
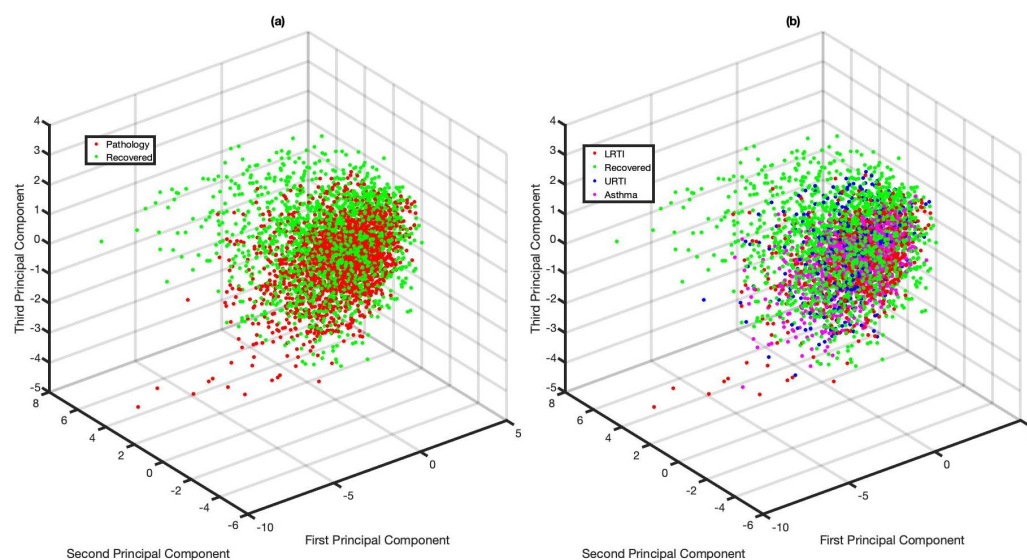
Figure 5. MFCC feature evolution from pathological to recovery. The plot is created after transforming the original 42 dimensions into three using PCA: (**a**) Recovered vs. Pathology, (**b**) Recovered vs. LRTI vs. URTI vs. Asthma.

With no clear clusters visible in the feature space analysis discussed in Sections 6.2 and 6.3, our classification problem may require the introduction of non-linearity, so as to uncover more complex, hidden, relationships. This thus presents an additional motivation for choosing a deep neural network.

6.4. Model Performance

6.4.1. Healthy vs. Pathology Model

The cough classification accuracy (i.e., accuracy in classifying each cough segment) and the healthy-pathology classification accuracy (i.e., accuracy in classifying entire cough epochs to a particular respiratory pathology) on our test set are shown in Table 3. The BiLSTM has resulted in good performance when classifying the pathological cough sounds from healthy-voluntary cough sounds, with an accuracy of 84.5%. Furthermore, when respiratory pathology classification of subject was made (by considering the entire cough epochs) based on the most frequent (mode) prediction outcome of coughs from a subject for an entire cough epoch, the accuracy is even higher (91.2%). This is to be expected, e.g. if one assumes there are n coughs available per subject, even though model misclassifies individual cough sounds, the respiratory pathological classification result will be wrong only when $(n/2) + 1$ out of the n coughs belonging to a particular patient are misclassified (or in other words respiratory pathological classification is more robust). Given an accuracy rate of 84.5% for individual cough prediction, this would be very rare.

Table 3. Accuracy of the *Healthy vs. Pathology Model.*

Model	Individual Cough Classification Accuracy (in %)	Respiratory Pathology Classification of Subject Accuracy Based on Entire Cough Epoch (in %)
Healthy vs. pathology Model	84.5	91.2

A confusion matrix was created to further analyse the results of this model (see Figure 6). The percentage of healthy-voluntary coughs misclassified as pathological coughs is higher compared to pathological coughs misclassified as healthy-voluntary coughs (23.8% misclassified compared to 7.1%, see Figure 6a). This higher healthy-voluntary

cough misclassification rate further resulted in a relatively large number of healthy subjects misclassified as having a pathology (15.6 % subjects were misclassified, see Figure 6b).
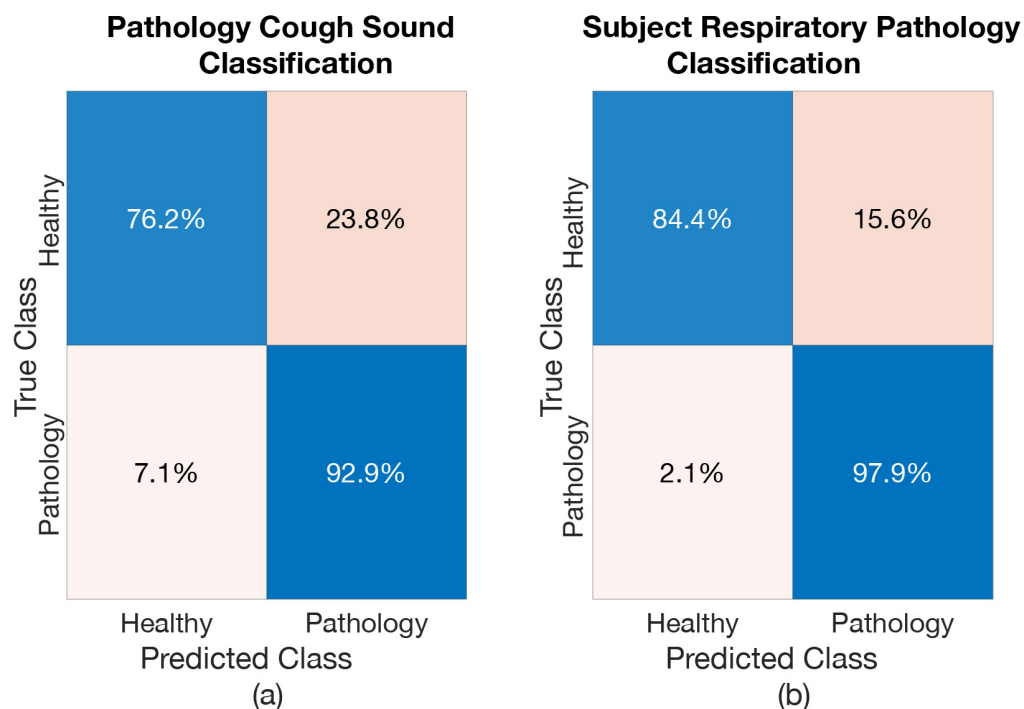

Figure 6. Confusion Matrix of *Model Healthy vs. pathology*: (**a**) when classifying coughs, (**b**) when classifying subject for respiratory pathology.

The receiver operating characteristic of this model is shown in Figure 7, along with the corresponding AROC value. The resulting AROC values are 0.84 for cough classification and 0.91 for respiratory pathology classification of subject, see Table 4). The AROC is convincingly high, which means that the model has delivered good separability between two classes. Additionally shown in Figure 7, is the optimum threshold, co-located in the nearest point to (0, 1), which maximizes the sensitivity and specificity values (shown as a red cross).
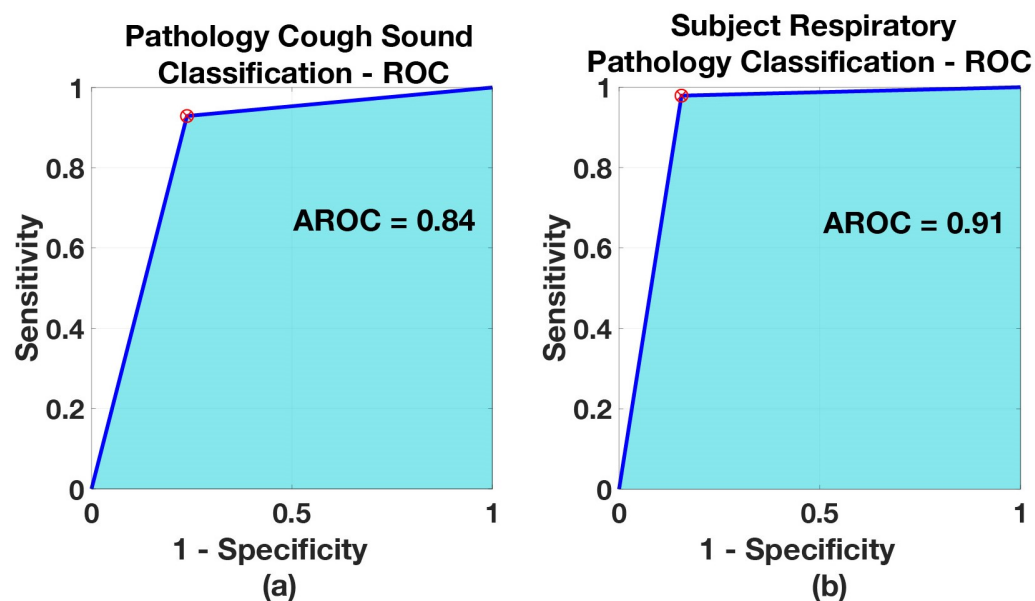

Figure 7. ROC of *Model Healthy vs. pathology*: (**a**) when classifying coughs, (**b**) when classifying subject for respiratory pathology.

Table 4. AROC of *Healthy vs. pathology model.*

Model	Individual Cough Classification AROC	Respiratory Pathology Classification of Subject AROC Based on Entire Cough Epoch
Healthy vs. pathology Model	0.84	0.91

6.4.2. *Healthy vs. LRTI Model*, *Healthy vs. URTI Model*, *Healthy vs. Asthma Model*

The resulting cough classification accuracy and the respiratory pathology classification of subject accuracy when considering one respiratory pathology at a time is shown in Table 5. Again, the deep BiLSTM was able to produce good results when differentiating the healthy-voluntary coughs from those resulting from various respiratory conditions. This resulted in classification accuracy exceeding 85% for every investigated scenario. Respiratory pathology classification of subjects, as expected, result in even higher accuracy (exceeding 92% for every case).

Table 5. Experimental results in terms of accuracy for *Healthy vs. LRTI Model*, *Healthy vs. URTI Model*, *Healthy vs. Asthma Model*.

Model	Individual Cough Classification Accuracy (in %)	Respiratory Pathology Classification of Subject Accuracy Based on Entire Cough Epoch (in %)
Healthy vs. LRTI Model	86.3	94.5
Healthy vs. URTI Model	86.5	92.7
Healthy vs. Asthma Model	85.9	94.2

Confusion matrices were produced to further analyse the results from each of these models. Figures 8–10 show the confusion matrices for *Healthy vs. LRTI Model*, *Healthy vs. URTI Model* and *Healthy vs. Asthma Model*, respectively. The performance of *Healthy vs. LRTI Model* and *Healthy vs. Asthma Model* when it comes to correctly classifying healthy coughs from pathological coughs is comparable (see Figures 8a and 10a). *Healthy vs. URTI Model* has a slightly larger number of misclassifications when predicting healthy coughs; however, its performance on pathological coughs detection (URTI in this case) is better compared to the other two models (see Figure 9). When it comes to respiratory pathology classification of subject based on the entire cough epochs, as expected, the classification models have resulted in higher correct classification rate compared to the individual cough classification model (see Figures 8b, 9b and 10b).

Receiver operating characteristics were created for all three models, both for the case of cough and pathology classification. The ROCs are shown in Figures 11–13 and the resulting AROC is shown in Table 6. The AROC values are convincingly higher for all the pathology screening results (exceeding 93%) compared to the individual cough classification models. They support the finding from Table 5 and the corresponding confusion matrices.

Table 6. Area under the receiver operating curve (AROC) for *Healthy vs. LRTI Model*, *Healthy vs. URTI Model*, *Healthy vs. Asthma Model*.

Model	Individual Cough Classification AROC	Respiratory Pathology Classification of Subject AROC Based on Entire Cough Epoch
Healthy vs. LRTI Model	0.87	0.95
Healthy vs. URTI Model	0.87	0.93
Healthy vs. Asthma Model	0.86	0.94

LRTI Cough Sound Classification

	Healthy	LRTI
Healthy	90.5%	9.5%
LRTI	16.7%	83.3%

True Class / Predicted Class

(a)

Subject LRTI Pathology Classification

	Healthy	LRTI
Healthy	95.6%	4.4%
LRTI	6.4%	93.6%

True Class / Predicted Class

(b)
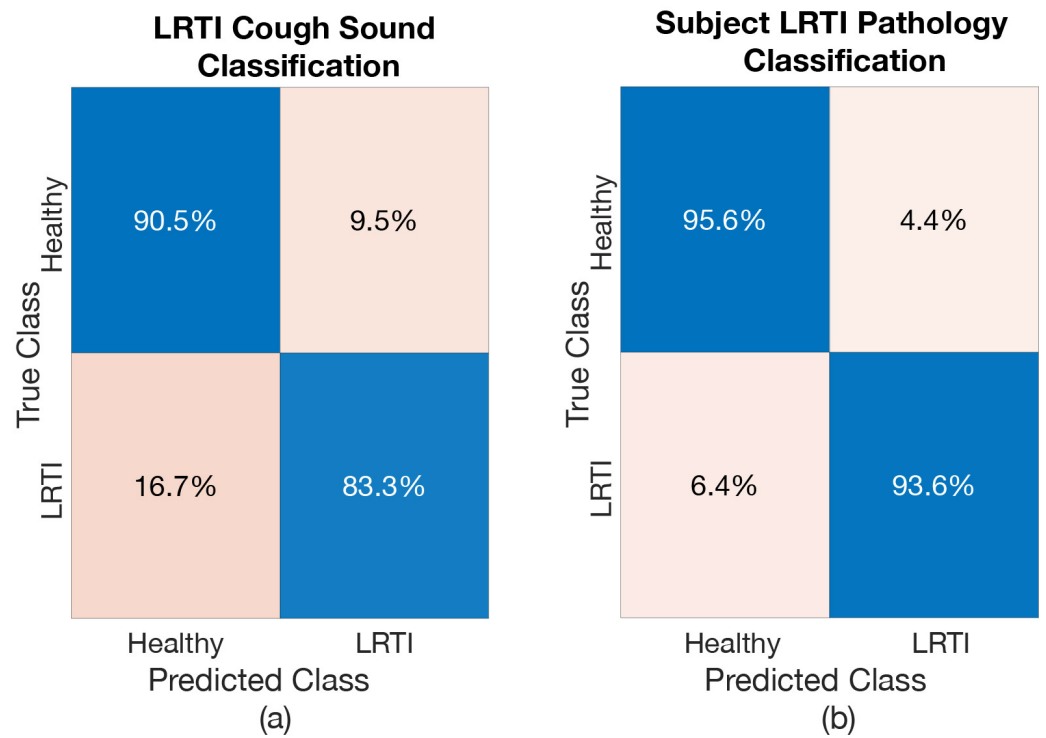
Figure 8. Confusion Matrix of *Healthy vs. LRTI Model*: (**a**) when classifying LRTI coughs, (**b**) when classifying subject for LRTI.

URTI Cough Sound Classification

	Healthy	URTI
Healthy	84.1%	15.9%
URTI	10.2%	89.8%

True Class / Predicted Class

(a)

Subject URTI Pathology Classification

	Healthy	URTI
Healthy	91.1%	8.9%
URTI	4.2%	95.8%

True Class / Predicted Class

(b)
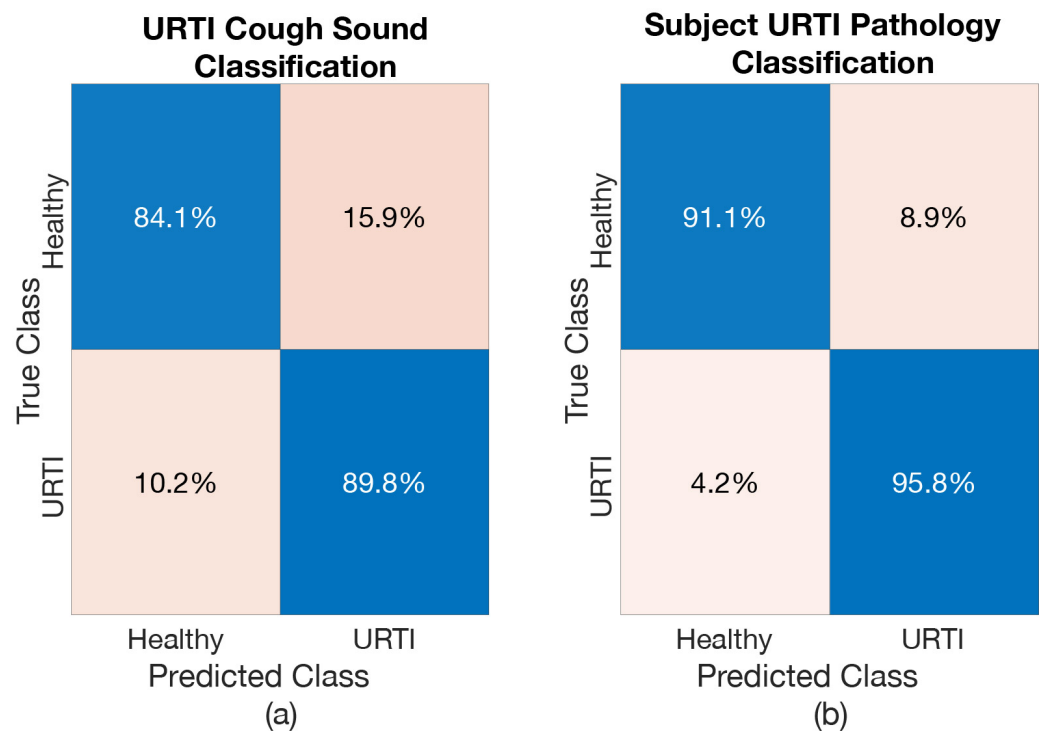
Figure 9. Confusion Matrix of *Healthy vs. URTI Model*: (**a**) when classifying URTI coughs, (**b**) when classifying subject for URTI.
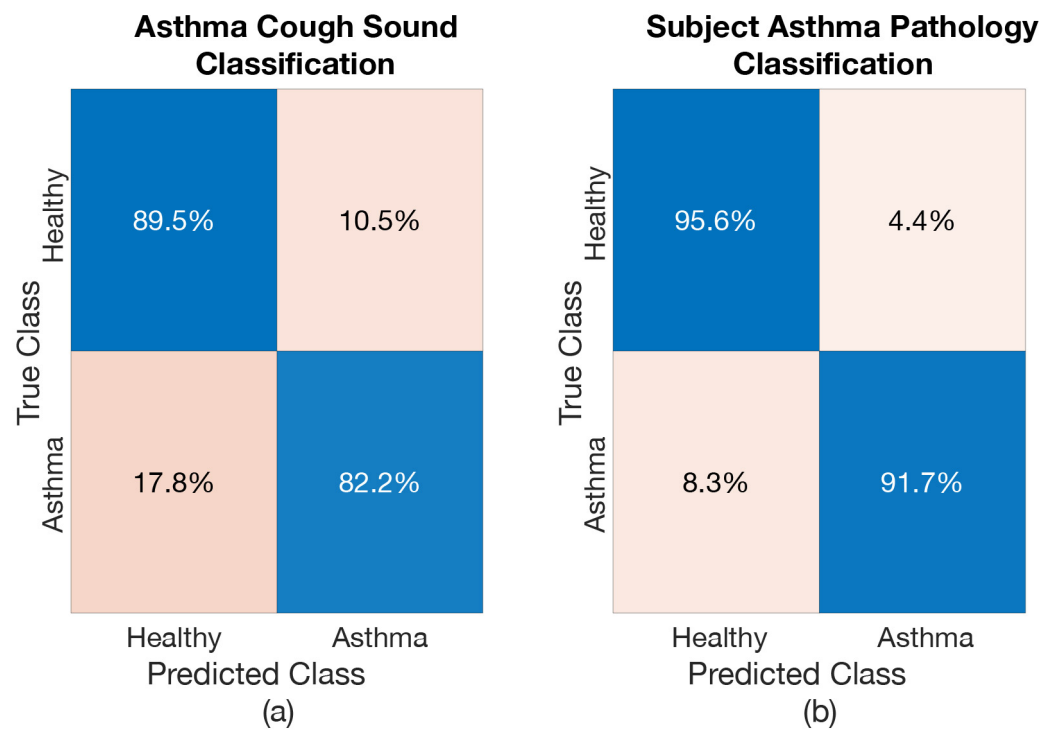
Figure 10. Confusion Matrix of *Healthy vs. Asthma Model*: (**a**) when classifying Asthmatic coughs, (**b**) when classifying subject for Asthma.
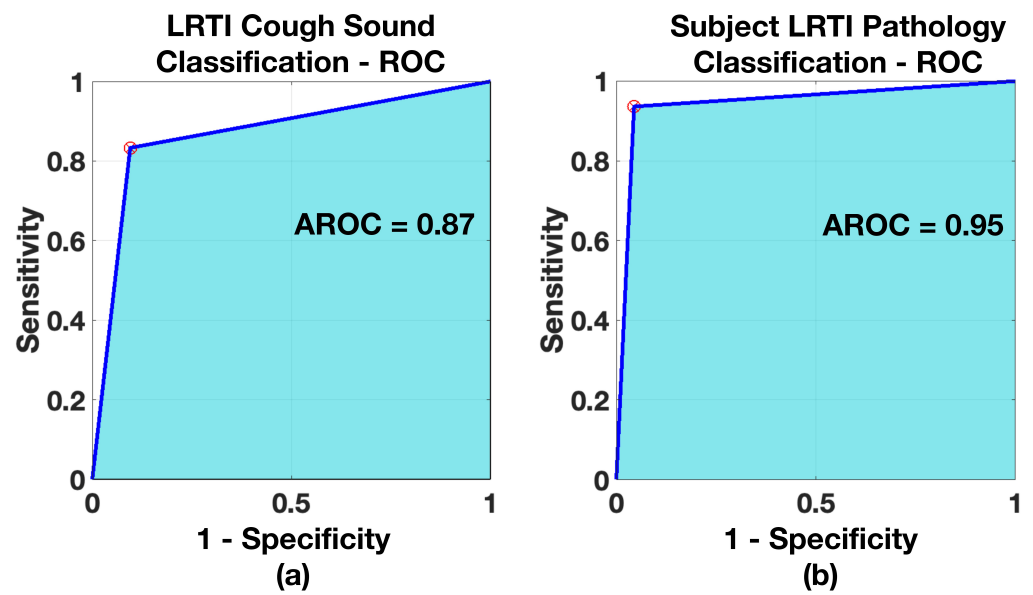

Figure 11. ROC—*Healthy vs. LRTI Model*: (**a**) when classifying LRTI coughs, (**b**) when classifying subject for LRTI.
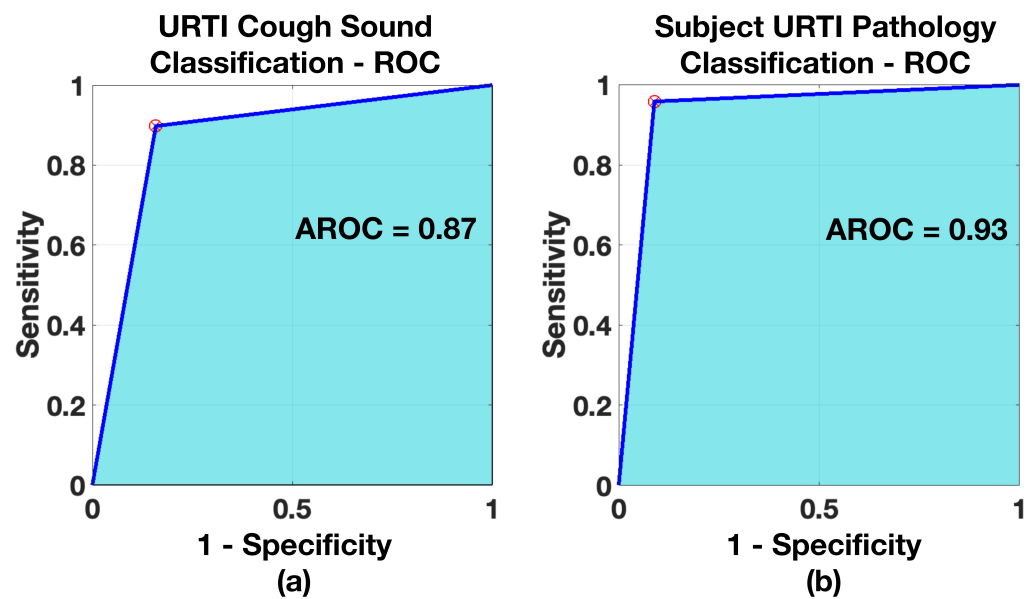
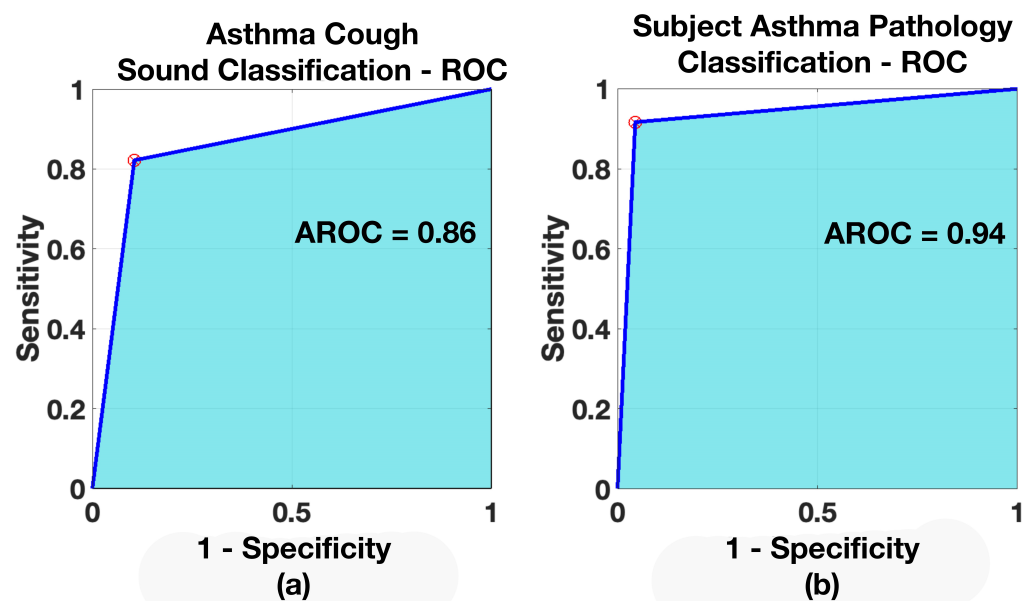**Figure 12.** ROC—*Healthy vs. URTI Model*: (**a**) when classifying URTI coughs, (**b**) when classifying subject for URTI.

Figure 13. ROC—*Healthy vs. Asthma Model*: (**a**) when classifying Asthmatic coughs, (**b**) when classifying subject for Asthma.

6.4.3. *Healthy vs. Pathology (4-Class) Model*

The resulting performance of the proposed model when trained to classify different respiratory pathological coughs and healthy-voluntary coughs (i.e., 4-Class model) is shown in Table 7. The subject respiratory pathology classification result for this four-class classification, based on the most frequent (mode) prediction outcome for all cough epochs of a subject, is shown in Table 8. The overall classification accuracy of both cough classification and each pathology classification is lower compared to the results shown in Tables 3 and 5. The classification accuracy for the healthy-voluntary cough class and the subsequent respiratory pathology classification is relatively high (71.2% and 84.4%, respectively). However, the classification accuracy of pathological cough classes is relatively low. The Asthma class has the highest misclassification rate among the three investigated respiratory conditions. The confusion matrices are shown to further understand this classification result (see Figure 14a). It is interesting to note in the respiratory pathology

classification results (see Figure 14b) that none of the subjects with LRTI and asthma are misclassified as healthy and only one subject with URTI is misclassified as healthy (4.2% out of 24 subjects with URTI tested will be one subject). However, seven healthy subjects were misclassified to have some kind of respiratory problems (of these seven, two were misclassified as having URTI, another two were misclassified as having LRTI and another three misclassified as having asthma). Among the three respiratory conditions, as mentioned earlier, asthma was the most misclassified pathology (15 subjects out of 24 with asthma were misclassified as having LRTI). Even though there is high misclassification rate among the three investigated respiratory conditions, in summary, this four-class classification model has a classification accuracy of 84.4% for correctly identifying healthy subjects and 95.8% accuracy for identifying subjects with respiratory issues , see Table 9.
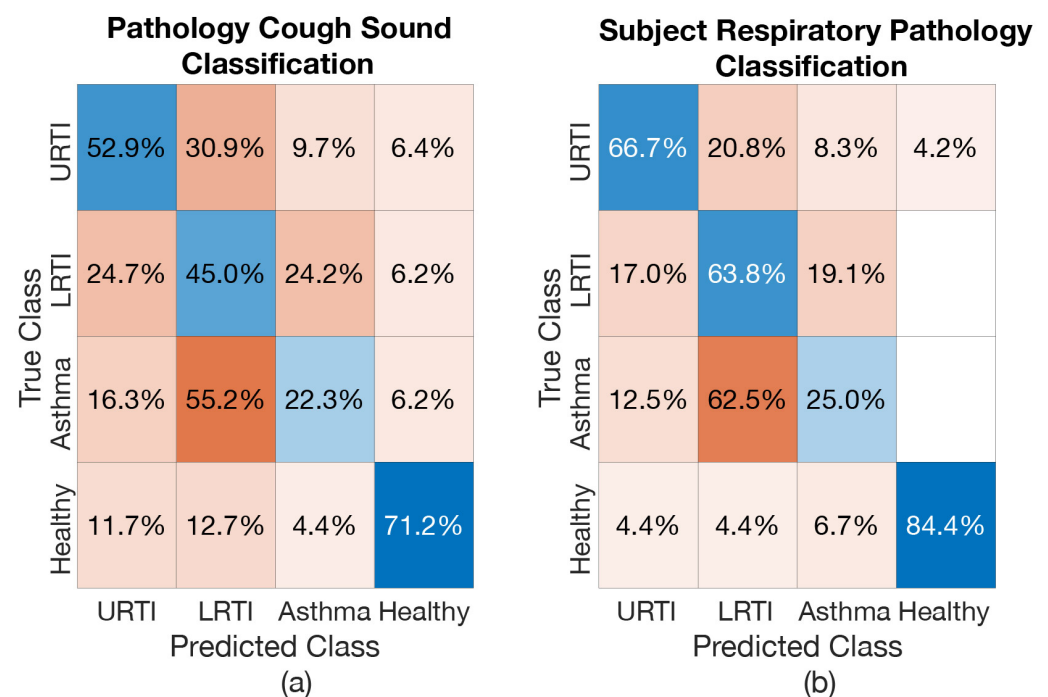

Figure 14. Confusion Matrix—*Healthy vs. Pathology (4-Class) Model*: (**a**) when screening coughs, (**b**) when screening for pathology of the subjects.

Table 7. Cough classification accuracy of *Healthy vs. Pathology 4-Class Model*.

Overall Cough Classification Accuracy (in %)	Healthy Cough Classification Accuracy (in %)	Asthma Cough Classification Accuracy (in %)	URTI Cough Classification Accuracy (in %)	LRTI Cough Classification Accuracy (in %)
47.9	71.2	22.3	52.9	45.0

Table 8. Pathology classification accuracy for *Healthy vs. Pathology 4-Class Model*.

Overall Respiratory Pathology Classification of Subject Accuracy Based on Entire Cough Epoch (in %)	Healthy Subject Classification Accuracy Based on Entire Cough Epoch (in %)	Asthmatic Subject Classification Accuracy Based on Entire Cough Epoch (in %)	URTI Subject Classification Accuracy Based on Entire Cough Epoch (in %)	LRTI Subject Classification Accuracy Based on Entire Cough Epoch (in %)
60.0	84.4	25.0	66.7	63.8

Table 9. Accuracy of *Healthy vs. Pathology 4-Class Model.*

Model	Healthy Subjects Classified as Healthy (in %)	Subjects With Respiratory Conditions and Classified to Have Some Kind of Respiratory Conditions (in %)
Healthy vs. pathology 4-Class Model	84.4	95.8

6.4.4. App Rendering

We expect that such cough classification methodology should be eventually applied to support clinicians "in the field", if at least as a simple triage or as a preliminary screening tool. However, explicit discussions of a smartphone-deployed application (App) are premature for the scope of the current paper. However, if allowed to speculate, we see two possible pathways towards implementation: (1) port the whole algorithm into the smartphone and perform all the computational heavy lifting using the smartphone hardware to generate the prediction result; (2) the app simply collects audio data (via the onboard microphone) and communicates with a centralized server to perform the prediction and the results are returned to the user. Both pathways have their operational considerations, such as processing hardware available on the smartphone (the developer must consider the number of floating-point operations needed to make the prediction), availability and connection speed of the Internet (a consideration if remote deployment in rural communities is expected), among other issues. For the current setup running in a NVIDIA TITAN Xp Graphics Card, it takes almost three hours to train a particular deep neural network model and requires less than half a second to perform the prediction for a particular cough sample (timings include audio preprocessing and feature extraction steps). Given the fact that a clinical usage scenario only needs to be "quasi-real time" (a few seconds delay is usually tolerated—clinicians are accustomed to waiting longer for other screening tests), the second approach seems prudent for contexts with ready internet connection, so that the App would be lighter in terms of mobile phone hardware usage.

7. Conclusions

A classifier was developed based on a BiLSTM model trained using Mel-Frequency Cepstral Coefficient features that can differentiate cough sounds from healthy children with no active respiratory pathology to those with active pathological respiratory conditions such as asthma, URTI and LRTI. Four classifiers were trained as part of this investigation. The resulting trained model that classifies cough sounds into healthy/pathological in general or healthy/belonging to LRTI, URTI and asthma resulted in classification accuracy exceeding 84% when predicting a clinician's diagnosis. When a respiratory pathology classification of subject was performed using the mode of the prediction results across the multiple cough epochs from a particular subject, the resulting classification accuracy exceeded 91%. The classification accuracy of the model was compromised when trained to classify all the four classes of cough categories in one shot. However, most of the misclassification happened within the pathological classes where one class of pathological cough was often misclassified as having another pathology. If one ignores such misclassification and considers healthy cough to be that from a healthy subject and pathological cough to have come from subject with some kind of pathology, then the overall accuracy of the classifier is above 84%. This is a first step towards developing a highly efficient deep neural network model that can differentiate between different pathological cough sounds. Such a model could support physicians in creating a differential screening of respiratory conditions that present with cough, and will thus add value to health status monitoring and triaging in medical care, and potentially be deployed to support tele-medicine in remote and developing communities.

References

1. Shields, M.D.; Bush, A.; Everard, M.L.; McKenzie, S.; Primhak, R. Recommendations for the assessment and management of cough in children. *Thorax* **2008**, *63*, iii1–iii15. [CrossRef]
2. Shields, M.D.; Thavagnanam, S. The difficult coughing child: Prolonged acute cough in children. *Cough* **2013**, *9*, 11. [CrossRef]
3. Oren, E.; Rothers, J.; Stern, D.A.; Morgan, W.J.; Halonen, M.; Wright, A.L. Cough during infancy and subsequent childhood asthma. *Clin. Exp. Allergy* **2015**, *45*, 1439–1446. [CrossRef]
4. Abaza, A.A.; Day, J.B.; Reynolds, J.S.; Mahmoud, A.M.; Goldsmith, W.T.; McKinney, W.G.; Petsonk, E.L.; Frazer, D.G. Classification of voluntary cough sound and airflow patterns for detecting abnormal pulmonary function. *Cough* **2009**, *5*, 8. [CrossRef] [PubMed]
5. Murata, A.; Taniguchi, Y.; Hashimoto, Y.; Kaneko, Y.; Takasaki, Y.; Kudoh, S. Discrimination of productive and non-productive cough by sound analysis. *Intern. Med.* **1998**, *37*, 732–735. [CrossRef] [PubMed]
6. Abeyratne, U.R.; Swarnkar, V.; Setyati, A.; Triasih, R. Cough sound analysis can rapidly diagnose childhood pneumonia. *Ann. Biomed. Eng.* **2013**, *41*, 2448–2462. [CrossRef] [PubMed]
7. Swarnkar, V.; Abeyratne, U.; Tan, J.; Ng, T.W.; Brisbane, J.M.; Choveaux, J.; Porter, P. Stratifying asthma severity in children using cough sound analytic technology. *J. Asthma* **2021**, *58*, 160–169. [CrossRef] [PubMed]
8. Schröder, J.; Anemiiller, J.; Goetze, S. Classification of human cough signals using spectro-temporal Gabor filterbank features. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6455–6459.
9. Botha, G.; Theron, G.; Warren, R.; Klopper, M.; Dheda, K.; Van Helden, P.; Niesler, T. Detection of tuberculosis by automatic cough sound analysis. *Physiol. Meas.* **2018**, *39*, 045005. [CrossRef] [PubMed]
10. Nemati, E.; Rahman, M.M.; Nathan, V.; Vatanparvar, K.; Kuang, J. A comprehensive approach for cough type detection. In Proceedings of the 2019 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Arlington, VA, USA, 25–29 September 2019; pp. 15–16.
11. Sharma, N.; Krishnan, P.; Kumar, R.; Ramoji, S.; Chetupalli, S.R.; Ghosh, P.K.; Ganapathy, S. Coswara—A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. *arXiv* **2020**, arXiv:2005.10548.
12. Cohen-McFarlane, M.; Goubran, R.; Knoefel, F. Novel coronavirus cough database: Nococoda. *IEEE Access* **2020**, *8*, 154087–154094. [CrossRef]
13. Orlandic, L.; Teijeiro, T.; Atienza, D. The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. *arXiv* **2020**, arXiv:2009.11644.
14. Wei, W.; Wang, J.; Ma, J.; Cheng, N.; Xiao, J. A Real-time Robot-based Auxiliary System for Risk Evaluation of COVID-19 Infection. *arXiv* **2020**, arXiv:2008.07695.
15. Imran, A.; Posokhova, I.; Qureshi, H.N.; Masood, U.; Riaz, M.S.; Ali, K.; John, C.N.; Hussain, M.I.; Nabeel, M. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform. Med. Unlocked* **2020**, *20*, 100378. [CrossRef]
16. Laguarta, J.; Hueto, F.; Subirana, B. COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *IEEE Open J. Eng. Med. Biol.* **2020**, *1*, 275–281. [CrossRef]
17. Bagad, P.; Dalmia, A.; Doshi, J.; Nagrani, A.; Bhamare, P.; Mahale, A.; Rane, S.; Agarwal, N.; Panicker, R. Cough against COVID: Evidence of COVID-19 signature in cough sounds. *arXiv* **2020**, arXiv:2009.08790.
18. Brown, C.; Chauhan, J.; Grammenos, A.; Han, J.; Hasthanasombat, A.; Spathis, D.; Xia, T.; Cicuta, P.; Mascolo, C. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 6–10 July 2020; pp. 3474–3484.

19. Wang, H.H.; Liu, J.M.; You, M.; Li, G.Z. Audio signals encoding for cough classification using convolutional neural networks: A comparative study. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; pp. 442–445.

20. Barry, S.J.; Dane, A.D.; Morice, A.H.; Walmsley, A.D. The automatic recognition and counting of cough. *Cough* **2006**, *2*, 1–9. [CrossRef] [PubMed]

21. Stegmaier-Stracca, P.A.; Tschichold-Gürman, N.N. Cough detection using fuzzy classification. In Proceedings of the 1995 ACM Symposium on Applied Computing, Nashville, TN, USA, 26–28 February 1995; pp. 440–444

22. Amoh, J.; Odame, K. DeepCough: A deep convolutional neural network in a wearable cough detection system. In Proceedings of the 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS), Atlanta, GA, USA, 22–24 October 2015; pp. 1–4.

23. Nemati, E.; Rahman, M.M.; Nathan, V.; Kuang, J. Private audio-based cough sensing for in-home pulmonary assessment using mobile devices. In *EAI International Conference on Body Area Networks*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 221–232.

24. Tracey, B.H.; Comina, G.; Larson, S.; Bravard, M.; López, J.W.; Gilman, R.H. Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; pp. 6017–6020

25. Larson, E.C.; Lee, T.; Liu, S.; Rosenfeld, M.; Patel, S.N. Accurate and privacy preserving cough sensing using a low-cost microphone. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; pp. 375–384.

26. Barata, F.; Tinschert, P.; Rassouli, F.; Steurer-Stey, C.; Fleisch, E.; Puhan, M.A.; Brutsche, M.; Kotz, D.; Kowatsch, T. Automatic recognition, segmentation, and sex assignment of nocturnal asthmatic coughs and cough epochs in smartphone audio recordings: Observational field study. *J. Med. Internet Res.* **2020**, *22*, e18082. [CrossRef] [PubMed]

27. Monge-Álvarez, J.; Hoyos-Barceló, C.; Lesso, P.; Casaseca-de-la Higuera, P. Robust detection of audio-cough events using local hu moments. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 184–196. [CrossRef]

28. Hee, H.I.; Balamurali, B.; Karunakaran, A.; Herremans, D.; Teoh, O.H.; Lee, K.P.; Teng, S.S.; Lui, S.; Chen, J.M. Development of Machine Learning for Asthmatic and Healthy Voluntary Cough Sounds: A Proof of Concept Study. *Appl. Sci.* **2019**, *9*, 2833. [CrossRef]

29. BT, B.; Hee, H.I.; Teoh, O.; Lee, K.; Kapoor, S.; Herremans, D.; Chen, J.M. Asthmatic versus healthy child classification based on cough and vocalised /aː/sounds. *J. Acoust. Soc. Am.* **2020**, *148*, EL253–EL259.

30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

31. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.

32. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]

33. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef]

34. Bt, B.; Lin, K.; Lui, S.; Chen, J.; Herremans, D. Towards robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access* **2019**, *7*, 84229–84241. [CrossRef]

35. Muda, L.; Km, B.; Elamvazuthi, I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *J. Comput.* **2010**, *2*, 138–143.

36. Rabiner, L.R.; Schafer, R.W. *Theory and Applications of Digital Speech Processing*; Pearson: Upper Saddle River, NJ, USA, 2011; Volume 64.

37. Kawakami, Y.; Wang, L.; Kai, A.; Nakagawa, S. Speaker identification by combining various vocal tract and vocal source features. In *International Conference on Text, Speech, and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 382–389.

38. Brown, C.D.; Davis, H.T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 24–38. [CrossRef]

39. Tarca, A.L.; Carey, V.J.; Chen, X.W.; Romero, R.; Drăghici, S. Machine learning and its applications to biology. *PLoS Comput. Biol.* **2007**, *3*, e116. [CrossRef]

40. McGill, R.; Tukey, J.W.; Larsen, W.A. Variations of box plots. *Am. Stat.* **1978**, *32*, 12–16.

41. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]