Sound-Based Cough Detection System using Convolutional Neural Network

1st Julia Zofia Tomaszewska
School of Computing and Engineering
University of West London
London, United Kingdom
21458327@student.uwl.ac.uk
https://orcid.org/0000-0002-9387-4350

2nd Christos Chousidis
School of Computing and Engineering
University of West London
London, Unitged Kingdom
christos.chousidis@uwl.ac.uk
https://orcid.org/0000-0003-3762-8208

3rd Eugenio Donati
School of Computing and Engineering
University of West London
London, United Kingdom
eugenio.donati@uwl.ac.uk
https://orcid.org/0000-0002-0048-1858

***Abstract*— Sound recording and processing techniques can be used in designing diagnostic solutions for a variety of medical conditions related to the respiratory system. In this spectrum, cough monitoring for chronic or seasonal conditions is a significant medical practice. In this paper, a precise cough identification and monitoring system is presented. The system is utilising a convolutional neural network as a feature extraction algorithm and classification system. Including several functions of loading the audio data into the system and converting it into a set of spectrograms, as well as the pre-segmentation stage function, the model retains its relatively low-complexity, which allows accelerating the learning process, also enhanced using dropout. Due to limited audio data available, the dataset dimension was established at 600 samples, split into two equal-numbered groups – 300 samples of "cough" samples, and 300 of "non-cough" samples. The validation accuracy (thus the percentage of samples labelled correctly by the system during the validation process) yielded over 84%, suggesting that this can be a successful cough detection method for future medical applications and devices, such as potential respiratory system condition diagnostic tool.**

***Index Terms*— artificial neural networks, classification, convolutional neural networks, deep learning, dropout, gradient descent, machine learning, optimisation, overfitting, ReLu, SoftMax, supervised learning.**

I. Introduction

Cough is a defence mechanism of a human body intended to clear the upper airways of the respiratory tract. Its lack or impairment can be significantly dangerous in effect. It is also considered one of the earliest and most common symptoms of various respiratory diseases and can be a crucial indicator during patients' diagnostic [1, 2].

Cough responses of a human body are triggered differently depending on a type of the irritant, as well as a specific "receptor" located in the respiratory tract that captures that irritant [3]. Those receptors are airways sensory nerves that induce an adequate response of the respiratory system to the irritant, thus showcasing the response can differ significantly depending on the stimuli. Therefore, different respiratory system conditions cause different cough responses, and thus, unalike cough sounds [3]. Consequently, the diagnosis of a particular respiratory system condition could be possible investigating a cough sound as an audio signal with an analytical and statistical approach.

The project presented here investigates the use of Convolutional Neural Networks (CNN) in a cough detection system, based on the analysis of patient's vocalisations. The system can identify a cough sound amongst other human-related audio events such as laughter, sneezing, or even speech. For the implementation of the Cough Detection System presented in this paper, MATLAB computing environment has been used. The system can accept an audio input of a specified length, transpose it into a pre-segmented spectrogram, assign a correct label to the input, declaring it as a "cough" or a "non-cough" sound, and output the outcome accordingly.

As an input the system uses specific type of spectrograms known as Mel-spectrograms. A Mel-spectrogram is a spectrogram where the frequencies are mapped to the logarithmic Mel-scale. The Mel-scale mimics the human auditory system resolution, and it is thus proven to be more successful when working with human-related sound recognition [4].

II. Existing work on Cough Detection

Most cough detection systems, as well as most human-related sound recognition systems, rely on the use of standard speech recognition algorithms as a fundamental feature extraction algorithm [5, 6]. The most commonly used methods include the Linear Predictive Coding (LPC), variations of Principle Component Analysis (PCA) with predominance of Mel Frequency Cepstral Coefficients analysis (MFCCs) [5, 6, 7, 8, 9]. The success rate of those studies is significantly dependent on the use of artificial neural networks (NN) during the classification stage of the cough data acquired. This leads to the certainty that the use of an artificial NN within a cough detection system can be accomplished, retaining a high rate of accuracy [5, 6].

The Automatic Recognition and Counting of Cough by Barry et al. was among the first studies completed within the computational cough detection and recognition research [9]. The cough monitoring framework implemented an alternative form of computation, combining the Linear Predictive Coding and Principal Component Analysis as the feature extraction algorithm, with the probability neural network. The application of so-called Hull Automatic Cough Counter (HACC) resulted in sensitivity of 80% and a specificity of 96%, which completed the construction of a robust and successful automated system for the analysis of cough and other human-related sounds [9]. Despite satisfactory results, HACC implemented the principal component analysis (PCA) and linear predictive coding (LPC) coefficients with a probabilistic neural network (PNN). Here we propose the use of Mel-spectrograms and CNN for both feature extraction and classification algorithms, ascertaining more straight forward process and thus less computing power required.

Pramono et al. investigated the use of computational methods in cough diagnostics by implemented MFCCs as the fundamental feature extraction algorithm [8]. Their cough detection part of the diagnostic system, using logistic regression model (LRM), indicates sensitivity of 89% and specificity of 92%. The pertussis diagnosis algorithm proposed by Pramono et al. successfully identifies all testing cases, resulting in 100% accuracy. This however was achieved using extremely restricted data of 10 pertussis and 11 non-pertussis audio recordings. Although limited data was used, the results of this study prove that a cough sound caused by a specific respiratory disease maintains enough characteristic audio features to be successfully processed by a computational diagnostic model.

The study conducted by Porter et al. uses MFCCs as feature extraction in combination with the use of artificial NN for classification [7]. The system was designed to detect the cough sound characteristic for asthma, pneumonia, lower respiratory tract disease, croup, and bronchiolitis. The accuracy achieved reaches a minimum of 80%, for each condition, which proves again the potential of artificial NN within cough recognition.

Bales at al. conducted a crucial study within the research of cough detection systems, focusing on the use of convolutional neural networks for feature extraction [10]. The cough detection system implemented by Bales at al. was the main inspiration for the system built for the purposes of this project, encompassing all outlined aims; low-complexity network's structure, limited dataset, and the use of audio recordings of patient's vocalisations. The system developed by Bales et al. also implements a diagnosis algorithm, successfully identifying three respiratory system conditions: pertussis, bronchiolitis, and bronchitis [10]. The success of the system suggests the potential that artificial neural networks may have for the diagnostics of airway conditions.

Despite some similarities between our research and the research conducted by Bales et al., the cough detector proposed in this project retains even simpler architecture. Bales et al. transposes generated Mel-spectrograms to grayscale to unify the intensity scaling. The system proposed in this research avoids this step, as it does not influence the accuracy of the system. Furthermore, the cough detector built in this project uses significantly smaller training dataset (only 300 samples from each group, where Bales et al. uses 993 samples per group) proving this architecture is sufficient for successful results.

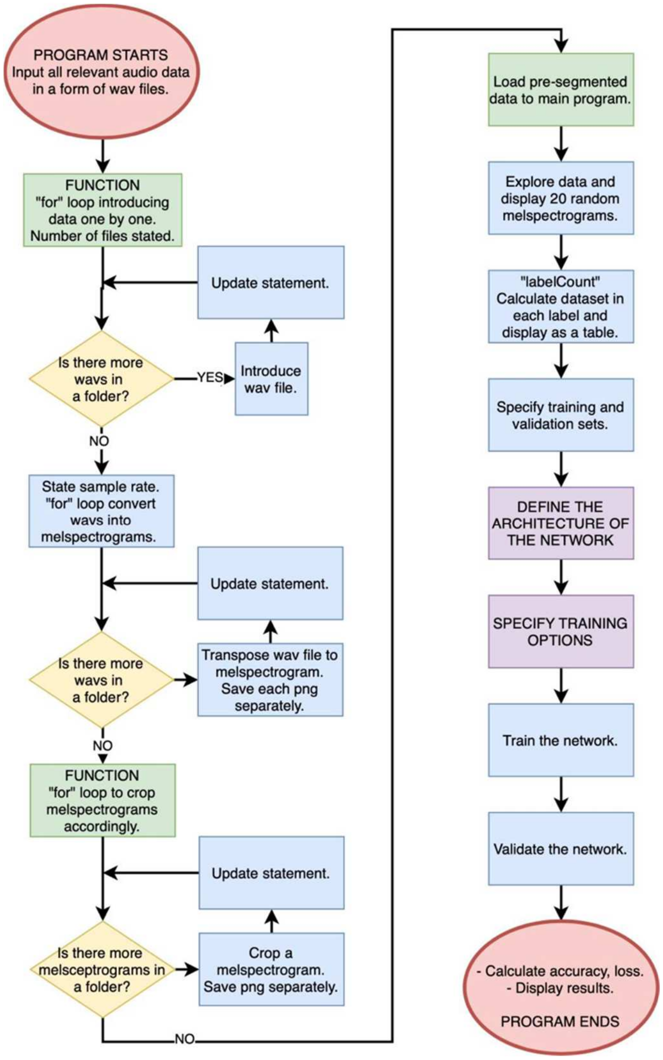

PROGRAM STARTS
Input all relevant audio data in a form of wav files.
FUNCTION "for" loop introducing data one by one. Number of files stated.
Update statement.
Is there more wavs in a folder?
YES
Introduce wav file.
NO
State sample rate. "for" loop convert wavs into melspectrograms.
Update statement.
Is there more wavs in a folder?
Transpose wav file to melspectrogram. Save each png separately.
NO
FUNCTION "for" loop to crop melspectrograms accordingly.
Update statement.
Is there more melsceptrograms in a folder?
Crop a melspectrogram. Save png separately.
NO
Load pre-segmented data to main program.
Explore data and display 20 random melspectrograms.
"labelCount" Calculate dataset in each label and display as a table.
Specify training and validation sets.
DEFINE THE ARCHITECTURE OF THE NETWORK
SPECIFY TRAINING OPTIONS
Train the network.
Validate the network.
- Calculate accuracy, loss.
- Display results.
PROGRAM ENDS

Fig. 1. A signal flow diagram representing the program implemented in this project.

III. Methodology and System Design

It is important to note the possible limitations encountered during the acquisition of audio samples, such as the background noise interfering with the relevant human-made vocalisations, or the availability of the data itself. During the execution of this project the available human-made vocalisation audio samples were highly limited. Due to this issue, the network's complexity was intended to remain as low as possible as a mean of preventing overfitting. To combat possible background noise interference, each sample was investigated before feeding it to the system, ensuring the relevant vocalisation was prominent.

For the development of the cough detection system in this project the CNN were chosen as main processing construct. As Convolutional Neural Networks are specifically suited for image recognition [11], audio files have been pre-segmented and transposed into Mel-spectrograms prior to feeding into the network. The choice of the system's structure was made with regards to:

- the high success rate of CNN-built systems within cough recognition [5],
- the simplicity of the network's structure, enabling quick and efficient application of the system [10],
- the limited dataset of cough audio samples available.

Figure 1 presents the operation stages of the system, beginning with introducing all audio samples from training dataset, pre-segmenting them and converting one-by-one into Mel-spectrograms. On completion of pre-processing, the dataset is fed into the main program, displaying 20 random Mel-spectrograms. The dataset is then split into two labels of "cough" and "non-cough", and number of samples in each label is calculated. Data is then split into training and validation datasets accordingly, the architecture of the network is defined, and training options are defined. Once trained and validated, the network calculates accuracy and loss, and displays the results acquired.

A. Data Acquisition and Preparation

The acquisition of required audio samples was one of the main challenges of this project. For the efficient and sustainable evaluation of the deep learning model built, the required dataset was established at 600 samples in total and split into two equal-numbered groups – 300 samples of "cough", and 300 samples of "non-cough" human-made vocalisations.

All samples were acquired from multiple random subjects, mostly from existing datasets. The dataset included cough and non-cough sounds of female, male, adolescent, and infant subjects. It was mainly acquired from free sound effect libraries such as *Zapsplat* and BBC Sound Effect Library. A significant number of both, "cough" and "non-cough" recordings, had been acquired from the ESC-50 dataset, previously used in successful deep learning sound recognition systems [12]. ESC-50 consists of 2000 samples of 5 major classes (animals, natural soundscapes and water sounds, domestic sounds, urban noises). Most importantly, it includes human non-speech sounds, that has been used for the evaluation of the deep learning model built for this project. Due to a higher number of "non-cough" than "cough" sound, acquired from existing libraries, some cough recordings of the main author had been made to enlarge the "cough" dataset and ensure both datasets are equal in number.

B. Data Preparation

Each sample acquired was split into 5s intervals and rendered into a mono WAV file with a sampling rate of 44100 Hz and 16-bit resolution. The 5s length was chosen to create an arbitrary window that could contain multiple vocalisations whilst increasing sample diversity. The final dataset consisted of 600 samples, including 300 cough recordings and 300 human non-cough vocalizations, such as laughter, sneeze, cry, and sounds caused by various emotions, unrelated to speech.

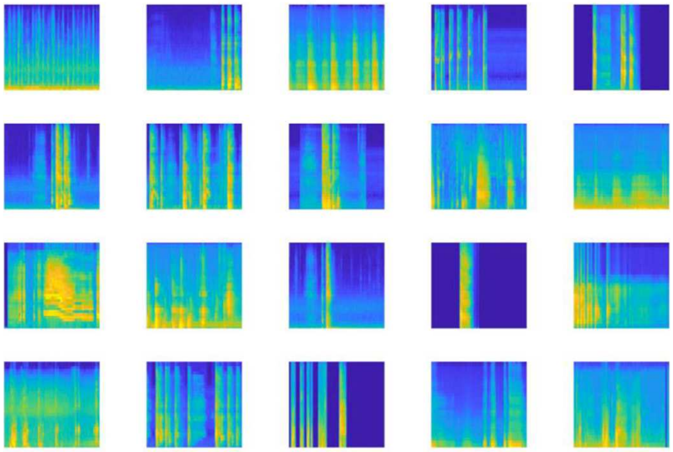


Fig. 1. Randomly selected Mel-spectrograms from the entire dataset.

The dataset was fed into the system in a form of spectrograms – visual representation of signal frequencies and its strength changing over time. For this project Mel-spectrograms were chosen, meaning spectrograms in Mel-scale. Mel scale is perceptual scale of signal frequencies' strength, perceived by a human listener to be of equal distance from one another. The choice was made over linear spectrograms, as they are better suited for human hearing perception and had been found to be most effective in audio classification applications [13].

To introduce the dataset into the system, two functions were created in MATLAB for both data groups. The first function loads dataset into the system in a form of WAV files and converts them into Mel-spectrograms with the application of Short Time Fourier Transform (STFT). Fourier Transform allows to convert a signal into individual spectral components, providing frequency information about that signal. The STFT allows to calculate the change of those frequencies over time by splitting them into shorter segments and performing Fourier Transform of each [13].

The second function executes the pre-segmentation stage. Mel-spectrograms are cropped accordingly to suit the pre-segmentation requirements of the CNN system built. *Figure 2*

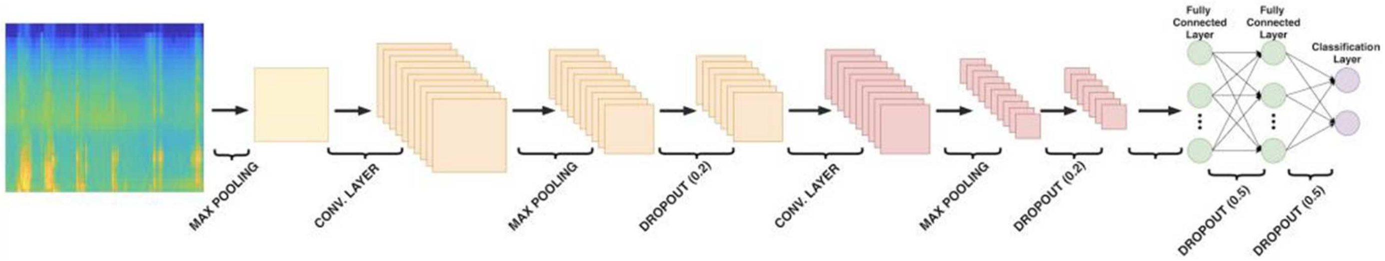
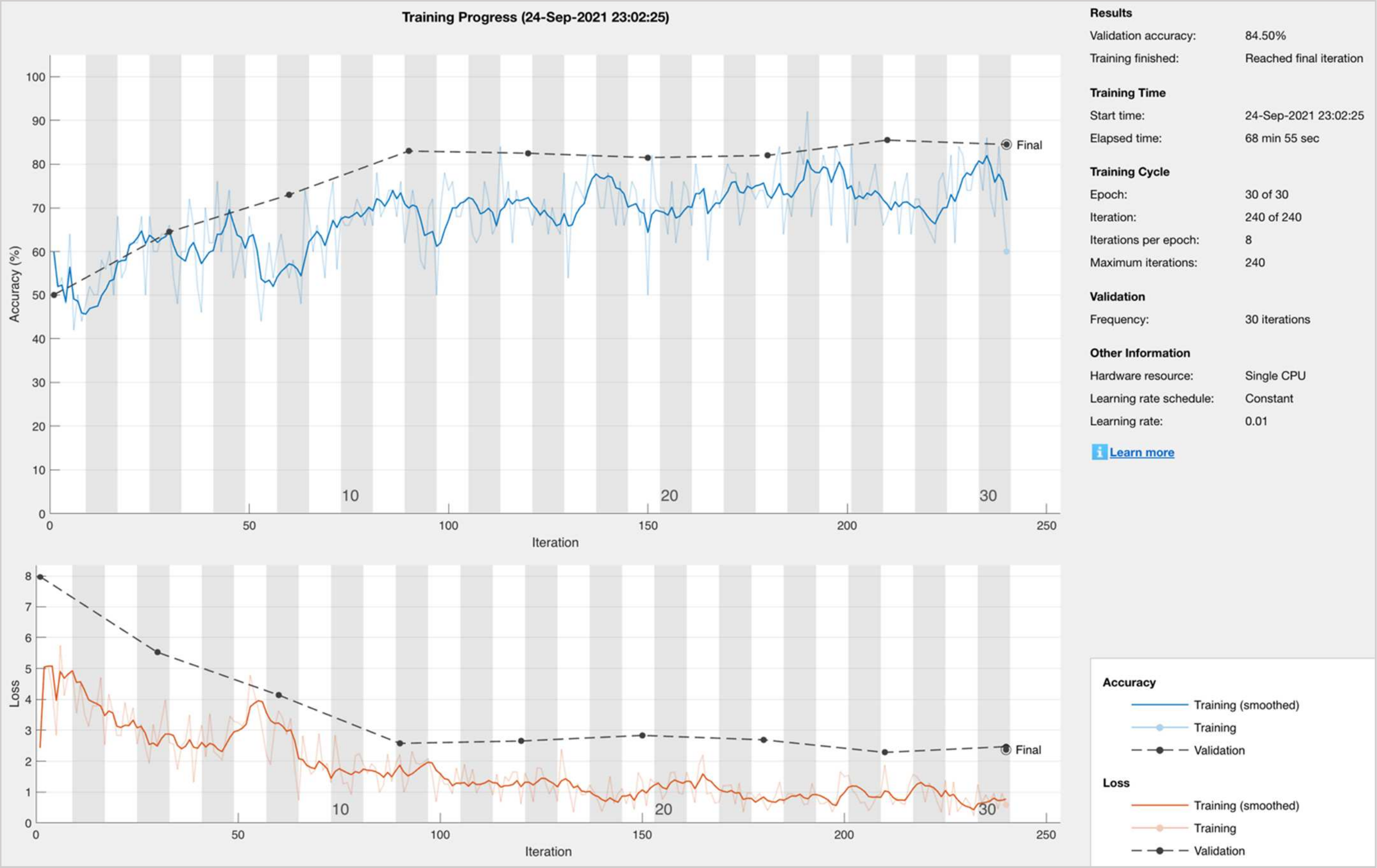


Fig. 2. The CNN structure applied in Cough Detection System.



Results
Validation accuracy: 84.50%
Training finished: Reached final iteration

Training Time
Start time: 24-Sep-2021 23:02:25
Elapsed time: 68 min 55 sec

Training Cycle
Epoch: 30 of 30
Iteration: 240 of 240
Iterations per epoch: 8
Maximum iterations: 240

Validation
Frequency: 30 iterations

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.01

Fig. 3. Training progress of the CNN structure of the Cough Detection System

shows a group of 20 randomly selected spectrograms form the entire dataset (both groups).

C. Network Architecture and Training

Due to the high resolution of input dataset, the model's structure begins with a 2×2 max-pooling layer, applied to lower the overall complexity of the system. Furthermore, the system implements dropout layers to prevent overfitting, alike the algorithm implemented by Bales et al. [10]. The overview of the system's structure can be seen on *Figure 3*.

Since the dataset is converted to Mel-spectrograms with a colour gradient the channel size of the input layer was set to 3, corresponding to RGB values, the size of each sample is

725×850×3, where 725×850 correspond to the number of pixels of each Mel-spectrogram. Each convolutional layer implements padding and 32 filters of size 5×5. Subsequently, data is normalised, and the ReLU function is added to activate the nodes according to the rectified linear function. Max-pooling layers are applied with the stride (the step of the filter) equal 2×2 to reduce the size of the input. Moreover, the dropout of probability equal 0.2 is applied. Between two fully connected layers, the dropout increases to 0.5 to acquire more specific results. SoftMax activation function is implemented right before the classification layer to represent the probability distribution over two classes of "cough" and "non-cough" sounds.

The optimisation algorithm chosen for this application was Adam, as it is significantly outperforming other optimisers within audio detection and recognition systems [5, 6, 10]. During its application as a mini-batch gradient descent algorithm, the size of one batch was set to 50, meaning the input data was introduced into the network in the number of 50 during each iteration. The number of maximum epochs, thus the runs through the entire input data provided, was set to 30, as this number allows the network to fully explore feature maps acquired. The data inputted into the system is shuffled for every epoch to avoid the risk of creating batches that may be unrepresentative of the entire dataset. Training options also declare the validation dataset and set the frequency of the validation to 30. This means the evaluation of validation metrics is to be repeated every 30th iteration.

The initial learning rate is specified as 0.01, which allows the parameters to adjust in relatively small steps, nevertheless, it is large enough to make significant progress and allow the network to run faster than with a smaller learning rate [14]. Once training options are specified, the accuracy is defined as the fraction of the dataset labels that are predicted correctly by the network. It is therefore calculated by dividing the classification of the validation dataset made by the model, by the actual labels of that data.

IV. Additional testing methods

To evaluate the unbiased estimate of the skill of the final model, additional testing (separate from the validation processes) was performed [14]. For that purpose, additional 20 audio samples – 10 cough sounds and 10 human-made non-cough sounds – were recorded and edited following a similar process to the network's training dataset. Providing that the designed system is an example of the supervised machine learning, all files were named according to the data contained therein ("cough" or "non-cough"). The testing images were then fed into the network and classified using the Cough Detection System designed. The results were collected and assessed manually.

All 20 audio samples intended for the final manual testing were fed into the network, and then detected and labelled by the network correctly, in accordance with its audio content. Thus, the additional manual testing was deemed 100.00% successful. Nevertheless, each sample fed into the system during the additional manual testing, was a premeditated recording made within a controlled environment. Testing stages using live recordings acquired in an uncontrolled environment were not concluded, as the research was intended as a prototype for audio data obtained in optimal conditions. The use of recordings acquired in an uncontrolled environment, however, is foreseen as part of future research.

V. Results Analysis

Following the testing and manual evaluations, the final version of the network was simplified by reducing the number of convolutional layers to two only. Furthermore, to avoid the overfitting, the dropout layers were added, following two middle max-pooling layers (probability of 0.2), as well as both fully connected layers (probability of 0.5). The input layer was also re-evaluated, determining that the network should apply an additional max-pooling layer following the input layer directly, due to the unnecessarily large size of input Mel-spectrograms (Fig. 3). The final version of the network build in this project was able to achieve a validation accuracy of 84.50% (Fig. 4).

VI. Conclusion

This project investigated the use of Convolutional Neural Networks as the main programming construct in the Cough Detection algorithm, able to identify a cough sound based on patient's vocalisations only.

Providing the limitations of dataset, the low complexity of the network was sustained as a mean of preventing the overfitting. Due to that, only two convolutional layers were implemented alongside dropout layers accelerating the training process, and thus, allowing the network to remain relatively simple and stimulating faster learning process. The simplified structure of the model allowed the validation accuracy to reach exactly 84.50%.

The accuracy of this value proves that minor background noise interference does not prevent CNN from detecting the relevant features of cough sound, however, elimination of the background noise could potentially improve the accuracy. Furthermore, no additional testing using live recordings of an uncontrolled environment was completed. In order to thoroughly test the system built, additional manual testing using live recordings shall be completed before further research.

Providing the limited dataset available, the accuracy of this value proves cough detection abilities of a system build solely of convolutional neural networks. Furthermore, it signifies large potential of a CNN-built cough detection system as a potential in further cough-recognition-related projects. Given those results, further research within cough detection and recognition systems is foreseen. The future research will focus particularly on the potential of artificial NN within recognition of various human-made vocalisations' characteristics, aiming to develop a potential diagnostic tool.

The work described in this paper shows that human cough audio samples can deliver enough audio features for an artificial NN to detect and recognise. Moreover, various respiratory system conditions can cause different vocalisation responses in a human subject [7]. It is thus reasonable to assume that further research into the properties of human vocalisations could help develop a standalone AI-based system as a support for medical professionals in the diagnosis of respiratory system conditions.

References

[1] S. S. Birring, S. Matos, R. B. Patel, B. Prudon, D. H. Evans, I. D. Pavord, "Cough frequency, cough sensitivity and health status in patients with chronic cough," *Respiratory medicine*, vol. 100, pp. 1105- 9, 2006.

[2] K. W. Altman, R. S. Irwin, "Cough: An Interdisciplinary Problem," *Otolaryngologic Clinics*, 43(1). Elsevier Health Sciences, 2010.

[3] B. J. Canning, "Afferent nerves regulating the cough reflex: mechanisms and mediators of cough in disease," *Otolaryngologic Clinics of North America*, vol. 43(1), pp. 15-25, 2010.

[4] B. D. C. Ramesh and R. S. Vishnu, "CNN and Sound Processing-Based Audio Classifier for Alarm Sound Detection," *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, pp. 365-375. Springer, Singapore, 2020.

[5] J. Amoh and K. Odame, "Deep Neural Networks for Identifying Cough Sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, pp. 1003–1011, IEEE, 2016.

[6] J. Amoh and K. Odame, "DeepCough: A Deep Convolutional Neural Network in A Wearable Cough Detection System," *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, IEEE, 2015.

[7] P. Porter, U. Abeyratne, V. Swarnkar, J. Tan, T. Ng, J. M. Brisbane, D. Speldewinde, J. Choveaux, R. Sharan, K. Kosasih, et al., "A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children," *Respiratory research*, vol. 20, no. 1, 2019.

[8] R. X. A. Pramono, S. A. Imtiaz, E. Rodriguez-Villegas, "A Cough- Based Algorithm for Automatic Diagnosis of Pertussis," *PLOS ONE*, vol. 11, no. 9, 2016.

[9] S. J. Barry, A. D. Dane, A. H. Morice, and A. D. Walmsley, "The automatic recognition and counting of cough." *Cough (London, England)*, vol. 2, p. 8, Jan., 2006.

[10] C. Bales, M. Nabeel, C. N. John, U. Masood, H. N. Qureshi, H. Farooq, I. Posokhova, A. Imran, "Can machine learning be used to recognize and diagnose coughs?," *International Conference on e-Health and Bioengineering (EHB), October 2020*, IEEE, pp. 1-4, 2020.

[11] P. Kim, *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*. Seoul, Korea: Apress, 2017.

[12] K. J. Piczak, "Environmental sound classification with convolutional neural networks," *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP),* pp. 1–6, 2015.

[13] A. A. Rahman, J. Angel Arul Jothi, "Classification of UrbanSound8k: A Study Using Convolutional Neural Network and Multiple Data Augmentation Techniques," *International Conference on Soft Computing and its Engineering Applications*, pp. 52-64. Springer, Singapore, 2020.

[14] J. Brownlee, "Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions," *Machine Learning Mastery*, 2018.