Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)
20-22 July 2022, Prague-Czech Republic

2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9873469

Automatic Tuberculosis and COVID-19 cough classification using deep learning

Madhurananda Pahar
Department of Electrical and Electronic Engineering, Stellenbosch University,
Stellenbosch, South Africa.
mpahar@sun.ac.za

Marisa Klopper
Division of Molecular Biology and Human Genetics, Stellenbosch University,
Stellenbosch, South Africa.
marisat@sun.ac.za

Byron Reeve
Division of Molecular Biology and Human Genetics, Stellenbosch University,
Stellenbosch, South Africa.
byronreeve@sun.ac.za

Rob Warren
Division of Molecular Biology and Human Genetics, Stellenbosch University,
Stellenbosch, South Africa.
rw1@sun.ac.za

Grant Theron
Division of Molecular Biology and Human Genetics, Stellenbosch University,
Stellenbosch, South Africa.
gtheron@sun.ac.za

Andreas Diacon
TASK Applied Science,
Cape Town, South Africa.
ahd@task.org.za

Thomas Niesler
Department of Electrical and Electronic Engineering, Stellenbosch University,
Stellenbosch, South Africa.
trn@sun.ac.za

***Abstract*—We present a deep learning based automatic cough classifier which can discriminate tuberculosis (TB) coughs from COVID-19 coughs and healthy coughs. Both TB and COVID-19 are respiratory diseases, contagious, have cough as a predominant symptom and claim thousands of lives each year. The cough audio recordings were collected at both indoor and outdoor settings and also uploaded using smartphones from subjects around the globe, thus containing various levels of noise. This cough data include 1.68 hours of TB coughs, 18.54 minutes of COVID-19 coughs and 1.69 hours of healthy coughs from 47 TB patients, 229 COVID-19 patients and 1498 healthy patients and were used to train and evaluate a CNN, LSTM and Resnet50. These three deep architectures were also pre-trained on 2.14 hours of sneeze, 2.91 hours of speech and 2.79 hours of noise for improved performance. The class-imbalance in our dataset was addressed by using SMOTE data balancing technique and using performance metrics such as F1-score and AUC. Our study shows that the highest F1-scores of 0.9259 and 0.8631 have been achieved from a pre-trained Resnet50 for two-class (TB vs COVID-19) and three-class (TB vs COVID-19 vs healthy) cough classification tasks, respectively. The application of deep transfer learning has improved the classifiers' performance and makes them more robust as they generalise better over the cross-validation folds. Their performances exceed the TB triage test requirements set by the world health organisation (WHO). The features producing the best performance contain higher order of MFCCs suggesting that the differences between TB and COVID-19 coughs are not perceivable by the human ear. This type of cough audio classification is non-contact, cost-effective and can easily be deployed on a smartphone, thus it can be an excellent tool for both TB and COVID-19 screening.**

***Index Terms*—tuberculosis, COVID-19, cough, transfer learning, deep learning, Resnet50**

I. Introduction

Tuberculosis (TB) is a bacterial infectious disease which affects the human lungs, prevalent in low-income settings and 95% of all TB cases are reported in developing countries [1], [2]. Modern diagnostic tests are costly as they rely on special equipment and laboratory procedure [3]–[5]. Suspected patients are tested when they show the symptom criteria of TB investigation and the results indicate that most of them cough due to other lung ailments; in fact most of those TB-suspected patients do not suffer from TB [6].

COVID-19 (**CO**rona **VI**rus **D**isease of 20**19**) was declared as a global pandemic on February 11, 2020 by the World Health Organisation (WHO). At the time of writing, there are 513.9 million COVID-19 global cases and sadly, the pandemic has claimed the life of 6.2 million [7]. Thus, many suspected TB patients are very likely to be suffering from COVID-19 in developing countries and experimental evidence suggests that healthy people cough less than those who are sick from lung ailments [8]. Therefore, there is a need for automated non-contact, low-cost, easily-accessible tools for both TB and COVID-19 screening based on cough audio.

One of the major symptoms of respiratory diseases like TB and COVID-19 is a cough [9], [10]. Depending on the nature

978-1-6654-7087-2/22/$31.00 © 2022 IEEE

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

of the respiratory disease, the airway is to be either blocked or restricted and this can affect the acoustic properties of the coughs, thus enabling the cough audio to be used by machine learning algorithms in many studies including our own [11]–[13] for discriminating both TB [14] and COVID-19 [15] from healthy coughs. As TB is mostly found in developing countries, the efforts to collect TB coughs are rare, thus TB cough data are small and not publicly available. Successful studies [8], [12], [16] have experimentally found that shallow classifiers such as a multilayer perceptron (MLP) or logistic regression (LR) model works well in detecting TB in cough audio. However, COVID-19 cough data are widely available [17]–[19] and many recent studies have successfully applied neural networks [20] including deep neural network (DNN) classifiers to detect COVID-19 in cough audio [15], [21], [22].

In this study, we present a deep learning based automatic cough classifier which discriminates TB coughs from COVID-19 coughs. We have used both public and private datasets and synthetic minority over-sampling technique (SMOTE) to create new datapoints to balance the datasets, as COVID-19 coughs are under-represented in our datasets. We have also used both AUC (area under the ROC curve) and F1-score as the performance metric for our three DNN classifiers: CNN, LSTM & Resnet50 and evaluated them using nested cross-validation to make the best use of our datasets. The highest F1-score of 0.9042 has been achieved from a Resnet50 classifier in discriminating TB coughs from COVID-19 coughs. Inspired by our previous research [13], we have made use of sneeze, speech and noise to pre-train these three deep architectures as well. This has improved the F1-score of this two-class classification task to 0.9259 with more robust performance across the cross-validation folds. The corresponding AUC has been 0.9245 with a 96% sensitivity at 80% specificity, exceeding the TB triage test requirement of 90% sensitivity at 70% specificity determined by WHO. We have further investigated these three DNN classifiers' performances in a three-class classification task, where we added healthy coughs as the third class. Initially, an F1-score of 0.8578 has been achieved from the Resnet50 and it has been improved to 0.8631 from the same architecture in discriminating TB, COVID-19 and healthy coughs after applying the transfer learning.

Section II will detail the datasets used for pre-training the DNN classifiers and the datasets used for both two-class and three-class classification and fine-tuning those three classifiers. Section III explains the features extracted from the audio and Section IV describes the classification and hyperparameter optimisation process. Section V summarises the results and Section VI discusses them. Finally, Section VII concludes this study.

II. Data

We have made use of both public and private data in this study. Previously, we have compiled TASK, Sarcos, Brooklyn, and Wallacedene datasets as part of the research projects concerning cough monitoring and cough classification. Coswara, ComParE, Google Audio Set & Freesound and LibriSpeech were compiled from publicly available data.

Coughs with labels 'TB', 'COVID-19' and 'healthy' are used for the classification task. Coughs were excluded from the data used for pre-training altogether as coughs without these three labels may originate from other diseases and we only classified disease in both classification (two-class and three-class) task and fine-tuning the pre-trained DNN classifiers on cough audio. All recordings were downsampled to 16 kHz.

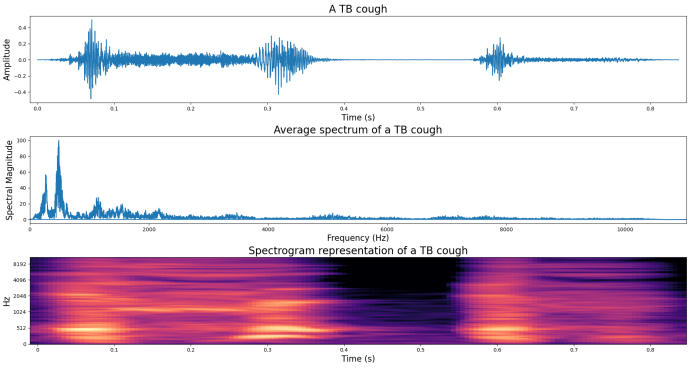

Fig. 1. The audio and spectrogram of a TB positive cough.

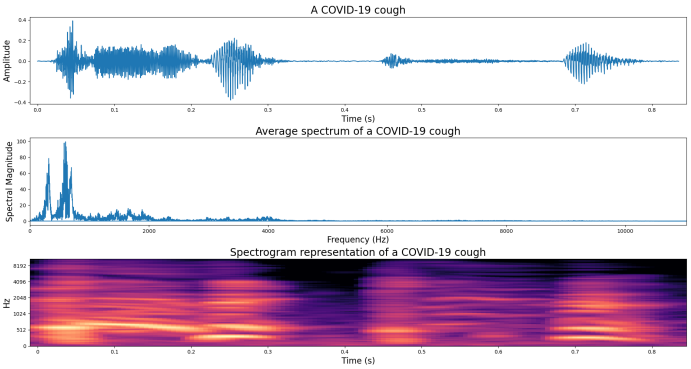

Fig. 2. The audio and spectrogram of a COVID-19 positive cough.

A. Cough audio data for classification

The following six datasets of coughs with TB, COVID-19 and healthy labels were available for experimentation and are described in Table I. A simple energy detector was applied to pre-process the audio recordings of Coswara, Sarcos and ComParE datasets by removing silence within a margin of 50 msec [11].

1) TASK dataset: This dataset contains 6000 continuous cough recordings and 11393 non-cough events such as laughter, doors opening and objects moving. It was collected at TASK, a TB research centre near Cape Town, South Africa from patients undergoing TB treatment [23]. Previous research indicates that cough-frequency decreases as patients' health conditions improve [24]. Thus, the TASK dataset was compiled to develop cough detection algorithms and

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

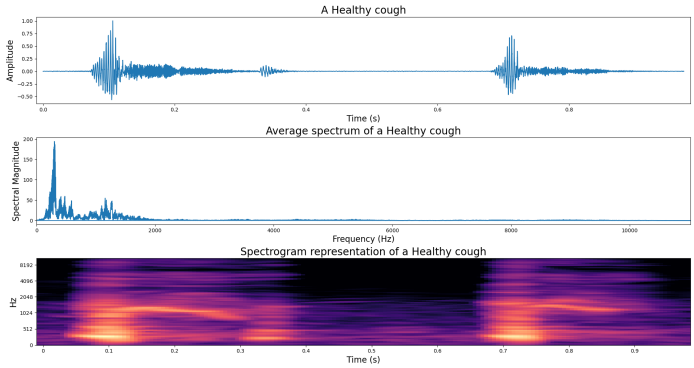

Fig. 3. The audio and spectrogram of a healthy cough which are not much different from those of the TB and COVID-19 coughs, shown in Fig. 1 and Fig. 2. A subjective test couldn't differentiate the sick coughs from the healthy coughs or diagnose the disease.

monitor patients' long-term health recovery in a multi-bed ward environment using an external microphone attached to a smartphone [25].

2) Sarcos dataset: This dataset was collected in South Africa as part of our own COVID-19 research [11], [13] and contains coughs from 18 COVID-19 positive subjects.

3) Brooklyn dataset: Cough audio was compiled from 17 TB and 21 healthy subjects to discriminate TB from healthy cough for developing a TB cough audio classifier [8]. The recordings were taken inside a controlled indoor booth, using an audio field recorder and a RØDE M3 microphone.

4) Wallacedene dataset: This dataset, containing 402 coughs from 16 TB patients, was collected to extend the previous TB cough audio classification study [8] to discriminate TB coughs from other sick coughs in a real-world noisy environment [12]. Here, the cough recordings were collected using an audio field recorder and a RØDE M1 microphone and the recording process took place in an outdoor booth, located next to a busy street [26].

5) Coswara dataset: This publicly available dataset (https://coswara.iisc.ac.in) is compiled to develop machine learning algorithms for the diagnosis of COVID-19 in vocal audio [18], [27], [28]. Participants from five different continents contributed their vocal audio including coughs using their smartphones. In this study, we used the deep coughs from 92 COVID-19 positives and 1079 healthy subjects.

6) ComParE dataset: This dataset was presented in the 2021 Interspeech Computational Paralinguistics ChallengE (ComParE) [19] and it contains 119 COVID-19 positives and 398 healthy subjects.

7) Summary of data used for disease classification: Table I demonstrates that our data contain only 18.54 minutes of COVID-19 cough audio, compared to 1.68 hours of TB coughs and 1.69 hours of healthy coughs, indicating COVID-19 labelled data are under-represented. As such a data-imbalance can affect the neural networks' performance negatively [29], [30], we have applied SMOTE [31]. This data balancing technique creates new synthetic samples to oversample the minor class during training. SMOTE has been successfully applied to address training set class imbalances in cough detection [32] and cough classification [11] in the past. TASK dataset contains only 14 patients but the length of cough audio per patient was much longer than the other two datasets.

The audio and spectrograms of a TB, COVID-19 and healthy cough are shown in Fig. 1, Fig. 2 and Fig. 3. There are very little obvious visual differences between these three coughs. An informal subjective test was conducted where approximately 20 university students were asked to spot the sick and healthy coughs just by listening to these cough audio and the results showed that human auditory system is unable to spot any disease or differentiate sick coughs from healthy coughs only by listening to the coughs.

B. Datasets without cough labels for pre-training

Our classifier training is limited as cough audio data is not available abundantly. Hence, we use three other types of audio data for pre-training and they include sneeze, speech and noise from Google Audio Set & Freesound, LibriSpeech and TASK datasets, as described in Table II.

1) Google Audio Set & Freesound: The Google Audio Set has manually-labelled excerpts from 1.8 million Youtube videos belonging to 632 audio event categories [33]. The Freesound audio database contains tagged audio with various noise levels, uploaded by subjects from many parts of the world under widely varying recording conditions [34]. From these audio, We have selected the recordings that include 1013 sneezes, 2326 speech excerpts and 1027 other non-vocal sounds such as restaurant chatter, running water and engine noise. This manually annotated dataset was successfully used in developing cough detection algorithms [35].

2) LibriSpeech: The LibriSpeech corpus [36] is freely available and contains very little noise. We have carefully selected utterances from 28 female and 28 male speakers.

3) Summary of data used for pre-training: In total, the data described in Table II includes 1013 sneezing sounds (13.34 minutes of audio), 2.91 hours of speech, and 2.98 hours of noise. As sneezing is under-represented, we have again applied SMOTE to create additional synthetic samples. In total, a dataset containing 7.84 hours of audio recordings with three class labels (sneeze, speech, noise) was used to pre-train three DNN classifiers.

III. Feature Extraction

Mel-frequency cepstral coefficients (MFCCs) along with their velocity and acceleration coefficients, zero-crossing rate and kurtosis [37] were extracted from the audio recordings and these features were used for both classification and pre-training task. The feature combination containing MFCCs rather than linearly-spaced log filerbanks [8] showed better performance in our previous TB [12] and COVID-19 [11], [32] classification tasks. MFCCs are the features of choice in detecting and classifying voice audio such as speech [38]–[40] and coughs [35].

Overlapping frames were used to extract features, where the frame overlap ensures that a certain exact number of

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

TABLE I:
DATASETS USED IN COUGH CLASSIFICATION. THESE DATASETS CONTAIN THREE COUGH CLASSES: TB, COVID-19 AND HEALTHY.

Type	Dataset	Sampling rate	No of subjects	Total audio	Average length	Standard deviation
TB Cough	TASK	44.1 kHz	14	91 mins	6.5 mins	1.23 mins
	Brooklyn	44.1 kHz	17	4.63 mins	16.35 sec	13 sec
	Wallacedene	44.1 kHz	16	4.98 mins	18.69 sec	4.95 sec
	Total (TB Cough)	**—**	**47**	**1.68 hours**	**2.14 min**	**28.37 sec**
COVID-19 Cough	Coswara	44.1 kHz	92	4.24 mins	2.77 sec	1.62 sec
	ComParE	16 kHz	119	13.43 mins	6.77 sec	2.11 sec
	Sarcos	44.1 kHz	18	0.87 mins	2.91 sec	2.23 sec
	Total (COVID-19 Cough)	**—**	**229**	**18.54 mins**	**4.85 sec**	**1.92 sec**
Healthy Cough	Coswara	44.1 kHz	1079	0.98 hours	3.26 sec	1.66 sec
	ComParE	16 kHz	398	40.89 mins	6.16 sec	2.26 sec
	Brooklyn	44.1 kHz	21	1.66 mins	4.7 sec	3.9 sec
	Total (Healthy Cough)	**—**	**1498**	**1.69 hours**	**4.05 sec**	**1.85 sec**

TABLE II:
DATASETS USED IN PRE-TRAINING. DNN CLASSIFIERS ARE PRE-TRAINED ON 7.84 HOURS OF AUDIO DATA WITH THREE CLASS LABELS: SNEEZE, SPEECH AND NOISE. THESE DATA DO NOT CONTAIN ANY COUGH.

Type	Dataset	Sampling rate	No of events	Total audio	Average length	Standard deviation
Sneeze	Google Audio Set & Freesound	16 kHz	1013	13.34 mins	0.79 sec	0.21 sec
	Google Audio Set & Freesound + SMOTE	16 kHz	9750	2.14 hours	0.79 sec	0.23 sec
	Total (Sneeze)	**—**	**10763**	**2.14 hours**	**0.79 sec**	**0.23 sec**
Speech	Google Audio Set & Freesound	16 kHz	2326	22.48 mins	0.58 sec	0.14 sec
	LibriSpeech	16 kHz	56	2.54 hours	2.72 mins	0.91 mins
	Total (Speech)	**—**	**2382**	**2.91 hours**	**4.39 sec**	**0.42 sec**
Noise	TASK dataset	44.1 kHz	12714	2.79 hours	0.79 sec	0.23 sec
	Google Audio Set & Freesound	16 kHz	1027	11.13 mins	0.65 sec	0.26 sec
	Total (Noise)	**—**	**13741**	**2.79 hours**	**0.79 sec**	**0.23 sec**

frames always represents the entire audio event. This way an image-like fixed input dimension feature matrix can be computed where the general overall temporal structure of the sound are also maintained. Such fixed two-dimensional features have been successfully used to train DNN classifiers in our previous experiments [11], [32].

The dimension of the input feature matrix has been $(3\mathbb{M}+2, \mathbb{S})$ for $\mathbb{M}$ MFCCs. The frame length ($\mathbb{F}$), exact number of frames ($\mathbb{S}$) and number of lower order MFCCs ($\mathbb{M}$) are used as the feature extraction hyperparameters, mentioned in Table III. The table shows that $\mathbb{M}$ lies between 13 and 65, which varies the spectral information in each audio event and each audio event is divided into between 70 and 200 frames, as different phases of coughs carry different information. Each frame consists of between 512 and 4096 samples, i.e. 32 msec and 256 msec of audio, as the sampling rate is 16 kHz in our experiments.

TABLE III:
FEATURE EXTRACTION HYPERPARAMETERS. BY VARYING THE MFCCS BETWEEN 13 & 65, FRAME LENGTHS BETWEEN 512 & 4096 AND NO OF FRAMES BETWEEN 70 & 200, THE SPECTRAL RESOLUTIONS OF THE AUDIO HAVE BEEN VARIED OVER A LARGE RANGE.

Hyperparameters	Description	Range
MFCCs ($\mathbb{M}$)	lower order MFCCs to keep	$13 \times k$, where $k = 1, 2, 3, 4, 5$
Frame length ($\mathbb{F}$)	into which audio is segmented	2^k, where $k = 9, 10, 11, 12$
Segments ($\mathbb{S}$)	no. of frames extracted from audio	$10 \times k$, where $k = 7, 10, 12, 15$

IV. Classification Process

A. Classifier Architectures

We have used only three DNN classifiers: CNN [41], LSTM [42] and Resnet50 [43] in this study. We have refrained from experimenting with any shallow classifier as using deep architectures along with SMOTE data balancing technique yielded better results in our previous experiments [11], [32]. For our initial set of experiments, we have used these three DNN classifiers for two-class (TB vs COVID-19) and three-class (TB vs COVID-19 and healthy) classification and the classifier hyperparameters are mentioned in Table IV. The classifier training process is stopped when the performance wasn't improved after 10 epochs. Finally, for the improved performance, we have applied the transfer learning.

B. Transfer Learning Architectures

The application of transfer learning has improved the classification performance in our previous studies [11], [13]. Hence, we have also applied transfer learning in this study to improve the classification performance, where the DNN classifiers are pre-trained on the dataset, explained in Section II-B, and then fine-tuned on the classification datasets, explained in Section II-A. The feature extraction hyperparameters are adopted from

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

TABLE IV:
CLASSIFIER HYPERPARAMETERS, OPTIMISED USING LEAVE-p-OUT NESTED CROSS-VALIDATION SCHEME.

Hyperparameters	Classifier	Range
No. of conv filters (α_1)	CNN	3×2^k where $k = 3, 4, 5$
Kernel size (α_2)	CNN	2 and 3
Dropout rate (α_3)	CNN, LSTM	0.1 to 0.5 in steps of 0.2
Dense layer units (α_4)	CNN, LSTM	2^k where $k = 4, 5$
LSTM units (α_5)	LSTM	2^k where $k = 6, 7, 8$
Learning rate (α_6)	LSTM	10^k where, $k = -2, -3, -4$
Batch Size (α_7)	CNN, LSTM	2^k where $k = 6, 7, 8$

our previous studies [11], [13] and the hyperparameters of the CNN and LSTM were determined during the cross-validation process. These hyperparameters are mentioned in Table V. A standard Resnet50, as explained in Table 1 of [43], with 512-unit dense layer has been used for the transfer learning. The transfer learning process for a CNN is explained in Fig. 4.

TABLE V:
FEATURE EXTRACTION AND CLASSIFIER HYPERPARAMETERS OF THE PRE-TRAINED NETWORKS: WE USED THE SAME FEATURE EXTRACTION HYPERPARAMETERS USED IN OUR PREVIOUS WORK [11], [13], WHILE CLASSIFIER HYPERPARAMETERS WERE OPTIMISED ON THE PRE-TRAINING DATA (TABLE II) USING THE NESTED CROSS-VALIDATION.

FEATURE EXTRACTION HYPERPARAMETERS		
Hyperparameters		**Values**
$\mathbb{M}$	MFCCs	39
$\mathbb{F}$	frame length	$2^{10} = 1024$
$\mathbb{S}$	no. of frames	150
CLASSIFIER HYPERPARAMETERS		
Hyperparameters	**Classifier**	**Values**
No. of conv filters (α_1)	CNN	256 & 128 & 64
Kernel size (α_2)	CNN	2
Dropout rate ((α_3))	CNN, LSTM	0.3
Dense layer units (α_4) (for pre-training)	CNN, LSTM, Resnet50	512 & 128 & 3
Dense layer units (α_4) (for fine-tuning)	CNN, LSTM, Resnet50	16 & 2 or 3
LSTM units (α_5)	LSTM	512 & 256 & 128
Learning rate (α_6)	LSTM	$10^{-3} = 0.001$
Batch Size (α_7)	CNN, LSTM, Resnet50	$2^7 = 128$

C. Hyperparameter optimisation

The feature extraction process and classifiers have a number of hyperparameters, listed in Table III and IV. They were optimised by using a leave-p-out cross-validation scheme [44]. The train and test split ratio was 4:1, due to its effectiveness in medical applications [45]. This 5-fold cross-validation process ensured the best use of our dataset by using all subjects in both training and testing the classifiers and implementing a strict no patient-overlap between cross-validation folds.

D. Classifier evaluation

The F1-score has been the optimisation criteria in the cross-validation folds and the performance-indicator of the classifiers [46]. We note the mean per-frame probability that a cough is from a COVID-19 positive subject is $\hat{C}$ in Equation 1.

$$\hat{C} = \frac{\sum_{i=1}^{\mathbb{S}} P(Y = 1|X_i, \theta)}{\mathbb{S}} \quad (1)$$

where $P(Y = 1|X_i, \theta)$ is the output of the classifier for feature vector X_i and parameters θ for the i^{th} frame.

The average F1-score along with its standard deviation (σ_{F1}) over the outer folds during cross-validation are shown in Tables VI, VII. Hyperparameters producing the highest F1-score over the inner loops of the cross-validation scheme have been noted as the 'best classifier hyperparameters' in Tables VI, VII.

V. Results

A. TB and COVID-19 cough classification

For the initial classification task in Table VI, the Resnet50 architecture has performed the best by producing the highest mean F1-score of 0.9042 and mean AUC of 0.9190 with the σ_{F1} of 0.83. Although the CNN and LSTM have produced a lower F1-score and AUC, the σ_{F1} has also been lower, 0.61 and 0.49 respectively, suggesting better generalisation and robustness over the folds for less deep architectures. This also indicates that the very deep architectures such as a Resnet50, although able to perform better, are prone to over-fitting. The best feature hyperparameters have been 26 MFCCs, 1024 sample-long frames and 150 frames per event such as a cough.

To prevent over-fitting, we have applied transfer learning and noticed a slight improvement in the DNN classifiers' performance. The F1-score and the AUC have increased to 0.9259 and 0.9245 and the σ_{F1} has decreased to 0.03 from the pre-trained Resnet50 classifier. A similar trend has also been noticed for CNN and LSTM classifiers, where their performance (F1-score and AUC) has also increased along with a lower σ_{F1}.

Although the CNN outperformed the LSTM initially, LSTM has outperformed the CNN after applying transfer learning. The mean ROC curves for the initial and pre-trained Resnet50 are shown in Fig. 5. These two systems achieve 96% and 93% sensitivity respectively at 80% specificity. Thus they exceed the community-based TB Triage test requirement of 90% sensitivity at 70% specificity set by WHO [47].

B. TB and COVID-19 and healthy cough classification

We observe a similar pattern in three-class classification as well. Table VII shows that the highest F1-score of 0.8578 has been achieved from the Resnet50 classifier with a σ_{F1} of 0.67 from the best feature hyperparameters of 39 MFCCs, 1024 sample-long frames and 120 frames per cough. At the same time, CNN and LSTM produce the F1-scores of 0.8220 and 0.8125 with σ_{F1} of 0.41 and 0.49 respectively. Both these F1-scores and σ_{F1} scores are lower than those produced by the Resnet50. As this is a three-class classification, we have replaced AUC with accuracy in Table VII. Again, the signs

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

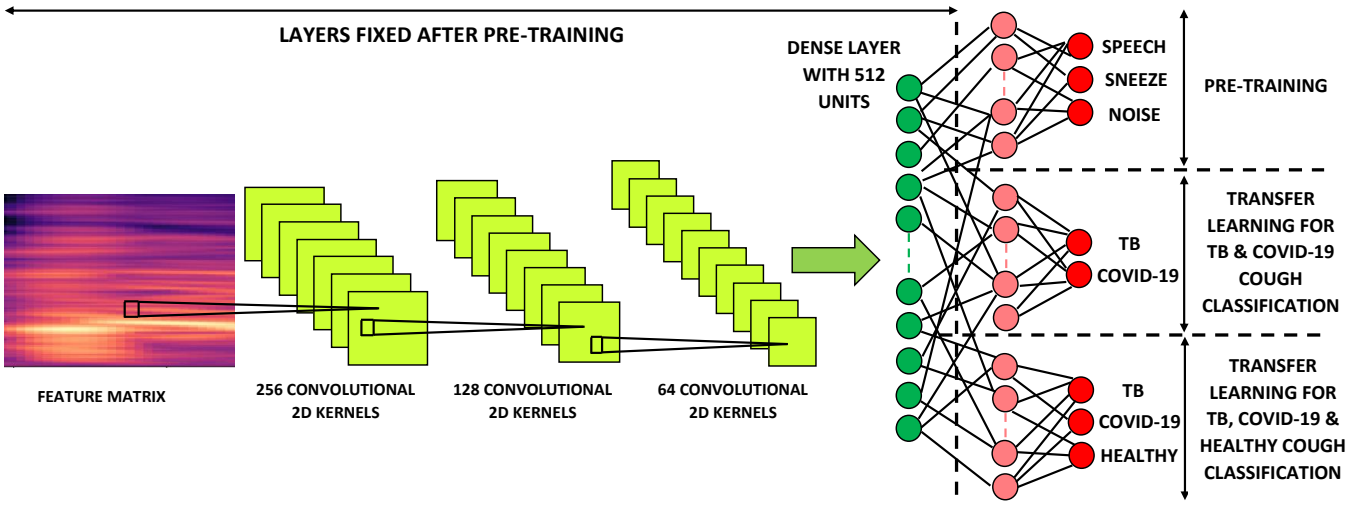

Fig. 4. **Transfer learning architecture for the CNN:** Cross-validation on the pre-training data produced optimal results when three convolutional layers (256, 128 & 64) with (2×2) kernels were used, followed by (2, 2) max-pooling. The outputs of these three convolutional layers were flattened and passed through two fully connected layers (with a dropout rate of 0.3), each consisting 512, 128 relu units. The final fully connected layer consists of 3 softmax units. To apply transfer learning, the final two layers were taken away and was replaced by two fully connected layers with 16 and 2 units for two-class (TB and COVID-19) cough classification and with 16 and 3 units for three-class (TB, COVID-19 and healthy) cough classification.

TABLE VI

CLASSIFYING TB AND COVID-19 COUGHS: RESNET50 HAS PERFORMED THE BEST IN DISCRIMINATING TB COUGHS FROM COVID-19 COUGHS. THE INITIAL EXPERIMENT ACHIEVED THE F1-SCORE OF 0.9042 AND THE AUC OF 0.9190, ALONG WITH THE σ_{F1} OF 0.83. AFTER APPLYING THE TRANSFER LEARNING, F1-SCORE AND AUC INCREASE TO 0.9259 AND 0.9245 AND σ_{F1} DECREASES TO 0.03.

Classifier	Best Feature Hyperparameters	Best Classifier Hyperparameters (Optimised inside nested cross-validation)	Performance		
			F1-score	σ_{F1}	AUC
Resnet50	$\mathbb{M}=26, \mathbb{F}=1024, \mathbb{S}=150$	Default Resnet50 (Table 1 in [43])	0.9042	83×10^{-2}	0.9190
CNN	$\mathbb{M}=26, \mathbb{F}=2048, \mathbb{S}=100$	α_1=256, α_2=2, α_3=0.3, α_4=32, α_7=256	0.8887	61×10^{-2}	0.8895
LSTM	$\mathbb{M}=39, \mathbb{F}=2048, \mathbb{S}=120$	α_3=0.1, α_4=32, α_5=128, α_6=0.001, α_7=256	0.8802	49×10^{-2}	0.8884
Resnet50 + Transfer Learning	*Table V*	*Default Resnet50 (Table 1 in [43])*	*0.9259*	3×10^{-2}	*0.9245*
LSTM + Transfer Learning	”	Table V	0.9134	4×10^{-2}	0.9124
CNN + Transfer Learning	”	”	0.9127	4×10^{-2}	0.9211

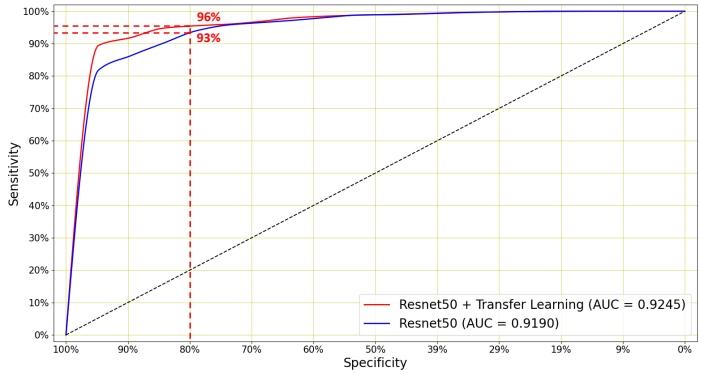

Fig. 5. **The ROC curves for discriminating TB coughs from COVID-19 coughs:** An AUC of 0.9190 is achieved from the Resnet50 and the highest AUC of 0.9245 is achieved after applying transfer learning to this Resnet50 architecture. Both systems achieve 96% and 93% sensitivity respectively at 80% specificity, thus exceed the community-based TB Triage test requirement of 90% sensitivity at 70% specificity set by WHO.

of overfitting are clear in these performances and we apply transfer learning next.

Application of the transfer learning has improved the classification performance by a small margin. The F1-score from the Resnet50 rose to 0.8631 and the σ_{F1} decreased to 0.11. The performances from CNN and LSTM have also improved, as the F1-scores of 0.8455 and 0.8427 have been achieved from these two DNN classifiers, respectively. Their σ_{F1} scores are also much lower: 0.07 and 0.09 respectively. Although, pre-trained CNN and LSTM models have produced lower F1-scores, their σ_{F1} is also lower, unlike in the previous two-class classification. This shows that the application of transfer learning helps the classifiers to be more robust in classification tasks.

VI. DISCUSSION

Although many previous studies have shown that both TB and COVID-19 can be discriminated from healthy coughs, here we show that there are unique disease signatures present in cough audio which is responsible for the machine learning classifiers to discriminate TB coughs from COVID-19 coughs. We have experimentally found that when the cough data are limited, classifier performance can be poor and they are also

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

CLASSIFYING TB AND COVID-19 AND HEALTHY COUGHS: RESNET50 HAS AGAIN BEEN THE CLASSIFIER OF THE CHOICE BY PRODUCING THE HIGHEST F1-SCORE OF 0.8578 WITH A σ_{F1} OF 0.67. THIS PERFORMANCE HAS BEEN IMPROVED TO AN F1-SCORE OF 0.8631 WITH A LOWER σ_{F1} OF 0.11 AFTER APPLYING THE TRANSFER LEARNING.

Classifier	Best Feature Hyperparameters	Best Classifier Hyperparameters (Optimised inside nested cross-validation)	Performance		
			F1-score	σ_{F1}	Accuracy
Resnet50	$\mathbb{M} = 39, \mathbb{F} = 1024, \mathbb{S} = 120$	Default Resnet50 (Table 1 in [43])	0.8578	67×10^{-2}	0.8662
CNN	$\mathbb{M} = 26, \mathbb{F} = 1024, \mathbb{S} = 150$	α_1=256, α_2=2, α_3=0.3, α_4=16, α_7=128	0.8220	41×10^{-2}	0.8311
LSTM	$\mathbb{M} = 26, \mathbb{F} = 2048, \mathbb{S} = 120$	α_3=0.1, α_4=32, α_5=128, α_6=0.001, α_7=256	0.8125	49×10^{-2}	0.8181
Resnet50 + Transfer Learning	*Table V*	*Default Resnet50 (Table 1 in [43])*	*0.8631*	*11*$\times10^{-2}$	*0.8689*
CNN + Transfer Learning	”	Table V	0.8455	7×10^{-2}	0.8564
LSTM + Transfer Learning	”	”	0.8427	9×10^{-2}	0.8490

prone to overfitting. Very deep architectures generally produce higher mean F1-scores, however with the expense of higher variances along the cross-validation folds. Our study shows that the application of transfer learning using vocal data which do not even include cough can be used to improve classifiers' performance in disease classification.

TB and COVID-19 are the two most deadly respiratory diseases transmitted via droplets that are coughed out. Thus, a contact-less diagnosis using a smartphone would be the most desirable solution in these conditions, as opposed to other common coughs from allergic asthma, chronic obstructive pulmonary disease (COPD), bronchitis, common colds, etc, that are not contagious. The deep learning classifiers presented in this study can be implemented in a smartphone, thus enabling the diagnosis process fully non-contact and without needing any expensive laboratory testing equipment, thus protecting the environment and the health care professionals from possible exposure to health risks.

VII. CONCLUSION AND FUTURE WORK

Here in this study, a deep learning based cough classifier which can discriminate between TB coughs and COVID-19 coughs and healthy coughs has been presented, where a subjective test confirmed that respiratory disease can't be confirmed just by listening to the cough audio. The cough audio recordings contain various types and levels of background noise as they were collected inside a TB research centre, recording booth and by using smartphones from subjects around the globe. This cough data include 47 TB subjects, 229 COVID-19 subjects and 1498 healthy subjects contributing 1.68 hours, 18.54 minutes and 1.69 hours of audio respectively. Application of transfer learning has yielded better performance in our previous studies, thus a separate data containing 2.14 hours of sneeze, 2.91 hours of speech and 2.79 hours of noise such as door slamming, engine running, etc, have been used to pre-train three deep neural networks: CNN, LSTM and Resnet50. The class-imbalance in our dataset was addressed by using SMOTE data balancing technique during the training process and using performance metrics such as F1-score and AUC. The classifiers were evaluated by using a 5-fold nested cross-validation scheme. The experimental results show that the highest F1-score of 0.9259 has been achieved from a pre-trained Resnet50 for the two-class (TB vs COVID-19) cough classification task and the highest F1-score of 0.8631 has been achieved from a pre-trained Resnet50 for three-class (TB vs COVID-19 vs healthy) cough classification task. The pre-trained Resnet50 architecture also produces the highest AUC of 0.9245 with 96% sensitivity at 80% specificity, which exceeds the TB triage test requirement of 90% at 70% specificity. The results also show that the application of transfer learning has improved the performance and generalises better over the cross-validation folds, making the classifiers more robust. The best feature hyperparameters also contain higher order of MFCCs, suggesting auditory patterns responsible for disease classification are not perceivable by the human auditory system. This type of cough audio classification is non-contact, cost-effective and can easily be deployed to a smartphone, thus it can be a useful tool for an automatic non-invasive TB and COVID-19 screening, especially in a developing country setting, where these two most deadly contagious diseases claim thousands of lives each year.

As for the future work, we are investigating the length of the coughs required for effective overall classification scores. We are also compiling a bigger dataset containing both TB and COVID-19 patients to improve the existing cough classification models and deploying the TensorFlow-based models on an Android and iOS platform.

ACKNOWLEDGMENT

This research was supported by the South African Medical Research Council (SAMRC) through its Division of Research Capacity Development under the SAMRC Intramural Postdoctoral programme and the Research Capacity Development Initiative as well as the COVID-19 IMU EMC allocation from funding received from the South African National Treasury. We also acknowledge funding from the EDCTP2 programme supported by the European Union (grant SF1401, OPTIMAL DIAGNOSIS; grant RIA2020I-3305, CAGE-TB) and the National Institute of Allergy and Infection Diseases of the National Institutes of Health (U01AI152087).

We would like to thank the South African Centre for High Performance Computing (CHPC) for providing computational resources on their Lengau cluster for this research and gratefully acknowledge the support of Telkom South Africa. We also thank the Clinical Mycobacteriology & Epidemiology (CLIME) clinic team for assisting in data collection, espe-

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

cially Sister Jane Fortuin and Ms. Zintle Ntwana. We also especially thank Igor Miranda, Corwynne Leng, Renier Botha, Jordan Govendar and Rafeeq du Toit for their support in data collection and annotation.

The content and findings reported are the sole deduction, view and responsibility of the researcher and do not reflect the official position and sentiments of the SAMRC, EDCTP2, European Union or the funders.

REFERENCES

[1] W. H. Organization. (2020, Mar.) Tuberculosis; who is most at risk? World Health Organization. Https://www.who.int/news-room/fact-sheets/detail/tuberculosis. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[2] K. Floyd, P. Glaziou, A. Zumla, and M. Raviglione, "The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era," *The Lancet Respiratory Medicine*, vol. 6, no. 4, pp. 299–314, 2018. [Online]. Available: https://doi.org/10.1016/S2213-2600(18)30057-2

[3] P. K. Dewan, J. Grinsdale, S. Liska, E. Wong, R. Fallstad, and L. M. Kawamura, "Feasibility, acceptability, and cost of tuberculosis testing by whole-blood interferon-gamma assay," *BMC Infectious Diseases*, vol. 6, no. 1, p. 47, 2006. [Online]. Available: https://doi.org/10.1186/1471-2334-6-47

[4] F. Bwanga, S. Hoffner, M. Haile, and M. L. Joloba, "Direct susceptibility testing for multi drug resistant tuberculosis: a meta-analysis," *BMC Infectious Diseases*, vol. 9, no. 1, p. 67, 2009. [Online]. Available: https://doi.org/10.1186/1471-2334-9-67

[5] A. Konstantinos, "Testing for tuberculosis," 2010. [Online]. Available: https://doi.org/10.18773/austprescr.2010.005

[6] A. Chang, G. Redding, and M. Everard, "Chronic wet cough: protracted bronchitis, chronic suppurative lung disease and bronchiectasis," *Pediatric Pulmonology*, vol. 43, no. 6, pp. 519–531, 2008. [Online]. Available: https://doi.org/10.1002/ppul.20821

[7] John Hopkins University. (2022, May) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE). John Hopkins University. Https://coronavirus.jhu.edu/map.html. [Online]. Available: https://coronavirus.jhu.edu/map.html

[8] G. Botha, G. Theron, R. Warren, M. Klopper, K. Dheda, P. Van Helden, and T. Niesler, "Detection of Tuberculosis by Automatic Cough Sound Analysis," *Physiological Measurement*, vol. 39, no. 4, p. 045005, 2018. [Online]. Available: https://doi.org/10.1088/1361-6579/aab6d0

[9] A. Carfì, R. Bernabei, F. Landi *et al.*, "Persistent symptoms in patients after acute COVID-19," *JAMA*, vol. 324, no. 6, pp. 603–605, 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.12603

[10] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China," *JAMA*, vol. 323, no. 11, pp. 1061–1069, 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.1585

[11] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, vol. 135, p. 104572, 2021. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104572

[12] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, "Automatic cough classification for tuberculosis screening in a real-world environment," *Physiological Measurement*, vol. 42, no. 10, p. 105014, oct 2021. [Online]. Available: https://doi.org/10.1088/1361-6579/ac2fb8

[13] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features," *Computers in Biology and Medicine*, vol. 141, p. 105153, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.105153

[14] C. Infante, D. Chamberlain, R. Fletcher, Y. Thorat, and R. Kodgule, "Use of cough sounds for diagnosis and screening of pulmonary disease," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 2017, pp. 1–10. [Online]. Available: https://doi.org/10.1109/GHTC.2017.8239338

[15] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020. [Online]. Available: https://doi.org/10.1109/OJEMB.2020.3026928

[16] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. López, and R. H. Gilman, "Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 6017–6020. [Online]. Available: https://doi.org/10.1109/IEMBS.2011.6091487

[17] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021. [Online]. Available: https://doi.org/10.1038/s41597-021-00937-4

[18] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara–A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proc. Interspeech 2020*, 2020, pp. 4811–4815. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2768

[19] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, M. R. Leon J. J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proc. Interspeech 2021*, 2021, pp. 431–435. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-19

[20] I. Miranda, G. Cardoso, M. Pahar, G. Oliveira, and T. Niesler, "Machine Learning Prediction of Hospitalization due to COVID-19 based on Self-Reported Symptoms: A Study for Brazil," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–5. [Online]. Available: https://doi.org/10.1109/BHI50953.2021.9508548

[21] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020. [Online]. Available: https://doi.org/10.1016/j.imu.2020.100378

[22] A. Tena, F. Clarià, and F. Solsona, "Automated detection of COVID-19 cough," *Biomedical Signal Processing and Control*, vol. 71, p. 103175, 2022. [Online]. Available: https://doi.org/10.1016/j.bspc.2021.103175

[23] M. Pahar, I. Miranda, A. Diacon, and T. Niesler, "Automatic Non-Invasive Cough Detection based on Accelerometer and Audio Signals," *Journal of Signal Processing Systems*, pp. 1–15, 2022. [Online]. Available: https://doi.org/10.1007/s11265-022-01748-5

[24] A. Proaño, M. A. Bravard, J. W. López, G. O. Lee, D. Bui, S. Datta, G. Comina, M. Zimic, J. Coronel, L. Caviedes *et al.*, "Dynamics of cough frequency in adults undergoing treatment for pulmonary tuberculosis," *Clinical Infectious Diseases*, vol. 64, no. 9, pp. 1174–1181, 2017. [Online]. Available: https://doi.org/10.1093/cid/cix039

[25] M. Pahar, I. Miranda, A. Diacon, and T. Niesler, "Accelerometer-based bed occupancy detection for automatic, non-invasive long-term cough monitoring," *arXiv preprint arXiv:2202.03936*, 2022. [Online]. Available: https://arxiv.org/abs/2202.03936

[26] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, A. Diacon, and T. Niesler, "Wake-Cough: cough spotting and cougher identification for personalised long-term cough monitoring," *arXiv preprint arXiv:2110.03771*, 2021. [Online]. Available: https://arxiv.org/abs/2110.03771

[27] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy *et al.*, "DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021. [Online]. Available: https://arxiv.org/abs/2103.09148

[28] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The Second Dicova Challenge: Dataset and Performance Analysis for Diagnosis of Covid-19 Using Acoustics," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 556–560. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747188

[29] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of*

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.

the 24th International Conference on Machine Learning, 2007, pp. 935–942. [Online]. Available: https://doi.org/10.1145/1273496.1273614

[30] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016. [Online]. Available: https://doi.org/10.1007/s13748-016-0094-0

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. [Online]. Available: https://doi.org/10.1613/jair.953

[32] M. Pahar, I. Miranda, A. Diacon, and T. Niesler, "Deep Neural Network based Cough Detection using Bed-mounted Accelerometer Measurements," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8002–8006. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9414744

[33] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952261

[34] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 411–412. [Online]. Available: https://doi.org/10.1145/2502081.2502245

[35] I. D. Miranda, A. H. Diacon, and T. R. Niesler, "A comparative study of features for acoustic cough detection using deep architectures," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2601–2605. [Online]. Available: https://doi.org/10.1109/EMBC.2019.8856412

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[37] R. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*. Springer, 2010, pp. 279–282. [Online]. Available: https://doi.org/10.1007/978-90-481-3660-5_47

[38] M. Pahar, "Recreating sounds & delay sensitivity in reconstruction of a multi-band spike based coding of an audio signal," 2011.

[39] M. Pahar and L. S. Smith, "Coding and Decoding Speech using a Biologically Inspired Coding System," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 3025–3032. [Online]. Available: https://doi.org/10.1109/SSCI47803.2020.9308328

[40] M. Pahar, "Reconstructing sound from its coded state in a spike based system," 2012.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: https://doi.org/10.1145/3065386

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[44] S. Liu, "Leave-p-Out Cross-Validation Test for Uncertain Verhulst-Pearl Model With Imprecise Observations," *IEEE Access*, vol. 7, pp. 131 705–131 709, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2939386

[45] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, p. 1111, 2021. [Online]. Available: https://doi.org/10.3390/molecules26041111

[46] D. Fourure, M. U. Javaid, N. Posocco, and S. Tihon, "Anomaly Detection: How to Artificially Increase Your F1-Score with a Biased Evaluation Protocol," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 3–18. [Online]. Available: https://doi.org/10.1007/978-3-030-86514-6_1

[47] W. H. Organization *et al.*, "High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28-29 april 2014, Geneva, Switzerland," World Health Organization, Tech. Rep., 2014. [Online]. Available: https://apps.who.int/iris/handle/10665/135617

Authorized licensed use limited to: University of Stellenbosch. Downloaded on December 01,2022 at 07:46:28 UTC from IEEE Xplore. Restrictions apply.