Differentiation of Dry and Wet Cough Sounds using A Deep Learning model and Data Augmentation

by

Saiful Huq

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Carleton University
Ottawa, Ontario

© September 2023,
Saiful Huq
Copyright

This thesis includes content (paragraphs, tables, figures, etc.) from the following:


The items listed above are conference proceedings published or submitted to the Institute of Electrical and Electronics Engineers (IEEE). The IEEE allows the reuse of published articles in a thesis or dissertation without requesting permission. I am the lead author in the above proceedings and will be referencing content from them in the chapters of this thesis.
Abstract

Automated classification of cough sounds has increased in importance due to factors such as the worldwide COVID-19 pandemic. To train such classification models requires a large dataset of cough sounds; however, it remains challenging to find sufficient expert-labelled training data. This thesis explores a novel form of audio data augmentation, where training cough sounds are corrupted with varying levels of reverberation and Gaussian noise. The combination of noise and reverberation is more effective than both traditional image-based augmentation techniques and either noise or reverberation alone, leading to near-human accuracy on a wet vs. dry cough classification task using a ResNet18 model across two cough datasets. Alignment between the training and testing environments is examined using the Speech-Commands audio dataset. While models trained with the closest reverb and noise level to the test environment gave the best results, the proposed audio augmentation technique produces models with robust performance across test environments.
Acknowledgements

I would like to thank AGE-WELL NCE Inc., the National Research Council (NRC) Canada, and Carleton University for their funding and support throughout my research.

I would like to extend my sincere thanks and gratitude to my academic advisors, Dr. James Green and Dr. Rafik Goubran, for their continued support throughout my research program. I am grateful for the guidance, understanding, patience, and time that they have so kindly and readily offered. I would like to thank my friends and lab-mates for their feedback and support.

Finally, I would like to thank my family – your unconditional love and encouragement has been the fuel that powered me through this experience. To my wife, Fariha, thank you for your love and patience, especially the last months of the program. To my sister and family, appi (Rumpa), cb (Tinku), Rehan & Faraz, thank you for always being there to cheer me up and encourage me every time I needed it. To my mother and father, thank you for your love, kindness, patience, and your belief in me.
Table of Contents

Copyright...................................................................................................................................................... ii
Abstract........................................................................................................................................................ iii
Acknowledgements ....................................................................................................................................... iv
Table of Contents ........................................................................................................................................ v
List of Tables ................................................................................................................................................ vii
List of Figures ............................................................................................................................................... viii
Glossary of Terms & Acronyms................................................................................................................ x
Chapter 1: Introduction ................................................................................................................................ 1
  1.1 Introduction ........................................................................................................................................... 1
  1.2 Background and Motivation .................................................................................................................. 1
  1.3 Problem Statement ............................................................................................................................. 4
  1.4 Thesis Objective ................................................................................................................................... 5
  1.5 Thesis Outline ..................................................................................................................................... 5
Chapter 2: Background and Literature Review ............................................................................................. 7
  2.1 Introduction ........................................................................................................................................... 7
  2.2 Background .......................................................................................................................................... 7
    2.2.1 Cough Sounds ............................................................................................................................. 7
    2.2.2 Cough Types ............................................................................................................................... 9
    2.2.3 Non-Contact Medical Assessment ............................................................................................ 12
    2.2.4 Deep Learning ............................................................................................................................ 14
    2.2.5 Digital Signal Processing and Spectrogram Generation ............................................................ 16
    2.2.6 Noise and Reverberation ........................................................................................................... 18
    2.2.7 Deep Learning Models .............................................................................................................. 21
    2.2.8 Regularization and Overfitting .................................................................................................. 22
  2.3 Critical Literature Review ..................................................................................................................... 23
    2.3.1 Audio Classification Using Deep Learning ................................................................................ 23
    2.3.2 Cough Sound Classification ...................................................................................................... 26
    2.3.3 Data Augmentation for Training Deep Learning Models .......................................................... 28
    2.3.4 Research Gap ............................................................................................................................ 31
Chapter 3: Data Augmentation with Reverb and Noise for Cough Classification...33

3.1 Introduction ........................................................................................................................................33

3.2 Methods ............................................................................................................................................35

3.2.1 Dataset........................................................................................................................................35

3.2.2 Deep Learning Model ................................................................................................................35

3.2.3 Audio Pre-Processing..................................................................................................................36

3.2.4 Augmentations ..........................................................................................................................37

3.2.5 COUGHVID Dataset ..................................................................................................................42

3.3 Results ............................................................................................................................................45

3.3.1 Traditional Augmentation ........................................................................................................45

3.3.2 Novel Augmentation (Noise and Reverb) ................................................................................46

3.3.3 COUGHVID Dataset ..................................................................................................................48

3.4 Conclusion ......................................................................................................................................52

Chapter 4: Alignment Between Train and Test Environments ........................................54

4.1 Introduction ....................................................................................................................................54

4.2 Methods ........................................................................................................................................55

4.2.1 Dataset........................................................................................................................................55

4.2.2 Augmentations and Deep Learning Model...............................................................................55

4.3 Results ............................................................................................................................................58

4.3.1 Varying Noise Levels ................................................................................................................58

4.3.2 Varying Reverberation Levels ..................................................................................................59

4.3.3 Combined Noise and Reverberation in the Test Environment ...............................................59

4.4 Conclusions ........................................................................................................................................58

Chapter 5: Conclusions ..................................................................................................................66

5.1 Conclusions ....................................................................................................................................66

5.2 Summary of Research Contributions ........................................................................................67

5.3 Recommendations for Future Work ............................................................................................68

References ..........................................................................................................................................69
List of Tables

Table 1  Training and Test Dataset Distribution ......................................................... 38
Table 2  COUGHVID Expert Annotations Provided [34] .............................................. 43
Table 3  Noise-Augmentation Level Description .......................................................... 56
Table 4  Reverb Augmentation Level Description ......................................................... 57
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Audio Processing Flowchart</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>The Human Respiratory System [38]</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Typical Time Domain Cough Sound Signal</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>Typical Time Domain Cough Sound Signal and Corresponding Mel-Frequency Spectrogram</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>Effect of White Gaussian Noise on Spectrograms</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>Reverberation and Space</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>Illustration of Various DL Architectures [20]</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Data Augmentation in Computer Vision [101]</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>Dry (left) and Wet Cough (right) Waveforms in Time Domain</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>Cough Spectrogram</td>
<td>34</td>
</tr>
<tr>
<td>11</td>
<td>Spectrograms of Training Set Dry Coughs</td>
<td>40</td>
</tr>
<tr>
<td>12</td>
<td>Spectrograms of Training Set Wet Coughs</td>
<td>41</td>
</tr>
<tr>
<td>13</td>
<td>Traditional Data Augmentation Techniques – Model Accuracy Boxplots</td>
<td>46</td>
</tr>
<tr>
<td>14</td>
<td>Reverb &amp; Noise Augmentation Techniques – Model Accuracy Boxplots</td>
<td>47</td>
</tr>
<tr>
<td>15</td>
<td>The ROC Curves for the COUGHVID Dataset</td>
<td>49</td>
</tr>
<tr>
<td>16</td>
<td>COUGHVID – Model Accuracy Boxplots using Expt4 Labels</td>
<td>50</td>
</tr>
<tr>
<td>17</td>
<td>COUGHVID – Comparison of Accuracies Using Each Expert-Label as Ground Truth and using Noise and Reverb Augmentation</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>DL Classifier Accuracies for Varying Noise Augmentation</td>
<td>58</td>
</tr>
<tr>
<td>19</td>
<td>DL Classifier Accuracies for Varying Reverb Augmentation</td>
<td>59</td>
</tr>
</tbody>
</table>
Figure 20. DL Classifier Accuracies for Varying Noise & Reverb Augmentation .......60

Figure 21. DL Classifier Accuracies for Varying Noise & Reverb Augmentation .......62

Figure 22. Ranked DL Classifier Accuracies for Varying Augmentation (1-Reverb, 2-Noise) ............................................................................................................................................................................63

Figure 23. Ranked DL Classifier Accuracies for Varying Augmentation (1-Noise, 2-Reverb) ..................................................................................................................................................................................................................................................64
## Glossary of Terms & Acronyms

<table>
<thead>
<tr>
<th>Term/Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam Optimizer</td>
<td>A DL algorithm that is an extension of the gradient descent algorithm used to optimize the training of the network and converge fast</td>
</tr>
<tr>
<td>AGE-WELL NCE Inc.</td>
<td>A Canadian Network of Centres of Excellence to develop technologies and services for healthy aging</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>Audiomentations</td>
<td>A Python library to apply audio transformations to sound clips</td>
</tr>
<tr>
<td>Bias</td>
<td>It is when a DL model wrongly puts more weight on certain features in a training set, that do not translate to the test-set/reality.</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>COPD</td>
<td>chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>COUGHVID</td>
<td>A large, crowdsourced audio-dataset of cough sounds</td>
</tr>
<tr>
<td>COVID-19</td>
<td>The coronavirus disease</td>
</tr>
<tr>
<td>Data leakage</td>
<td>Data leakage is when information from outside the training dataset gets used to create a model. This is a problem</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning is a subset of machine learning that employs the use of AI and neural-networks primarily for prediction tasks</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Networks</td>
</tr>
<tr>
<td>Dry cough</td>
<td>A type of cough that does NOT produce fluids and has 'dry' audible properties</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Networks</td>
</tr>
<tr>
<td>GERD</td>
<td>Gastroesophageal reflux disease</td>
</tr>
<tr>
<td>GM</td>
<td>Generative Models</td>
</tr>
<tr>
<td>GRU</td>
<td>Gated-Recurrence Unit</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Models</td>
</tr>
<tr>
<td>Label/well-labelled</td>
<td>In classification terms, this is the categorical outcome that a model is trying to predict</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long-Short-Term Memory</td>
</tr>
<tr>
<td>Mel Scale</td>
<td>A perceptual scale defined by listeners as equidistant from one another. The formula is defined here: <a href="https://en.wikipedia.org/wiki/Mel_scale">https://en.wikipedia.org/wiki/Mel_scale</a></td>
</tr>
<tr>
<td>MeMeA</td>
<td>An IEEE International Symposium on Medical Measurements and Applications</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MLP</td>
<td>MultiLayer Perceptron</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>NRC</td>
<td>National Research Council</td>
</tr>
<tr>
<td>Overfitting</td>
<td>When a model performs well with training data, but poorly on the test data.</td>
</tr>
<tr>
<td>PSD</td>
<td>Power Spectral Density</td>
</tr>
<tr>
<td>Regularization</td>
<td>A technique in ML used to calibrate models to minimize the adjusted loss function, lowering the affects of bias and variance</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>Variance</td>
<td>It is the changes in a model when different portions of training data are used</td>
</tr>
<tr>
<td>Wet cough</td>
<td>A type of cough that produces fluids and has 'wet' audible properties</td>
</tr>
<tr>
<td>ZCR</td>
<td>Zero-Crossing Rate</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

1.1 Introduction

This chapter presents the background, thesis motivation, problem statement, and objectives. The thesis structure and contributions are also listed here.

1.2 Background and Motivation

Audio analysis is a rapidly evolving field that involves the extraction of meaningful information from audio signals using various computational techniques. It plays a vital role in a wide range of applications and domains, making it an important area of research and development [1]–[3].

One of the key reasons why audio analysis is important is its ability to enable machines to understand and interpret audio data, similar to how humans perceive and comprehend sound [3]. By extracting relevant features and patterns from audio signals, it enables applications such as speech recognition, music information retrieval, environmental sound classification, and more [3], [4].

Increasingly, methods from artificial intelligence and deep learning (DL) are being developed and applied to audio analysis [1], [2], [5], [6]. Deep learning techniques, particularly convolutional neural networks (CNNs) have proven to be highly effective in modeling and extracting useful representations from audio data[7]. With their ability to automatically learn hierarchical features, deep learning models have achieved state-of-the-art performance in various audio analysis tasks, including speech recognition, music classification, and sound event detection[7], [8].

DL is a subset of machine learning (ML) that involves training artificial neural networks to perform these complex tasks. DL models often require very large training
datasets to accomplish these tasks [9], [10]. Curating such large datasets, while ensuring they are well-labeled, is often a challenging endeavor and, in such cases, data augmentation techniques can be utilized to supplement smaller datasets [11]–[13]. Data augmentation is a technique to enhance the size and diversity of a dataset by applying data-specific transformations to existing samples [12]. For instance, for image-type data, transformations such as random, rotations, flips, zoom, and brightness adjustments can be applied to augment the dataset. Data augmentation not only increases the dataset size, but also helps the generalizability of a DL model by preventing overfitting in the training phase [11], [12].

One prominent application of audio analysis is environmental sound classification. This involves classifying sounds from the environment, such as traffic noise, bird songs, sirens, or machinery sounds [1]. Environmental sound classification has numerous practical applications, including soundscape analysis for urban planning, soundscape monitoring for noise pollution assessment, and audio surveillance for security purposes [3]. By accurately classifying environmental sounds, it becomes possible to monitor and understand the acoustic environment, leading to improved decision-making and resource allocation [3], [14].

In the medical field, audio analysis has shown great promise in a variety of applications. One notable area is respiratory sound analysis, where the acoustic characteristics of cough sounds, breathing sounds, and lung sounds can be analyzed to diagnose respiratory conditions, such as asthma, chronic obstructive pulmonary disease (COPD), or pneumonia [15], [16]. Additionally, cardiac auscultation is a process that involves analyzing heart sounds and murmurs to detect cardiovascular abnormalities and
assess cardiac health [17], [18]. By leveraging audio analysis techniques, healthcare professionals can obtain valuable insights for diagnosing and monitoring respiratory and cardiac conditions.

One of the most common symptoms of respiratory disease is coughing [19], [20]. It takes many years of careful training by healthcare professionals to properly extract useful clinical information from cough sounds [21]–[23]. Relying on patients’ self-reporting of cough severity and nature is not recommended, as it is often difficult or impossible for patients to be accurate due to their lack of medical knowledge [19], [21], [24].

Cough sound analysis can be typically separated into two parts: cough sound detection, and cough sound classification [22], [25]. Cough sound detection involves the use of microphones and digitization of the signal into audio files that are segmented into single cough sound audio clips. For this thesis, the focus is on the second part – the cough classification – using datasets that have the audio detection and cough segmentation steps already completed (see process in Figure 1).

![Audio Processing Flowchart](image)
Leveraging audio analysis techniques to evaluate cough sounds has many desirable applications in patient health assessments. It not only promotes non-contact evaluation of respiratory health but can also be included in smart home systems for remote health monitoring of patients in need of such medical support [26], [27]. As respiratory sounds are one of the first indications of respiratory illness, respiratory sound analysis can help with early detection and diagnosis of disease [1], [28], [29].

Characterizing coughs plays an important role in diagnosing respiratory diseases [30], [31]. By obtaining information on cough sounds from their spectrograms, a DL model can be used to classify coughs [2], [20]. However, there is a limitation on the amount of available, expert-labelled cough sounds – not enough to train a complex DL model from scratch [32]. Combining deep learning and data augmentation techniques may help address this problem and be effectively applied to cough classification.

1.3 Problem Statement

While data augmentation is widely used for maximizing utility of available data for training deep learning methods, it is unclear which types of augmentation are most effective for spectrograms used for cough sound classification. This thesis will examine various augmentation techniques, including combinations of techniques, to maximize cough sound classification accuracy. In particular, augmentations that are ‘natural’ to the problem domain (i.e., additive noise and reverberation) will be investigated and compared with traditional image-based augmentation techniques.

The alignment between training and testing environments for audio sound classification will be examined in terms of audio noise and reverberation. This thesis will
examine whether a single classifier, trained on coughs with a random mixture of noise and reverberation levels, can perform robustly across a range of test environments.

1.4 Thesis Objective

The primary objective of this thesis is to use data augmentation techniques to improve a DL models ability to differentiate between wet and dry cough sounds. A CNN-based cough classification method is developed and evaluated. Data augmentation techniques are systematically evaluated. The use of noise and reverberation as a combined augmentation is shown to be a novel and robust data augmentation technique for spectrogram data. Results are demonstrated on a pretrained CNN model on multiple audio datasets (two cough datasets and one speech command dataset).

1.5 Thesis Outline

Chapter 2 provides a detailed discussion on cough sound classification, including cough sound detection and past feature extraction approaches. The chapter also includes a literature review of previous applications of machine learning and DL for cough classification.

Chapter 3 discusses the build up to the cough sound classification that was done. This includes a short discussion on the dataset used, the augmentation methods evaluated, and the comparison of the performance of these methods. Portions of this chapter were previously published in proceedings of the 2023 Institute of Electrical and Electronics Engineers (IEEE) Medical Measurements and Applications (MeMeA) conference on which I was the primary author [33]. The chapter ends with detailed work on a pretrained model on cough classification on a separate database of cough sounds (COUGHVID [34]).
Chapter 4 explores the alignment between training and test environments in terms of noise and reverberation levels. Models trained on datasets with various augmentation techniques are evaluated on the test sets with varying levels of corruption by Gaussian noise and reverberation.

Chapter 5 presents the thesis conclusions and discussions for future directions in which to expand the work.
Chapter 2: Background and Literature Review

2.1 Introduction

This chapter first presents background material required to understand the detailed research contributions described later in the thesis. The second half of the chapter summarizes the state of the field through a critical review of the literature in the area of audio classification, with emphasis on cough sound classification.

2.2 Background

2.2.1 Cough Sounds

Coughing is a vital defense mechanism of the respiratory system which functions to protect the airways and lungs from irritants and harmful substances [35]. It is a reflex action that involves a series of complex physiological processes in the human body [35]–[37]. Coughing can be both voluntary and involuntary. Cough sounds are the audible outcome of this intricate mechanism, providing valuable clues for diagnosing various respiratory conditions [22].

Figure 2 shows the detailed anatomical features that are related to respiration and linked to the generation of cough sounds. Coughs can arise from various sources, and understanding the underlying causes is crucial for accurate diagnosis and treatment [38], [39]. Common triggers for coughs include respiratory infections, such as the common cold, influenza, or pneumonia, which lead to the irritation of airway linings [19]. Other causes may involve allergic reactions, chronic respiratory conditions like asthma, chronic obstructive pulmonary disease (COPD), or bronchitis [40]. Gastroesophageal reflux disease (GERD) and exposure to environmental irritants, such as smoke, dust, or pollution, can also induce coughing [40], [41]. Additionally, coughs can be psychogenic, where
emotional factors play a significant role in their manifestation [42], [43]. Coughs serve an essential purpose in the respiratory system. By forcibly expelling irritants, mucus, and foreign particles from the airways, coughing protects the delicate tissues of the lungs [41]. The expulsion of excess mucus from the airways also keeps them relatively open and unobstructed for ease of breathing [44]. Finally, foreign objects such as food particles or water that accidentally enter the trachea are also expelled by coughing, reducing health risks such as aspiration [45]–[47].

Figure 2. The Human Respiratory System [38]
Cough sounds are produced by the rapid expulsion of air from the lungs, facilitated by the coordinated action of several anatomical structures [47]. The process involves the respiratory muscles, including the diaphragm and intercostal muscles, working in tandem to create a sudden increase in intra-thoracic pressure [22]. This pressure surge forces the glottis to close momentarily, trapping air within the lower respiratory tract. When the glottis opens, the compressed air rushes out, generating the characteristic cough sound. Additionally, the vocal cords may also contribute to the quality and tone of the cough sound, depending on the specific condition or underlying cause [15], [38].

2.2.2 Cough Types

Coughs can be broadly classified into two main types: wet and dry coughs. These categories are based on the nature of the cough sound and the underlying conditions that trigger them [15], [48]. Distinguishing between wet and dry coughs is essential for accurate diagnosis and appropriate treatment of respiratory ailments [35], [49].

A wet cough, often referred to as a productive cough, is characterized by the presence of mucus or phlegm in the airways [48]. During a wet cough, this excess mucus is expelled, which may appear thick and discolored. The production of mucus is typically a response to an infection or inflammation in the respiratory tract, such as in cases of bronchitis, pneumonia, or COPD [40], [50]. The accumulation of mucus can cause discomfort and difficulty breathing, and the act of coughing helps to clear the airways, facilitating easier breathing and aiding in the body’s efforts to remove harmful pathogens and irritants. A wet cough may sound “rattling” or “chesty” due to the presence of mucus in the lower airways [50].
Conversely, a dry cough, also known as a non-productive cough, does not produce mucus or phlegm [50]. It is often caused by irritation or inflammation in the upper respiratory tract, including the throat and larynx. Common triggers for a dry cough include viral infections like the common cold or influenza, exposure to environmental irritants such as smoke or dust, allergies, or even certain medications. Unlike a wet cough, a dry cough may not provide immediate relief to the irritation in the airways since there is no mucus to be cleared [2], [15]. The lack of mucus in the cough sound can give it a “hacking” or “itchy” quality [50].

The cough type of an individual often changes over time due to a variety of factors [50], [51]. Respiratory conditions and infections may evolve, initially presenting as a dry cough and later becoming wet as mucus production increases. Conversely, as the body fights off infections, a wet cough may transition into a dry cough as mucus production decreases. Environmental factors, medication use, or shifts in allergies can also contribute to these changes[44], [50]. Monitoring the cough type and its progression are essential for healthcare professionals to adjust treatment plans and ensure effective management of the individual’s respiratory health [44].

Figure 3 below shows the time domain representations of typical cough sound. Many digital signal processing techniques have been used to study and extract different quantifiable features from audio signals such as cough sounds to characterize and classify them [17].

One crucial group of features includes time-domain features, which focus on the temporal aspects of the cough sound. Duration, the time between coughs, and the cough rate are some of the time-domain features that help understand the frequency and intensity
of coughing episodes [35], [49]. Cough rise time, the time taken for a cough sound to reach its maximum amplitude, and cough decay time, the time taken for the sound to dissipate, provide insights into the dynamics of the cough. Additionally, amplitude-related features, such as peak amplitude and energy distribution across different frequency bands (such as signal RMS – Root Mean Square, or ZCR – Zero Crossing Rate), are crucial in understanding the intensity and spectral content of the cough [52].

**Figure 3. Typical Time Domain Cough Sound Signal**

Frequency-domain features are another significant category, where cough sounds are transformed into the frequency spectrum [35], [53]. Fundamental frequency, dominant frequency, spectral centroid, as well as spectral bandwidth, slope, and skewness are among the frequency-domain features that provide information about the pitch and frequency characteristics of the cough sound [54]. Furthermore, advanced features like Mel-frequency cepstral coefficients (MFCCs) or power spectral densities (PSDs) can be used to capture the spectral characteristics of the cough sound, which is particularly useful in differentiating between types of audio signals [54], [55]. One of the major challenges with cough type classification, however, is that such traditional pattern recognition features fall
short when applied to a large population of audio sounds, possibly due to the inherent environmental noise in the cough recording [7], [56].

2.2.3 Non-Contact Medical Assessment

Non-contact clinic and hospital visits have gained significant attention in recent years, driven by the need for remote healthcare solutions and the desire to minimize physical contact in medical settings [57], [58]. These visits utilize technology to enable patients to receive medical consultations, evaluations, and diagnoses from healthcare professionals without being physically present. In this context, audio analysis holds great potential as a valuable tool for enhancing non-contact healthcare applications [57].

Audio analysis can play a crucial role in non-contact clinic and hospital visits by enabling remote monitoring and assessment of patients’ health conditions. By leveraging audio signals, such as speech or respiratory sounds, healthcare professionals can gain valuable insights into patients’ symptoms and physiological states [26], [58]. For instance, in telemedicine consultations, patients can describe their symptoms, and audio analysis algorithms can help identify specific patterns or characteristics in their speech, aiding in the diagnosis or assessment of their condition. Similarly, analyzing respiratory sounds can provide valuable information about lung function, allowing healthcare providers to remotely monitor respiratory health and identify potential abnormalities. [57], [58]

In the context of non-contact visits, audio analysis can also contribute to the screening and monitoring of patients’ health remotely. By analyzing audio signals captured through devices such as smartphones or wearables, healthcare professionals can detect and monitor various medical conditions [57], [59]. For example, audio analysis algorithms can be employed to identify cough sounds associated with respiratory infections or detect heart
murmurs indicative of cardiac issues. These capabilities enable early detection and intervention, leading to improved patient outcomes[58].

Moreover, audio analysis can enhance the efficiency and accuracy of non-contact healthcare visits by automating certain aspects of the diagnosis process [51], [58]. Through machine learning and deep learning techniques, audio analysis algorithms can be trained on large datasets to recognize patterns, classify medical conditions, and provide decision support to healthcare professionals [1], [20], [32], [60], [61]. This automated analysis can help streamline the diagnosis process, enabling healthcare providers to make informed decisions more efficiently and effectively [3], [51], [62].

Furthermore, audio analysis can contribute to patient prioritization and triage in non-contact healthcare settings. By analyzing audio signals, algorithms can identify urgent or critical cases based on specific audio patterns or characteristics associated with severe medical conditions. This capability enables healthcare professionals to prioritize patients who require immediate attention, ensuring timely intervention and appropriate allocation of resources [18], [62].

The potential applications of audio analysis in non-contact clinic and hospital visits extend beyond diagnostics and monitoring. Audio analysis can also be leveraged for emotional and psychological assessment. By analyzing speech patterns, prosody, and voice characteristics, algorithms can detect signs of emotional distress or mental health disorders remotely [63], [64]. This information can aid healthcare professionals in understanding patients’ emotional well-being, enabling appropriate interventions or referrals to mental health specialists [57].
It is important to note that the successful implementation of audio analysis in non-contact healthcare visits relies on addressing privacy and security concerns. Strict protocols and encryption techniques should be employed to ensure the confidentiality and integrity of patients’ audio data [65]. Compliance with privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, is crucial to maintain patient trust and protect sensitive health information.

In summary, audio analysis has great potential both in non-contact clinic and hospital visits, and in remote healthcare delivery. By analyzing audio signals, healthcare professionals can gain valuable insights into patients’ health conditions, aid in diagnosis, monitor physiological states, and provide decision support. Audio analysis enables efficient patient screening, triage, and prioritization, ensuring timely interventions for those in need. Furthermore, it can contribute to emotional and psychological assessment, enhancing holistic care. As technology advances and audio analysis algorithms improve, the potential for audio-based healthcare applications in non-contact settings is set to expand, offering increased accessibility and convenience for patients while maintaining high standards of care.

2.2.4 Deep Learning

Deep learning has revolutionized the fields of computer vision, speech recognition, and environmental sound analysis, leading to significant advancements in these areas [8]–[10], [60], [66]. In computer vision, deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable performance in tasks such as image classification, object detection, and image segmentation [67]. These models can automatically learn hierarchical features from raw pixel data, enabling them to
recognize complex patterns and objects in images with unprecedented accuracy [3]. Deep learning has powered advancements in facial recognition, autonomous vehicle piloting, tumor and disease detection in the medical imaging field, and more [3], [8], [9], [63], [68].

Similarly, in speech recognition, deep learning models, especially CNNs and transformer-based architectures, have made substantial strides. These models can process sequential data, making them adept at capturing temporal dependencies in speech signals [3], [9], [69]. DL methods have outperformed traditional methods and significantly improved automatic speech recognition (ASR) systems [13]. As a result, virtual assistants, speech-to-text systems, and voice-controlled devices have become increasingly reliable and accessible.

Deep learning has also found applications in environmental sound analysis, where it excels at identifying and classifying various sounds, such as bird calls, car noises, or sirens. CNNs have been applied to spectrograms (using 2D CNN) or to raw audio waveforms (using 1D CNN), enabling the models to discern unique patterns and extract discriminative features from the sounds [1], [70]. This has led to the development of sound event detection systems, environmental monitoring tools, and applications in acoustic ecology and wildlife conservation.

Feature engineering and deep learning play complementary roles in the field of computer vision, particularly when dealing with image data [8]. Feature engineering involves manually designing and selecting relevant features from raw image data to represent specific visual patterns or characteristics. These handcrafted features serve as inputs to traditional machine learning algorithms, such as support vector machines or random forests, for tasks like image classification or object detection[22], [60], [71].
However, feature engineering can be a labor-intensive and domain-specific process, often requiring expert knowledge and fine-tuning. Deep learning, on the other hand, has emerged as a powerful alternative by automating feature extraction through the use of neural networks. CNNs in particular have shown a tremendous ability to learn hierarchical features from raw pixel data [60].

Deep learning models automatically discover intricate patterns and representations from vast amounts of labeled image data during the training process. This allows CNNs to capture both low-level features, like edges and textures, and high-level concepts, such as object shapes and semantics. As a result, deep learning models have outperformed traditional feature engineering approaches in various computer vision tasks [8], [70].

2.2.5 Digital Signal Processing and Spectrogram Generation

Spectrograms are graphical representations of audio signals that provide valuable insights into the signal’s frequency-content and variations over time. They are widely used in audio analysis, including speech recognition and music processing to better visualize both time and frequency content of an audio clip. Mel-frequency spectrograms or mel spectrograms, are a type of spectrogram used to represent the frequency content of audio signals in a way that aligns more closely with human auditory perception [72]. Unlike traditional spectrograms, where frequency bins are linearly spaced, mel spectrograms utilize a mel-scale, which is a perceptually uniform scale that emphasizes lower frequencies. Fig 4 shows both the mel-spectrogram and time domain representations of a cough audio signal.
The process of generating mel-spectrograms from audio signals involves the following steps [52], [54], [72]:

1. Pre-processing: The raw audio signal is first pre-processed to remove noise and ensure a consistent sampling rate. It may also be divided into smaller segments using a windowing function to reduce spectral leakage.

2. Short-time Fourier Transform (STFT): Similar to generating traditional spectrograms, the STFT is applied to the windowed segments of the audio signal to transform them from the time domain to the frequency domain. This results in the computation of the magnitude of the frequency components (spectral bins) over time.
3. Mel filterbank application: The mel filterbank is a set of triangular filters spaced along the mel scale. Each filter represents a different frequency band, and their application to the STFT magnitude values integrates the energy within each mel frequency bin.

4. Logarithm transformation: The mel spectrogram is obtained by taking the logarithm of the filterbank frequency values. This transformation compresses the dynamic range and better aligns with human perception, as our hearing is more sensitive to changes in lower frequencies.

5. Visualization: The resulting mel-spectrogram is typically visualized as a 2D matrix, where the x-axis represents time, the y-axis represents mel frequency bins, and the color or intensity indicates the energy or magnitude of each mel frequency component. (see Figure 4)

Mel-spectrograms are widely used in various audio-related applications, such as speech recognition, music analysis, and sound classification, as they provide a more intuitive representation of the audio signal’s frequency content in accordance with how humans perceive sound [1].

2.2.6 Noise and Reverberation

The presence of noise can have detrimental effects on various audio analysis tasks, such as speech recognition, music classification, and acoustic event detection [35]. In speech recognition, for example, the accuracy of recognizing spoken words can be significantly reduced in the presence of noise. Background noise can mask important speech features, making it challenging to accurately identify and transcribe spoken words. Similarly, in music classification, noise can obscure crucial musical features, leading to
misclassification or inaccurate genre identification. Figure 5 shows the effects of random additive white Gaussian noise on the spectrogram of an audio clip. Notice the overlap in frequency content of the noise and the original audio signal. Due to the additive nature of audio signals, the presence of noise can degrade the quality of the target audio signal.

Human breathing and cough sounds have a wide frequency range of approximately 200-1500Hz [15], [73]. The energy distribution is similar to that of white noise. Other than speech, generic environmental sound events may be captured in cough recordings, and have very different characteristics (duration, spectral content, and volume) with respect to background noise[74], [75]. Typically, low-pass filtering is performed to reduce the effects of noise from common environmental sources with higher frequency components. However, the effects of noise cannot be completely removed [74].

Reverberation, also known as reverb, is a common phenomenon in audio signals that can significantly impact audio analysis tasks. It refers to the persistence of sound in an enclosed space after the sound source has stopped emitting [49]. Understanding and managing the effects of reverberation are crucial for accurate and reliable audio analysis [33].

Figure 5. Effect of White Gaussian Noise on Spectrograms
When an audio signal is emitted in a space, it interacts with surfaces such as walls, floors, and ceilings, leading to multiple reflections. These reflections create a series of delayed and attenuated echoes that blend with the direct sound, resulting in a prolonged decay of sound. Reverberation can distort the temporal and spectral characteristics of the audio signal, making it challenging to accurately analyze and interpret the signal [49]. Figure 6 illustrates the process of reverberation and its effect on recorded sounds.

The impact of reverberation can be observed across various audio analysis tasks. In speech recognition, for example, reverberation can degrade the intelligibility of speech and reduce the accuracy of speech recognition systems [76]. The reflections caused by reverberation can overlap with the direct sound, creating a blurred and smeared effect on the speech signal. Similarly, in music analysis and classification, reverberation can introduce unwanted coloration and affect the timbral characteristics of musical instruments. This can lead to misclassification or inaccurate identification of musical genres or instruments.

![Figure 6. Reverberation and Space](image)

Both noise and reverb are naturally occurring in the audio domain and cannot be fully eliminated in audio recordings [77]. There are many ways to reduce the effects of noise and reverb both physically (e.g., acoustically non-reflective panels for sound booths)
and digitally (e.g., noise reduction techniques). Although they are often undesirable artifacts that are picked up by the recording equipment, they can sometimes be desired, for example in music recordings. There are many software toolboxes that allow control of the magnitude of acoustic effects that can be digitally added to audio clips, including noise and reverb. Both natural and synthetic noise and reverb have their applications and can be used creatively in various audio-related tasks to provide a more immersive auditory experience [78].

2.2.7 Deep Learning Models

DL systems can learn meaningful information from a variety of types of data. DL models have been successfully used in text or natural language processing (NLP), speech recognition, and computer vision/image-processing applications with great success [8], [20]. DL models are at the forefront of pattern recognition research and they continue to help extract information and inferences from text, images, video, speech, sound, and other complex datasets [56]. DL models are being used in real-world applications such as autonomous vehicle control systems, factory assembly lines, general sciences, genetic sequencing, healthcare, defense, and many more [10], [20], [79].

One of the key advantages of DL models is transfer learning [56]. Models can be pretrained on large, available datasets, with huge computational power, and the knowledge gained by these models can be harnessed on a new problem domain where only limited labeled data are available. By fine-tuning the pretrained DL model on the new dataset, one can achieve domain adaptation to the new problem. This step of fine-tuning is often less computationally costly than the initial pretraining and requires fewer labeled training exemplars [66], [80].
2.2.8 Regularization and Overfitting

When a classifier is trained on a given dataset, it is hoped that it will perform just as well on unseen data. This is not always the case. Regularization is a technique in ML and statistics used to calibrate models to minimize the adjusted loss function and prevent overfitting or underfitting [81]. Overfitting is when a model performs well on training data but has a drop in performance on unseen data. This is not preferred as the resulting model is not generalizable to new data. There can be many reasons for overfitting, including poor training data quality, small datasets, or if the model is too complex. Regularization introduces a penalty term to a model’s objective function to discourage it from assigning high importance to any one feature or group of features. There are different types of regularization that can be performed, the choice of which depends on the specific use case. Modified loss function method (MLF) is when the loss function of a model is modified to account for the output distribution. Ridge (L2) and Lasso (L1) are two types of MLF regularization techniques [8], [70], [81]. Modified training algorithms such as ‘dropout’ or ‘injecting-noise’ are regularization methods that, as the name suggests, modify the training of the classifier. Dropout is when connections between nodes in consecutive layers are randomly dropped based on a dropout-rate. These are randomly adjusted for each iteration of the training cycle. In the injection method, another parameter called weights, is adjusted in the training cycle. Modified sampling method (MSM) is a regularization method that is applied when limited datasets are available. The input data to a DL model is first adjusted to provide a fair representation of general input distribution that the model may face. Data augmentation and k-fold cross validation are two types of MSM regularization techniques. This thesis focuses on the data augmentation method.
2.3 Critical Literature Review

2.3.1 Audio Classification Using Deep Learning

People today are surrounded by audio processing technology, which is part of everyday objects such as smart-home devices and smartphones with AI assistants [1], [70]. The process of making meaningful information out of speech commands involves AI technology and requires the audio to first be processed into a form that can be used for input to the DL models. A high-level flowchart of the process is shown in Figure 1.

The initial step of the process is often raw audio samples that need to get pre-processed. This step often involves digitization, down-sampling, normalization, windowing, noise-cancellation, and audio segmentation [20]. Different methods, all falling under the digital signal processing umbrella, can be used for each of these steps (e.g., various windowing techniques are available). The choice of method used often depend on the application. Following this audio-preprocessing step, different audio features are extracted from the audio segments that are then used as inputs to the DL models. The DL models are often trained on a subset of these features. Once trained, the DL models can be used to make predictions and be used as classifiers to identify events occurring from the test inputs.

Once the raw signal is pre-processed, useful information must be extracted to use as input to any DL model. There are several methods (e.g., principal component analysis or gradient descent algorithms) that are used to identify the most useful features for a specific task [56], [70]. Similarly, the type of DL model to use is also important, as certain architectures lend themselves better to certain types of data and so are better suited for
certain tasks. Figure 7 below shows different DL architectures that have been used in DL audio classification tasks [20].

![Figure 7. Illustration of Various DL Architectures [20]](image)

Generally, DL models can be trained using supervised learning, unsupervised learning, and reinforcement learning [56], [70], [82]. Supervised learning requires labeled datasets to train the models for classification and prediction tasks [1], [70]. The labeled training datasets are used to adjust model weights to tune the DL models to their specific task. Typically, in such cases, the training datasets must be large (n=1,000+), and it is often a challenge to acquire such data. When labeled data are not available, unsupervised learning methods can sometimes be employed. These types of DL models use unlabeled datasets, without human input, and can often extract patterns and clusters that were otherwise hidden in the data. Such DL models are often used in many real-world applications where very large unlabeled datasets are available, or where labeling the datasets manually become very resource intensive [1]. Finally, reinforcement learning includes types of DL models that are trained based on a reward or punishment basis. A model must select actions in the environment and an (often delayed) reward or punishment either encourages or discourages the learned behaviour. Such DL models have been
explored in many areas, such as game design, autonomous control systems, statistics, and multi-agent systems [1], [83]. This thesis uses supervised machine learning exclusively.

Deep Neural Networks (DNNs) are networks that involve more than three layers from the input to the output of the model/network [1], [84]. DNNs are able to process large datasets and determine patterns quicker than older pattern-recognition algorithms [67]. Some examples of DNNs include CNNs, recurrent neural networks (RNNs), and generative models (GMs) such as variational autoencoders (VAEs) and generative adversarial networks (GANs) [67], [82].

GANs and VAEs have been used extensively in audio and speech processing and have been used to synthesize speech. Similarly, due to the temporal nature of audio signals, RNNs have also been used to study audio signals and analyze the sequential nature of audio [3]. Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are two RNN structures that have been used in speech recognition tasks. However, LSTMs are often computationally complex and require substantial time and data to train. Similarly, due to their sequential architecture, RNNs have inherent sequential dependencies, complicating deployment on parallel computing to reduce computational cost/time.

Finally, CNNs have been extensively used in image processing with exceptional performance. Unlike RNNs, the sequential dependency is absent in CNNs [32]. Since CNN models have multiple convolutional layers followed by a dense layer, CNNs often require less memory or less parameters for their classification tasks. CNN models have been used in different audio applications such as automatic speech recognition, music classification, and many more [2], [32], [82]. This makes CNN models ideal for studying audio signals, as is the main focus of this thesis work.
2.3.2 Cough Sound Classification

As discussed, cough sound analysis provides healthcare professionals with important clinical information to better diagnose and treat patients [25], [85]. Cough sound analysis research can be broadly separated into two parts: cough detection and cough classification.

Cough detection primarily refers to the identification of cough sounds from background noise or from other environmental sounds. There have been a variety of studies that have had meaningful success using various models to detect cough sounds. Matos et al. used a hidden Markov model (HMM)-based method to detect cough sounds from other sounds with 82% accuracy [86]. In 2011 Tracey et al. demonstrated accuracies of over 88% using MLP (multi-layer perceptron) and SVM (support vector machine) models to detect coughs from pulmonary tuberculosis patients [51]. In 2018, Hoyos-Barcei et al. developed a k-nearest-neighbour detection method that was shown to have 93% accuracy [87]. Monge-Álvarez et al. used frequency-band-specific features to get 92.7% sensitivity and 88.6% specificity in cough detection by 2019 [6]. Orlandic et al. developed a similar cough detection algorithm in 2021 that was able to detect COVID-19 coughs with a probability of above 80% [34]. Over the years there has been a steady improvement on the ability to accurately detect cough sounds to near perfect levels. Classification of cough type, however, is a different task.

Cough classification involves the investigation of the cough sounds themselves to extract useful information from them. In 2015, Amrulloh et al. demonstrated the use of time domain audio features to differentiate between pneumonia and asthma in patients by observing their cough sound characteristics [88]. In 2017, Sharan et al. used Mel-
Frequency Cepstral Coefficients (MFCCs) from cough sounds to identify pediatric croup patients [89]. There can be many physiological or pathological causes of cough that can be used to describe them [27]. Characteristics from the cough sounds themselves provide important information to physicians about respiratory health and the underlying cause, helping with clinical diagnosis. In 2022, Cohen-McFarlane et al. successfully identified patients with adventitious sounds, such as whooping cough and stridor, from other cough sounds [28]. However, these methods involved traditional feature engineering and pattern recognition methods and were not necessarily generalizable to other respiratory diseases or coughs [15]. They further investigated the use of face-masks on cough sound characteristics and how it may affect cough classification [2], [28], [90].

A more challenging problem is distinguishing between two typical cough types that have far more similarity in their audio spectral content, such as wet and dry coughs. Murata et al. applied MFCCs to discriminate between wet and dry coughs [91]. Chatrzarrin proposed the development of two engineered features from the cough sounds that were able to distinguish between wet and dry coughs in a small dataset [48]. Schroder et al. used Gabor filters banks and an HMM classifier to show better performance than using MFCCs [92]. Moradshahi et al. studied the effects of noise and reverberation on cough classification and proposed a microphone array system to help with noise reduction for better cough classification results [35], [49]. Similarly, in 2019, Bhide et al. developed a method for noise suppression in cough sounds using SVM [93]. These all focused on traditional machine learning methods and the extraction of features that require a good understanding of the field and data. In recent years, with the development of DL models, the research community is moving away from traditional pattern recognition and machine
learning techniques such as feature engineering. The use of DL models reduces the reliance on domain expertise or human intervention as long as large sufficient data are available for training purposes.

In a clinical setting, many sensor technologies have been adopted to monitor respiration and cough characteristics and frequency. These systems have an inherent expectation of idealized environments [19], [55], [79]. Generally, the distinction between wet and dry cough types has been studied extensively due to its clinical value [41], [91]. Many patients with respiratory diseases such as asthma, chronic bronchitis, COPD, and pulmonary fibrosis have had their coughs studied to identify key features in the spectral content of their cough sounds [47]. Moreover, data augmentation techniques must be adopted to account for the relatively small training dataset sizes for cough sound classification [11], [79].

With the spread of the COVID-19 pandemic, the research community put a large emphasis on contactless-cough-classification, and it was clear the technology had to be robust, generalizable to a global scale, fast with predictions, and indifferent to various types of recording devices, patient types/demographics, and environmental influences [19], [61], [94].

2.3.3 Data Augmentation for Training Deep Learning Models

It is well known that, for DL models to learn, a large training dataset is vital. Even though audio classification datasets are increasing, it is still far behind the DL vision datasets such as ImageNet [95]. Publicly available datasets such as AudioSet and Speech-Command-dataset include many hundreds of audio classes, but relatively few labeled examples of each class [95], [96]. Nonetheless, expert-labeled datasets for the specific
medical task of cough classification are not readily available. Many have used the AudioSet or Environmental Sound Classification (ESC-50) datasets and extracted the cough class for analysis [5], [97]. Efforts have been made in recent years to accumulate datasets of respiratory sounds. The Coswara dataset was developed by the Indian Institute of Science (IISc) Bangalore as a large-scale, publicly available dataset that focuses on respiratory sounds and speech recordings [98]. Collected during the COVID-19 pandemic, it comprises of audio samples from individuals with respiratory conditions of either COVID-19 positive or healthy subjects. Since the focus of this thesis is broader respiratory health conditions and cough types, this dataset was outside the scope of this thesis. The COUGHVID dataset was compiled with the same intentions of studying respiratory sounds, particularly COVID-19 [34]. This is one of the datasets that is used in this thesis. Finally, other smaller datasets such as the NoCoCoDa and MIA-COVID-19 are also publicly available to further the understanding of disease monitoring through the development of machine learning models [99], [100]. Such datasets are often far too small to properly train DL models, and it is well known that inter-rater reliability is often very low, as the task of cough classification is very subjective [28], [90].

Data Augmentation is the task of increasing the size of a dataset for the purpose of training a DL classifier. There are often two paths for doing this: 1) generating new exemplars based on the existing data or 2) generating synthetic data from scratch.

The latter approach often uses GAN-type networks to generate synthetic datasets [9], [64]. Typically, such models can generate new data that are different from the original, natural dataset. Conversely, data augmentation from existing data involves the use of data transformations. Data transformations are the process of providing slight changes to a
datapoint such that it can be considered a different datapoint, without affecting the
underlying pattern of the data such that the label stays true. For example, Figure 8 shows
the same image of a cat with many random image transformations (shifts, rotations, zoom,
flip, etc), such that each individual image is different enough to be considered a separate
datapoint, while still being recognizable as its label (‘cat’).

![Figure 8. Data Augmentation in Computer Vision][101]

As can be inferred from Figure 8, the data transformation methods suitable for
augmentation rely on the data domain; applying random rotations, zooms, and crops on
spectrogram images would change the nature of the underlying audio data, while such
transformations can be quite suitable for image analysis.

There are certain traditional transformations to audio signals that are suitable for
data augmentation. Time-stretching (TS) the audio signal involves lengthening the duration
of an audio signal by a certain factor, without changing the pitch of the audio. Conversely,
pitch-shifting (PS) scales the pitch of an audio by a certain factor without changing the
tempo of the clip. A spectral augmentation method called SpecAugment proposed by
Google, is a natural extension of the Cutout and Random Erasing methods [13], [81], [102].
This method involved visualizing the spectrogram of an audio clip and masking or blocking
random time and frequency segments on it to create an augmented exemplar [102]. The constituent parts of this method can be called frequency masking (FM) or time masking (TM).

2.3.4 Research Gap

DL models have shown much better, faster, and more accurate performance than traditional pattern-recognition algorithms and classifiers. However, the large dataset sizes required to train or fine-tune DL classifiers to specific tasks is a challenge. Datasets of 10,000+ datapoints are quite typical for developing DL models, and, although such datasets exist in the natural language processing (NLP), speech recognition, and computer vision analysis domains, there is a lack of such datasets in audio classification for medical purposes. Most curated, well-labelled cough datasets range in the hundreds of datapoints. So, until the research and healthcare community can curate sufficiently large datasets with which to train DL classifiers, alternative techniques must be explored to fully leverage the small datasets currently available.

This thesis plans to address this problem with audio data augmentation techniques in cough classification. Due to the scarcity of large, dependably labelled audio data, DL models often overfit and do not generalize well. Data augmentation methods have been shown to counter this effect, making DL models more generalizable, while solving the problem of small labelled datasets.

Another issue with large audio datasets is the variability in recording equipment, environmental effects and audio quality. In real-world applications, when collecting cough-sound recordings, it is often impossible to gather studio-quality audio clips. DL classifiers must account for the noise and reverberations that may affect the quality of the audio clips.
In the literature today, traditional audio data augmentation methods are often applied, but they are not natural to the problem domain. The novel methods of augmentations examined in this thesis consider the nature of the audio signals themselves. Noise and reverberation are unavoidable artifacts that are picked up from environmental factors when trying to record an audio clip; thus, they are ‘natural’ to the problem domain. Typically, efforts are made in the pre-processing stage of audio processing to reduce noise and reverberation. However, this work treats noise and reverb as data augmentations. Adding synthetic noise and/or reverb to create, in effect, new datapoints, may help train a better DL model that is generalizable.

This thesis aims to study the effects of reverberation and noise in audio classification and how this affects the choice of DL classifiers.
Chapter 3: Data Augmentation with Reverb and Noise for Cough Classification

Portions of this chapter appear in the following publication, on which I am the first author:

[33]

3.1 Introduction

Figure 9. Dry (left) and Wet Cough (right) Waveforms in Time Domain

The previous chapter provided detailed information about the literature on cough sound classification and the research gap that needs to be addressed. A trained physician can often determine the underlying condition by listening to a patient’s cough sounds. Typical auscultation includes the use of a stethoscope to listen to chest sounds between coughs, however, the cough sounds are often heard by the naked ear. The automated analysis and classification of cough sounds, or modeling the physician’s ability to determine cough type, has many health applications including remote health care for people aging in place [2], [103]. Cough analysis can also be used to monitor the effectiveness of treatment and track changes in a patient’s health over time. Figure 9 shows the time-domain plot of a typical wet and dry cough. Recent advancements in machine learning, coupled with the pandemic, have shown an increased interest in developing DL models for cough
analysis that can be implemented in real-world applications. These systems have the potential to significantly improve patient screening and assessment for respiratory illnesses [6], [22], [23], [37], [72], [94].

A spectrogram provides a visual representation of the frequency content of an audio signal over time. Spectrograms can be used to identify and isolate specific frequency components of a sound, and to measure their relative strength, duration, and timing [72]. In this thesis, all spectrograms use the log-mel-scale. Figure 10 shows the spectrograms of the same two wet and dry coughs as shown in Figure 9. Most of the high intensity components of the cough occur at lower frequencies and in the early stages of the cough. There is substantial overlap between these intensities in both cough types. However, higher frequency components tend to attenuate more quickly for wet coughs than for dry coughs [16]. Additionally, resonance can be exhibited in wet cough sounds, which typically does not appear in dry coughs (note the three whiskers/tails visible in the right/wet spectrogram in Figure 10) [22], [48], [104].

![Figure 10. Cough Spectrogram](image)

Typical spectrograms of dry cough (left) and wet cough (right). Spectrograms have the time along the x-axis and frequency along the y-axis, and the color represents the signal strength (lighter colors indicate high power, while darker colors show lower power values). These spectrograms have the axes removed as they are used as input to the ResNet18 model.
Any recorded sound will include some degree of noise and reverberation. In an acoustical sense, these undesirable perturbations are always present in the natural environment and cannot be avoided. Imperfect transduction and recording of sound as well as the recording equipment or the room where the sound is captured, all generate noise and reverb. For recorded coughs, noise may also be generated by other human activities or environmental sources. Some typical sources of noise in cough sound datasets include air conditioners, fans, television noises, typing, other individuals speaking, etc. This chapter focuses on comparing data augmentation methods and how they performed on cough classification when used to generate augmented training data. A comparison will be made between traditional augmentations and natural augmentations. The methods used are described in the following.

3.2 Methods

3.2.1 Dataset

The data used in this work was collected from two main sources [28], [48]. The cough sounds were obtained from Chatrzarrin et al., whose dataset contains physician and expert labelled events [48]. This dataset contains 27 physician-labelled wet coughs and 19 dry coughs, along with expert labelling of whooping coughs and restricted breathing sounds that were further added to the dataset by a technician trained in respiratory sound analysis [28], [31]. The focus of this thesis was primarily on the wet and dry cough labels from the dataset.

3.2.2 Deep Learning Model

All analysis was done in Python (v3.9). The audio signals were sampled with a sampling frequency of 44.1kHz. Audio samples were lowpass filtered with a cut-off
frequency of 4kHz, before converting to a mono WAV file. This follows standard preprocessing of cough sounds [22], [35], [48], [104], [105].

This study used ResNet18, a deep CNN model that was pre-trained on the ImageNet dataset [29], [106]. This is a residual network model with 18 layers, which is one of the smaller ResNet models. It was chosen for its size, as it allows the system to be computationally less expensive and makes it more favourable for embedded systems and future real-world applications. The final fully-connected layer was adjusted from its built-in 1000 classes to the 2-class (‘dry cough,’ ‘wet cough’) task. Transfer learning was used to fine-tune the model for the classification of cough sound spectrograms.

3.2.3 Audio Pre-Processing

The audio files were pre-segmented into individual cough sounds and spectrograms were generated from the audio signals. When working with audio inputs to DL models, the duration of the segments needs to be considered. All segments must be the same length so that the DL model does a fair comparison. Additionally, they should be long enough to include sufficient data for the model to learn from the spectrogram. Following the standard protocol for audio datasets in this field, a length of 1 second was chosen for the audio segments and zero-padding was used where necessary. To generate the spectrograms, the python librosa library was used with the following parameters: (n_fft=2048, hop_length=128, n_mels=224) and the power spectrum was converted to decibel units to get the final ‘log-mel-spectrogram.’ These generated spectrograms were a single channel image of 224x224 pixels. However, for ease of viewing, the spectrogram plots in this thesis use the ‘viridis’ colormap. ResNet18 models require a three-channel
input (224x224x3), and so the same single-channel spectral images were stacked to generate the three-channel form of the desired input dimensions for the ResNet18 model.

The original dataset and the class size distributions are shown in Table 1. The classes are relatively balanced. The overall dataset size is relatively small for DL applications (n=46), hence the use of a pretrained ResNet18 model and transfer learning, rather than training a DL model from scratch. A repeated hold-out test protocol was used throughout this study. In order to avoid data leakage, care was taken to ensure that coughs from the same patients did not appear in both the test and train splits prior to separating the dataset. Subsequently, in each iteration, the data were first split into training and testing sets using a uniform random 80-20 split. The test data was set aside and untouched until the evaluation of the model. The train-test splits were all done prior to any augmentations.

### 3.2.4 Augmentations

The *Torch-Audiomentations* Python library ([https://github.com/asteroid-team/torch-audiomentations](https://github.com/asteroid-team/torch-audiomentations)) was used to add all data augmentations to the data. This python library was designed to add audio augmentation to test audio products. It standardizes the testing, streamlines the use of GPU-resources in the process. The library has several pre-built functions that can be used to apply the desired audio transformation. Furthermore, multiple transformations can be applied in series to given data. Two sets of experiments were run using data augmentation on the training set. The first used traditional data-augmentation techniques including:

- Time-Masking (TM): This involves randomly masking a portion of the audio signal in the time domain (range for band part: [0.1s, 0.15s] – This defines the min and
max of the masked portion. So for TM, a single portion of up to 50ms of time is masked at a time to generate the augmented datapoint).

- Frequency-Masking (FM): This method randomly masks a portion of the frequency spectrum of the audio signal (range for fraction: [0.05, 0.10] – This defines FM to have a single masked range, between 5-10% of the entire frequency range of the audioclip).

- Pitch-shifting (PS): This involves shifting the pitch of the audio signal by a random factor (range for semitones: [-4, 4] – semitones range from -12:12. This defines a single pitch-shift applied that can range from min -4: max 4).

- Time-Stretching (TS): This method randomly stretches or compresses the audio signal by a factor in the time domain (range for rate: [0.8, 1.25] – this is also a uniteless rate, that ranges from 0.1:10, where a rate >1 means the audio is sped-up. This given range was determined from the available literature, and provides a single, random scale factor for the TS-augmented clip).

Table 1 Training and Test Dataset Distribution

<table>
<thead>
<tr>
<th>Class</th>
<th>Traditional Augmentations</th>
<th>‘Natural’ Augmentations (Reverb and Noise)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No Augment.</td>
<td>TM</td>
</tr>
<tr>
<td>Dry cough</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>Wet cough</td>
<td>21</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>Test Set</td>
<td></td>
</tr>
<tr>
<td>Dry cough</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Wet cough</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
The original training dataset with no augmentation (N/A set) was kept as a control against which to compare the augmentation methods. Each of the traditional data augmentation methods was applied to the training dataset to form four separate datasets (TM, FM, PS, and TS datasets). The augmentations were run until the dataset size reached the size shown in Table 1. The augmented training datasets were all kept of the same size to ensure that there were no confounding factors due to dataset size when comparing the augmentation methods.

In the second set of experiments, reverberation and Gaussian noise were used as data augmentation techniques on the audio signals. The augmentations that were applied were:

- Only Reverberation (OR): RT60 is an acoustic measure defined as the time it takes for sound pressure to reduce by 60dB. The minimum and maximum RT60 for three ranges that were selected were equivalent to a small studio, an average room and a large room (ranges used: [0.15s, 0.3s], [0.4s, 0.6s], [1s, 1.5s]). Equal importance was put on each of these ranges when generating the augmented datasets.

- Only Noise (ON): Two ranges of Gaussian noise were defined by signal-to-noise-ratio (ranges used: [0dB, 5dB], [5dB, 20dB]). The given ranges defined the min and max of the possible SNR value in decibels that was selected to generate the ON-augmented datapoint. Equal importance was placed on both ranges, to result in a balanced ON-dataset.

- Reverb and Noise (RN*): All combinations of reverb and noise defined before were applied to this training dataset. This set defined the final training dataset size, and all other augmentation methods were applied multiple times to balance the different training datasets.
The training sets for each of these groups were kept separate. A ResNet18 model was trained on each of these training sets separately and evaluated on the test sets with a repeated hold-out method. The models were named using the data-augmentation method used to train them, for ease of reference. Table 1 shows the resulting training dataset sizes that were used in the second experiment in each iteration of the repeated hold-out test. Test data were not augmented. The augmented audio files inherited the class labels from their original source clip. Spectrograms were generated from all audio clips. Figure 11 (dry) and Figure 12 (wet) show examples of the spectrograms with and without the random augmentations.

The Adam optimization algorithm was chosen for finetuning the ResNet18 models, as it has been shown to surpass the other optimizers in audio classification domain [22], [23]. The maximum epochs were set to 15, as the dataset is relatively small, and this
allowed the network to sufficiently explore the feature map from the provided inputs. The order in which training samples were presented was shuffled for each epoch. The learning rate was set to 0.01 and was determined empirically.

![Figure 12. Spectrograms of Training Set Wet Coughs.](image)

From top to bottom, the images are a) with no augmentation, b) with random reverb augmentation, c) Gaussian noise-only augmentation with random SNR value, and d) both random Gaussian noise and reverb augmentation. Note the tapering of high frequency content and the presence of resonance at low frequencies.

Based on the different training sets, ResNet18 models were fine-tuned with the training data from each of them. Once training was completed, each model (TM, FM, PS, TS, OR, ON, RN* and N/A) was evaluated on the test set, and the model predictions were compared to the actual class labels from the test set. The accuracy of the models was noted, and the process was repeated 20 times. With each iteration, the train-test splits were randomly assigned, as were the random augmentation values. The overall accuracies are reported in the Results section.
3.2.5 COUGHVID Dataset

With a CNN model developed, fine-tuned on an augmented training set of cough sounds, its performance on a larger, publicly sourced database of cough sounds was evaluated. The COUGHVID dataset is a collection of crowd-sourced cough sounds, created with the aim of supporting research and development related to respiratory health, particularly in the context of the COVID-19 pandemic [34]. The dataset was compiled during the early stages of the pandemic to aid in the study of cough sounds as a diagnostic tool for COVID-19 and other respiratory diseases.

The dataset consists of over 25,000 cough recordings obtained from individuals with a variety of respiratory conditions, and healthy individuals [34], [107]. The dataset was crowdsourced using an online public platform in 2020, ensuring a diverse set of samples representing different cough types and variations. Users were asked about their age, gender, geo-location, and to self-report their respiratory health condition.

From this crowdsourced dataset, a subset of under 3000 cough sounds were assessed by 4 experienced physicians. These experts scored the cough sounds, providing their medical diagnosis and assessment of the cough sounds. Table 2 below shows the type of annotations provided by the experts. Since the data was crowdsourced and the user identity was not provided with the dataset, it was not possible to ensure data leakage in the subsequent train-test splits. However, since the subset of cough sounds is so much larger in this case, it was of lesser concern.

It is well known that such labelling, although being the gold standard of respiratory health assessment, is highly subjective and based on physician assessment. The inter-rater reliability is often low. In the COUGHVID dataset, the inter-rater reliability was assessed
for the four experts on the cough sounds for which all experts had provided labels, and this finding was corroborated by calculating the kappa value between the expert labelers in the COUGHVID dataset (Cohen’s kappa for cough_type, $K_{\text{Fleiss}} = 0.26$) [19], [34].

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Options</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>Good</td>
<td>A general assessment of the cough sound recording</td>
</tr>
<tr>
<td></td>
<td>Ok</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No_cough</td>
<td></td>
</tr>
<tr>
<td>Cough_type</td>
<td>Wet</td>
<td>Cough type assessment</td>
</tr>
<tr>
<td></td>
<td>Dry</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>Dyspnea</td>
<td>True</td>
<td>Dyspnea detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Wheezing</td>
<td>True</td>
<td>Wheezing detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Stridor</td>
<td>True</td>
<td>Stridor detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Choking</td>
<td>True</td>
<td>Choking detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Congestion</td>
<td>True</td>
<td>Congestion detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Nothing</td>
<td>True</td>
<td>Nothing detected</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Upper_infection</td>
<td>An assessment of the patient’s respiratory health condition.</td>
</tr>
<tr>
<td></td>
<td>Lower_infection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Obstructive_disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COVID-19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Healthy_cough</td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td>Pseudocough</td>
<td>A general impression from the expert about the severity of the cough</td>
</tr>
<tr>
<td></td>
<td>Mild</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

Despite the low inter-rater reliability, the COUGHVID dataset is attractive due to its relatively large size. The exploration of data augmentation conducted on the cough dataset described in Section 3.2.1 was repeated on the COUGHVID dataset, where
Cough\textunderscore type was chosen as the class label to be predicted. The ResNet18 models used in this section were all pretrained on ImageNet data and further retrained on the COUGHVID data from the training subset; the trained model in the previous section was not used here. All augmentations were performed after the random train-test split of the dataset.

In order to evaluate the performance of augmentation-trained models to classify cough\textunderscore type, the subset of cough sounds evaluated by all four experts was extracted from the COUGHVID dataset, and the CNN model was used to predict cough\textunderscore type on it. This subset contained labels by each of the 4 experts (union set) for each datapoint and was used for all the COUGHVID evaluations in this thesis. Since there was no singular ground-truth for the cough sounds, the CNN model classification was compared against each expert assessment individually as ground truth labels and the Receiver Operating Characteristics curve’s Area-Under-the-Curve, (ROC)-AUC, for all four comparisons are shown and discussed in the Results section.

The expert label with the highest AUC from this comparison was chosen as the overall ground-truth label and the earlier experiments were repeated. Using this ground-truth label, the COUGHVID dataset had traditional data augmentations along with the proposed noise-and-reverb augmentation, and their overall performances are discussed in the following section. Finally, the proposed noise-and-reverb augmentation method was applied on each of the four expert-labels, and the performances were compared.
3.3 Results

Figures 11 and 12 show samples of the data-augmented spectrograms that were used as input to the DL model. The amount of reverb and/or noise added to the audio clip is random. The addition of noise is often more clearly visible, as it affects the entire frequency range in the spectrogram. The effect of reverberation is harder to visualize on the spectrogram; with sufficiently large reverb, the higher amplitude components of the audio clip become visibly expanded along the time-axis.

The addition of augmented data increases the size of the training datasets. It was observed that the models trained on augmented data took slightly longer to train than the set with no-augmentation. Once trained, the models were evaluated on the test set and the accuracy noted for comparison before repeating the process. With each iteration, a randomized 80-20 train-test split was used.

3.3.1 Traditional Augmentation

The first set of experiments compared the traditional augmentation methods against using no-augmentation. The aggregate accuracies of these models are shown as boxplots in Figure 13. The average accuracy of the model trained on N/A was 0.76, TM was 0.90, FM was 0.85, PS was 0.77, and TS was 0.82.
3.3.2 Novel Augmentation (Noise and Reverb)

The second set of experiments used reverb and noise augmentations. The aggregate accuracies of these models are shown in the boxplot in Figure 14. The average test accuracies of the model trained on N/A was 0.76, OR was 0.78, ON was 0.81, and RN* was 0.95.
The boxplots show the accuracies of the different models. The four models where training data were augmented in different ways: No augmentation (N/A), Reverb augmentation only (OR), Noise augmentation (ON), Reverb&Noise augmentation (RN*). The experiment was repeated twenty times and the resulting accuracies are shown in the boxplots.

The performance of the N/A model was the lowest. This may be due to the small training dataset size that resulted in the ResNet18 model having insufficient information from which to learn. Moreover, the resulting accuracies had one of the widest 95%-confidence intervals, ranging from 0.72-0.81. This suggests that there is a significant amount of uncertainty in the performance of this model.

All models that used augmentation showed improvement in the model performance. Among the traditional augmentation techniques, the TM and FM models had the highest average accuracy. Even though the PS and TS models improved on the accuracy of the N/A model, their confidence interval is comparably wide (0.76-0.84, 0.78-0.88).
Both noise and reverb augmentation were shown to be separately effective, increasing average test accuracies to a comparable degree (OR and ON models). Even though the noise-augmented training set had a slightly higher average accuracy, the 95% confidence intervals of both these models were similar (0.75-0.84 and 0.76-0.82, respectively). The wide range suggests some uncertainty in the model’s performance and that the result may be dependent on the train-test split of the original dataset.

The average accuracy of the RN* model (using both reverb and noise augmentation) was the highest overall. Moreover, this method had the narrowest confidence interval, ranging from 0.93-0.97. This indicates that the model is more robust and is not dependent on the train-test split of the original dataset. It suggests that the ResNet18 model has learnt the underlying features from the spectrograms to reliably classify wet and dry coughs.

### 3.3.3 COUGHVID Dataset

The ROC curves for each of the four experts’ labels against the trained-CNN model’s predictions are shown in Figure 15. The model had highest agreement with Experts 2 and 4, with AUCs of 0.82 and 0.90 respectively. This indicates that, if these experts’ labels were considered ground-truths, then there is an over 82% chance that the CNN model is able to distinguish between the cough types. The lowest AUC was 0.64, with Expert 3. As established earlier, the inter-rater reliability of these experts’ labels in the dataset is low, as is the case in most crowd-sourced, large datasets of cough sounds. A high correlation with each labeler cannot be achieved with low inter-rater agreement.
Since Expert 4 had the highest AUC with the DL model predictions, it was chosen as the ground-truth labels in the COUGHVID dataset. The earlier experiments that compared traditional augmentation with the proposed noise-and-reverb augmentation methods were repeated on this larger dataset, and the results are shown in Figure 16 below. The average accuracies of the model trained on N/A was 0.73, TM was 0.72, FM was 0.72, PS was 0.74, TS was 0.73, and novel RN* was 0.87. The traditional augmentation methods did not increase the performance of the models significantly and they were not consistent (increased 95%-confidence intervals). The average performance of all models that used traditional data augmentation techniques was comparable to the model with no-augmentation. The best traditional augmentation was PS (Pitch-shifting), with an average accuracy of 0.74. The proposed RN* augmentation method, on the other hand, had a higher average accuracy (0.87). Similar to the findings earlier in Chapter 3, the RN* augmented
model had the highest overall average accuracy, and a narrow confidence interval (0.81-0.91). This emphasizes the robustness of the RN*-augmented model and its potential generalizability to other, larger datasets.

Figure 16. COUGHVID – Model Accuracy Boxplots using Expt4 Labels
Model accuracies: The boxplots show the accuracies of the different traditional augmented DL models using Expert 4 as ground-truths. Training data were augmented in different ways: No augmentation (N/A), Time Masking (TM), Frequency Masking (FM), Pitch Shifting (PS), and Time Stretching (TS) and the novel combined reverb and noise (RN*) augmentation. The experiment was repeated ten times and the resulting accuracies are shown in the boxplots.

Figure 17 displays the performance comparisons of using each of the expert labels as ground truths. The average accuracy in pair-wised comparisons of no-augmentation and novel RN* augmentation for Expert 1 is 0.70 – 0.74, Expert 2 is 0.66 – 0.81, Expert 3 is 0.75 – 0.72, and Expert 4 is 0.72 – 0.87. As seen before, the highest performance is with Expert 4’s labels as ground truth (0.87), followed by Expert 2 (0.81), then Expert 1 (0.74) and, finally, Expert 3 (0.72). In three out of four of the comparisons, the proposed RN* augmentation method showed improvement over no-augmentation. Only in the case of
Expert 3’s labels did the augmentation method produce lower average accuracy than the no-augmentation method (0.72 vs 0.75).

The best performances were seen with the RN* augmentation method using Experts 4 and 2s’ labels as ground-truths. This is indicative that the DL augmentation-trained model was able to closely emulate the behaviors of these two human experts in their cough-type determination. The repeatability of this experiment on a larger dataset of cough sounds, coupled with its high performance on two different human-expert labels, highlights the robustness and generalizability of the novel RN* augmentation approach proposed in this thesis.

![COUGHVID Dataset - RN Augmentations](image)

Figure 17. COUGHVID – Comparison of Accuracies Using Each Expert-Label as Ground Truth and using Noise and Reverb Augmentation.

The boxplots show the accuracies of the different models. The reverb and noise augmentation technique was compared to the no-augmentation model, when using each Expert label information as ground truth. The average accuracies for Expert 1 (NA1: 0.70, Ex1: 0.74), Expert 2 (NA2:0.66, Ex2:0.81), Expert 3 (NA3: 0.75, Ex3: 0.72), and Expert 4 (NA4: 0.72, Ex4: 0.87) are compared.
3.4 Conclusion

A novel approach to spectral data augmentation for training DL cough sound classifiers was described and compared against traditional forms of audio data augmentation. The combined use of both reverb and noise augmentation provided the best results, indicating that the combination is particularly effective in improving the model performance. Not only did the RN* model have the highest average accuracy, but it also had the narrowest 95%-confidence interval. This implies that the combined augmentation allowed the model to robustly learn the underlying features to classify cough types. Considering that reverberation and environmental noise are naturally occurring data transformations, it is expected that models trained using these forms of augmentation will generalize to audio recordings in real-world scenarios.

As mentioned in sections 2.2.7 and 2.3, the development of DL models has made significant advancements in machine learning methods. Traditional ML models used feature engineering methods to develop features that were calculated from the audio clips. The use of a DL model, negates the need of traditional feature engineering methods that were previously applied to traditional ML research in this area. Instead of fitting an array of features as inputs to the DL models, the entire spectrogram is used as an image input to the classifier, and the task of feature extraction is handled within the DL model. Madison et. al [28] used the same dataset that was investigated in chapter 3. The cough classification accuracy reported using an AlexNet model was 83%. The findings in this paper, with the use of data-augmentation have improved on previous DL model analysis on the same dataset (accuracy= 95% with RN*).
The RN* augmentation-trained CNN model was further investigated by comparing its predictions with multiple expert-human raters on a different dataset called COUGHVID. Performance of a model that is trained on one dataset tends to decrease when applied to a different dataset that it has not been trained on before. Furthermore, the agreement between expert-physician assessments of cough_type is often low, as it is a subjective, qualitative assessment gained from years of experience. Nevertheless, the DL classifier was able to achieve good agreement with the human expert predictions, as shown by the high AUC values with Experts 2 and 4. Additionally, on repetition of the experiments on the larger COUGHVID dataset, the novel RN* augmentation method displayed higher performance the majority of the time. It showed notable improvement on average accuracies over no-augmentation methods with Experts 2 and 4. This confirms that the augmented DL classifier was able to distinguish between wet and dry coughs at a human-expert level on a blind-dataset of cough sounds.

The present research describes a novel improvement upon the application of machine learning for audio signal classification for cough classification. Incorporating these types of augmentation in the training process better prepares a model to handle these conditions in real-life situations. It makes the resulting model more robust to such naturally occurring phenomena. Overall, this demonstrates the importance of the underlying nature of the dataset and of reflecting it in data augmentation to achieve high performance for audio-based classification tasks. These insights can be applied to other similar tasks to improve the performance of deep learning models.
Chapter 4: Alignment Between Train and Test Environments

4.1 Introduction

The previous chapter demonstrated that training models using both noise and reverberation as means of data augmentation led to highly accurate classifiers for a noise-free test environment. However, any such system would ultimately be deployed into realistic environments, characterized by potentially high levels of both environmental noise and reverberation. This chapter explores the performance of DL models in a variety of test environments, representing realistic deployment environments. Furthermore, it explores the alignment between training and testing environments in terms of audio noise and reverberation levels.

Since the focus of this chapter is broader, with a focus on realistic environments, a larger dataset of definitively labeled audio sounds, called the “Speech Commands” dataset, was used throughout the experiments [96]. This is a large, mature dataset of keywords spoken by different individuals that has been curated for keyword spotting. This is a much larger dataset (100,000+ datapoints) and so is very suitable for DL models. Additionally, there is no ambiguity with the labels and they can be easily verified by listening to the audio clip. There is also a lot of overlap in the frequency and power distribution of speech and cough sounds [103], [108]. All these reasons combined makes Speech Commands an ideal dataset to use for DL models. The dataset, augmentation techniques, and model training and performance evaluation are described in the following sections of this chapter.
4.2 Methods

4.2.1 Dataset

Speech-Commands is an audio dataset publicly released by Google in 2017 to help further research in DL model training [96]. It includes over 65,000 audio clips of individual words spoken by many different people. The audio clips are each 1 second long, and include a singular class per clip, such as “yes,” “no,” “up,” ‘down,” etc., along with their class labels. The dataset’s main target is to improve keyword spotting models [96]. Since cough sounds are extremely short in nature, and cough types are often a binary classification problem, for the purposes of this thesis, two short classes were selected from speech command for experimentation (‘yes’ and ‘no’).

4.2.2 Augmentations and Deep Learning Model

The experiments were run on the same ResNet18 classifier that was pretrained on ImageNet data as used in Chapter 3. Similar to the previous chapter, the final layer was adjusted to the two-class problem (‘yes’ and ‘no’) for the speech-command classification task. Transfer learning was used to further fine-tune the model for classification of the speech-command classes.

The audio clips were already set to 1-second-long segments, so further segmentation and zero-padding was not necessary in this case. The pre-processing steps to generate the final log-mel-spectrograms that were used as input to the DL models are the same as described in the methods section in Chapter 3.

The torch-audiomentations Python library was also used here for the reverberation and noise augmentations for these experiments. There were three sets of experiments conducted. The idea was to explore how well a DL classifier performs at different levels
of noise and/or reverberation present in the test environment, and how that performance relates to the data-augmentation strategy used in the fine-tuning stage of the DL classifiers. This was then further compared to the novel reverb and noise (RN*) augmentation that represents one of the main contributions of this thesis.

In the first experiment, only noise augmentations were considered and in the second experiment, only reverberation augmentations were observed. In the third, different combinations of reverb and noise augmentations were studied. Tables 3 and 4 show the levels of noise and reverberation that were defined for these experiments using the torch-audiomentations library.

For each experiment, the dataset was first split into training and testing sets using a uniform random 80-20 split. The test and training sets were kept separate so that the classifier was blind to the test set datapoints.

Table 3  Noise-Augmentation Level Description

<table>
<thead>
<tr>
<th>Label</th>
<th>Noise Augmentation (SNR range)</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>N0</td>
<td>–</td>
<td>No noise</td>
</tr>
<tr>
<td>N1</td>
<td>10 dB-20 dB</td>
<td>Low noise</td>
</tr>
<tr>
<td>N2</td>
<td>5 dB-10 dB</td>
<td>Medium noise</td>
</tr>
<tr>
<td>N3</td>
<td>0 dB-5 dB</td>
<td>High noise</td>
</tr>
</tbody>
</table>

There were four levels of noise considered; N0: no noise augmentation, N1: low noise augmentation, N2: moderate/medium noise, and N3: high noise. In Experiment 1, for each of these augmentation levels, a separate copy of the training set was defined and used to fine-tune four individual DL classifiers: DL-N0, DL-N1, DL-N2, and DL-N3. The Adam optimization algorithm was used for fine-tuning (learning rate = 0.01, epochs = 40). In order
to investigate all possible noise-levels, the test-sets also had four noise-augmentations performed, to generate four separate test sets, and each DL classifier was evaluated on each test-set. Model accuracy was selected as the evaluation metric and it is shown in the color chart in Figure 18.

Experiment 2 followed the exact steps of Experiment 1, but with reverb augmentation instead of noise. The input variables for audio augmentations’ reverb augmentation are shown in Table 4. Similarly, the accuracy metric was noted and the results are shown in Figure 19 and discussed in the Results section.

Table 4 Reverb Augmentation Level Description

<table>
<thead>
<tr>
<th>Label</th>
<th>Reverb Augmentation (RT60 range)</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0</td>
<td>–</td>
<td>No reverb</td>
</tr>
<tr>
<td>R1</td>
<td>0.15-0.3 s</td>
<td>Low reverb (Studio)</td>
</tr>
<tr>
<td>R2</td>
<td>0.4-0.6 s</td>
<td>Medium reverb (Small room)</td>
</tr>
<tr>
<td>R3</td>
<td>1.0-1.5 s</td>
<td>High reverb (Concert hall)</td>
</tr>
</tbody>
</table>

Finally, in Experiment 3, combination of both reverb and noise augmentation was considered. In this case, there were a total of 16 possible permutations for levels of synthetic noise and reverb. The accuracy metric results are shown in Figures 20 and 21, organized by quantization of reverberation and noise, respectively. This would give a more granular view of the DL classifiers’ performance on different levels of reverb and noise present in the test dataset. Additionally, the smaller permutations of augmentations in the test-set were compared with the overall proposed RN* augmentation (that used the original wider ranges of noise and reverberation as in Chapter 3), and the results are displayed in the figures as well.
To further investigate the performance of the overall RN* classifier in the different test-set scenarios, the accuracies of the classifiers were ranked within each scenario, and the results are displayed graphically in Figures 22 and 23 below.

4.3 Results

4.3.1 Varying Noise Levels

The first experiment compared the alignment of noise in the test and train environments. The accuracies are shown graphically in Figure 18. A general conformation along the diagonal is observed, with the highest accuracy value of 0.95 seen when both training and test environments have N3 (high noise) augmentation. The lowest accuracy values (0.43 and 0.48) are observed when the training and test environments have the most opposing combination of noise present: training=N0 and test=N3, or train=N3 and test=N0. This reinforces that, if a DL classifier is trained on the same/similar level of noise as the test-environment, it will have better performance in classification.

Figure 18. DL Classifier Accuracies for Varying Noise Augmentation
Four levels of Noise Augmentation were applied to four DL classifier training sets, and their prediction accuracies on four test-environment augmentations are displayed. The Noise Augmentation labels are described in Table 3.
4.3.2 Varying Reverberation Levels

The second experiment compared the same alignment, but with reverb augmentation (Figure 19). The same diagonal dominance is observed here; however, the accuracies along the diagonal are much higher (ranging from 0.8-0.97) on average. This indicates two things. First, in both cases, fine-tuning a DL classifier to some value of noise or reverb improves the chances of better predictions by the classifier. Second, the CNN DL classifier being assessed is more sensitive to reverberations than to noise.

![Figure 19. DL Classifier Accuracies for Varying Reverb Augmentation](image)

Four levels of Reverb Augmentation were applied to four DL classifier training sets and their prediction accuracies on four test-environment augmentations are displayed. The Reverb Augmentation labels are described in Table 4.

4.3.3 Combined Noise and Reverberation in the Test Environment

Experiment 3 did a comparison of the combination of noise and reverb augmentations in the train and test environments. Figures 20 and 21 show the results of the experiment. This gives a detailed picture of the relationship between data augmentation
and the train-test environments, more specifically about the quantization of reverb and noise augmentations. Overall, there are 16 possible combinations of data-augmentation and test scenarios. The higher accuracy along the diagonal emphasizes the same relationship as was seen in Figures 18 and 19 but provides a much larger picture. With the combination of noise and reverb, the accuracies of the DL model are higher in a wider range of combinations than was seen with either data augmentation method alone.

Figure 20. DL Classifier Accuracies for Varying Noise & Reverb Augmentation
Augmentation were applied to 16 DL classifier training sets and their prediction accuracies on 16 test-environment augmentations are displayed. The Augmentation labels are described in Tables 3 and 4. The novel RN* augmentation and its predictions on the same 16 test environments are shown. The augmentations are sorted by increasing Reverb first, then Noise.
The highest accuracy values were seen in the following:

- training:R0N2 & testing:R0N2 (acc: 0.97),
- training:R1N1 & testing:R2N1 (acc: 0.97),
- training:R2N0 & testing:R2N0 (acc: 0.98),
- training:R2N1 & testing:R2N0 (acc: 0.96),
- training:R2N1 & testing:R3N1 (acc: 0.98),
- training:R3N1 & testing:R3N1 (acc: 0.96)

As can be seen from these results, the most performant models tend to have good alignment between the training and testing environments in terms of noise and reverb levels. This was further corroborated in Figure 21.

The next portion of the experiments focused on the overall RN* classifier and how that compared to the other fine-tuned classifiers’ performances in each of the 16 scenarios. To investigate this, the classifiers were ranked within each test scenario. The rank results are graphically displayed in Figures 22 and 23. The rank of the RN* classifier ranged from 10-16, staying in the top 60% overall. Moreover, the RN* classifier ranked between 14-16 (above 87.5%) in 50% of the scenarios. This was particularly noticeable in the higher reverb test-scenarios (R2 and R3) as well as the high noise scenario (N3). This highlights that, in half the possible scenarios, the RN* classifier is ranked in the top 3 classifiers, particularly in the non-zero noise and reverb scenarios. In the remaining scenarios, even though the RN* classifier was not the best, it still performed well compared to the other classifiers. This emphasizes the robustness of the RN* model and its potential in real-world scenarios where noise and reverb is ever-present.
Figure 21. DL Classifier Accuracies for Varying Noise & Reverb Augmentation

Augmentation were applied to 16 DL classifier training sets and their prediction accuracies on 16 test-environment augmentations are displayed. The Augmentation labels are described in Tables 3 and 4. The novel RN* augmentation and its predictions on the same 16 test environments are shown.

The augmentations are organized by Noise first, then Reverberation.
Figure 22. Ranked DL Classifier Accuracies for Varying Augmentation (1-Reverb, 2-Noise)
Figure 23. Ranked DL Classifier Accuracies for Varying Augmentation (1-Noise, 2-Reverb)
4.4 Conclusions

As expected, results indicate that models trained on data from environments most similar to the deployment environment exhibit the greatest performance. However, in practice, this requires accurate characterization of the test environment to select the optimal model. There can be a very large number of noise and reverb characterizations and it may not be feasible or practical to have trained/fine-tuned classifiers for each possible test environment. Furthermore, recording environments may have fast changing noise and reverb characteristics. In order for a DL classifier to correctly match the test-environment in such situations, it would require high computational power and storage on the deployment device, which is not possible in real life scenarios. As an alternative solution, these results show that DL models trained on a randomized mixture of augmented environments exhibit robust performance across a wide range of test environments. This latter solution is simpler and may be better suited to edge deployment on devices with limited computational power.
Chapter 5: Conclusions

5.1 Conclusions

This thesis focused on cough sound classification in the presence of noise and reverberation in the environment. Noise and reverb were selected as data augmentation techniques since these are inherent in any audio recording and are natural perturbations that are unavoidable in audio recordings. A pretrained-ResNet18 classifier was used as the backbone to the DL model studied in this research. Both noise and reverb were synthetically generated and added to the training dataset to fine-tune the CNN classifier to cough classification.

The novelty of this thesis is the use of reverb and noise as data augmentation techniques, to train data for a DL model. Although a pre-built python function was used for the actual augmentation (audiomentations), the library’s purpose was not for data augmentation. The application of the library’s functions is novel.

The performance of the CNN classifier was analyzed with various data augmentation methods, including traditional audio augmentation methods as well as the novel noise and reverb augmentations described in previous chapters. The experiments showed that the combination of noise and reverb augmentation greatly improved the performance of the CNN model to classify between wet and dry coughs. This was further studied in a larger public dataset of crowd-sourced cough sounds that were labeled by multiple expert physicians. Despite the inter-rater agreement being low in the human-evaluation of the cough sounds, the CNN classifier’s performance was comparable to the expert predictions.
Finally, the amount of synthetic data augmentation in the audio clips was studied to understand the relationship between them in the training and test environments. Reverb augmentation provided better accuracies than noise augmentation alone. Furthermore, if the CNN classifiers’ data-augmentation could match the test-environment, then it produced the best performances.

5.2 Summary of Research Contributions

The following are the research contributions made during the development of this thesis research:

Contribution 1: Developed a CNN-based cough classification model based on a pretrained ResNet18 architecture. Assessed the model on a well-curated, physician-assessed cough dataset.

Contribution 2: Evaluated several spectrogram data augmentation techniques, including traditional audio augmentation as well as noise and reverb augmentations in terms of their classification accuracy.

Contribution 3: Proposed a novel noise and reverb augmentation technique in spectral data augmentation for cough classification.

Contribution 4: Demonstrated that the combination of noise and reverberation as data augmentation for fine-tuning a CNN classifier is a highly effective method of spectral data augmentation, resulting in improved classification accuracy and stability. The combination of reverb and noise presents a novel approach to data augmentation for cough sound analysis.

Contribution 5: Evaluated the performance of DL models trained on different ranges of noise and reverb augmentation to match the noise and reverb augmentations in
the test-environment. The best performances can be seen when the DL models are trained on environments that closely resemble (or match) the test environment. A mixture of noise and reverberation data-augmentation in the training set provides better chances for the DL model’s higher performance across a wide range of test environments.

5.3 Recommendations for Future Work

Here is a list of possible future work that is relevant to the current research work:

- Exploring other natural noise types that can be leveraged as data augmentation to fine-tune DL classifiers.
- Expanding the research to other DL classifier architectures, including, but not limited to, transformer models and combination CNN-transformer models.
- Expanding the dataset of expert-labeled cough sounds.
- In the COUGHVID dataset, only the union of where all expert labels were available was investigated. The established RN* model’s performance can be further assessed on other subsets of the COUGHVID dataset, particularly those labeled by Expert 4.
- Collecting patient data, such as physician diagnosis, treatment, and follow-up, along with cough sounds throughout the treatment. This could provide two levels of analysis: 1) The physician diagnosis could be used instead of cough types as expert-labels to allow the DL classifiers to learn underlying/deeper patterns between cough spectrograms and diseases, and 2) the longitudinal collection of cough sounds may allow the DL classifiers to study respiratory disease progression.
References


