DESIGN OF MULTI-MODALITY DEEP FUSION ARCHITECTURE FOR DEEP ACOUSTIC ANALYTICS

A Dissertation
IN
Computer Science
and
Computer Networking and Communication System

Presented to the Faculty of the University
of Missouri–Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
ZEENAT TARIQ

M.S. Computational Science and Engineering,
National University of Sciences and Technology, Islamabad, Pakistan, 2012

Kansas City, Missouri
2021

DESIGN OF MULTI-MODALITY DEEP FUSION ARCHITECTURE FOR DEEP ACOUSTIC ANALYTICS

Zeenat Tariq, Candidate for the Doctor of Philosophy Degree

University of Missouri–Kansas City, 2021

ABSTRACT

There is increasing attention for audio classification research to support various emerging applications, including environmental monitoring, health care, and smart city. Audio classification is an important area of research that needs more dynamic and be adapted to possible changes in the environment, facilitate the adoption of enhancement techniques with innovative and effective solutions. While there is an increasing amount of audio data available for environmental audio classification, we still face significant challenges to conduct accurately deep learning in the environmental and health audio domain. These challenges may occur due to the various field and categories, e.g., environmental, animal sounds, noises, and human body sounds. Specifically, distortion, fracture, and audio data noise are the primary obstacles affecting the accuracy of environmental audio classification. There have been many advancements in the correct detection of audio sounds; the audio data depending on how the features are extracted from the audio, how modality represents the audio data, and how much noise is present in raw audio.

With the rapid increase of environmental datasets, extracting relevant data to create an adequate environmental sound classification is a crucial challenge. Deep learning is an advanced and promising solution to detect, predict and classify different types of sounds. Convolutional neural network is growing most in audio classification using environmental and health audio data. The classification technique distinguishes between different types of sounds by capturing different patterns across time and frequency and applying them to different features. For the neural network model to perform efficiently, the training requires a large amount of data to get better training. Many researchers are working on these ideas today, but the research is still not mature enough due to the lack of available datasets. Moreover, in audio classification domains, noise in sound data or unstructured data may affect the classifier's performance. The data for audio classification is too complicated to understand multiple characteristics and latent patterns of data.

We have developed unique fusion architectures based on convolutional neural networks for conducting multi-feature multi-modality fusion-based audio classification to solve these problems in acoustic classification. We have proposed the multi-modality fusion architecture with Deep Acoustics (DA) and Multimodal Deep Acoustics (MDA). The contributions are (1) The performance with multi-modality of the input audio clips was influenced by data issues such as noise or imbalance. (2) We have extracted various acoustic features for the multi-features fusion network, such as Log-Mel Spectrogram, Chromagram, and Mel Frequency Cepstral Coefficient. (3) We have developed effective data augmentation and normalization methods to enhance the quality and comparability of sparse audio samples with various extended features. (4) The proposed models have been evaluated with two types of fusion approaches: multi feature-based model fusion and network-based fusion. Our experimental results validate the benefits of our proposed work for audio classification tasks. Mainly, it is confirmed that our fusion models with multi-feature and multi-modality are very efficient. In numerous benchmark datasets, the suggested models outperformed state-of-the-art solutions, according to our comprehensive testing. The suggested deep acoustic analytics approaches have been used in the environmental sound detection and healthcare domains to detect lung and heart conditions. Furthermore, our models show better results even with small network models (i.e., less convolutional and hidden layers, fewer trainable parameters, a smaller number of epochs, and less time consumption) than the results of previous methods available in the literature.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled "Design of Multi-modality Deep Fusion Architecture for Deep Acoustic Analytics," presented by Zeenat Tariq, candidate for the Doctor of Philosophy degree, and hereby certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Yugyung Lee, Ph.D., Committee Chair
Department of Computer Science Electrical Engineering

Farid Nait-Abdesselam, Ph.D.
Department of Computer Networking and Communication System

Ghulam M. Chaudhry, Ph.D.
Department of Electrical Computing Engineering

Md Yusuf Sarwar Uddin, Ph.D.
Department of Computer Science

Ahmed M. Hassan, Ph.D.
Department of Electrical Computing Engineering

CONTENTS

ABSTRACT iii

ILLUSTRATIONS ix

TABLES xii

ACKNOWLEDGEMENTS xvi

Chapter

1 INTRODUCTION 1

2 DEEP ACOUSTIC MODEL 7

2.1 Related Work 9

2.2 Deep Audio Analytics 13

2.3 Experimental Works and Results 19

3 MULTIMODALITY FOR DEEP ACOUSTICS 31

3.1 Introduction 31

3.2 Related Work 33

3.3 Multimodal Acoustic Method 36

3.4 Experimental Works and Results 43

4 DEEP NETWORK-BASED FUSION FOR DEEP ACOUSTICS LEARNING . 53

4.1 Introduction 53

4.2 Related Work 56

4.3 Methodology 64

4.4 Results and Evaluation 79

4.5 Conclusion 102

5 FEATURE-BASED FUSION LEARNING FOR DEEP ACOUSTICS 111

5.1 Introduction 111

5.2 Milestone 112

5.3 Related Work 113

5.4 Methodology 118

5.5 Result and Evaluation 136

5.6 Comparison with State-of-the-Art Research 158

5.7 Conclusion 161

6 NETWORK MODEL DISCUSSION 165

6.1 Multimodal Deep Acoustic (MDA) 165

6.2 Network Fusion Model (FDA-NET) 166

6.3 Feature-Fusion Model (FDC-FS) 166

7 CONCLUSION AND FUTURE WORK 168

7.1 Summary of Deep Acoustics Learning 168

7.2 Summary of Multimodality for Deep Acoustics Learning 169

7.3 Summary of Network-based Fusion for Deep Acoustics Learning 169

7.4 Summary of Feature-based Fusion for Deep Acoustics Learning 170

7.5 Future Work 170

BIBLIOGRAPHY 172

VITA 184

ILLUSTRATIONS

Figure Page

1 Overall Workflow for Deep Acoustic Analytics 6

2 An overview of Proposed Deep Acoustic Model 14

3 Classification Model for Deep Acoustics 20

4 Classification Accuracy for Lung Sounds 25

5 Classification Accuracy for Heart Sounds 27

6 Classification Accuracy for UrbanSound8k 29

7 Workflow for Proposed Multi-modal Classification Model 38

8 STFT Obtained from Original Wav files a) Spectrogram Obtained from Original Lung Sound Data, b) Spectrogram Obtained from Normalized Lung sound data, c) Spectrogram Generated from Augmented Lung Sound Data 42

9 Spectrogram Visualization of Heart Sounds 45

10 Model Training Accuracy and Loss for Image-based Lung Sounds 47

11 Model Training Accuracy and Loss for Image-based Heart Sounds 51

12 Overall Architecture for our Proposed FDA-NET Model 69

13 Spectrogram for ESC-10 five categories 71

14 Spectrogram for ESC-50 five categories 72

15 Spectrogram for UrbanSound8k 74

16 Architectures: (a) FDA-NET-1, (b) FDA-NET-2, (c) FDA-NET-3 79

17 Fusion Network: (a) FDA-NET-1, (b) FDA-NET-2, (c) FDA-NET-3 . . . 80

18 Workflow of Fusion-based Audio Classification 84

19 FDA-NET Model Comparison (Precision, Recall, F1-Score) for Overall Models for Environmental Classification 88

20 Model Comparison for Fusion Based Environmental Classification 91

21 FDA-NET Audio Model Accuracy and Validation Accuracy for FDA-NET-1, FDA-NET-2 & FDA-NET-3 for Original Datasets 95

22 FDA-NET Audio Model Loss vs Validation Loss for FDA-NET-1, FDA-NET-2 & FDA-NET-3 for Original Datasets 96

23 Fusion Based Disease Classification Architecture 119

24 Lung Sound Features: (1) Wav (2) Spectrogram (3) MFCC (4) Chromagram123

25 Heart Sound Features: (1) Wav (2) Spectrogram (3) MFCC (4) Chromagram124

26 Overall FDC-FS Architectures (a) FDC-1 (b) FDC-2 (c) FDC-3 131

27 t-SNE Visualization: (a) Lung Sound Dataset (b) Heart Sound Dataset . . 144

28 Lung Condition Classification (Accuracy vs. Loss): Fusion Network Model (FDC-FS) with Features: (a) Spectrogram (b) MFCC (c) Chromagram 147

29 Accuracy for Lung/Heart Condition Detection (Original and Augmented Datasets) 150

30 Heart Condition Classification (Accuracy vs. Loss): Fusion Network Model (FDC-FS) with Features: (a) Spectrogram (b) MFCC (c) Chromagram 152

31 Accuracy for Lung and Heart Condition Detection 155

32 Class Wise Accuracy for Lung Condition Detection 162

33 Class Wise Accuracy for Heart Condition Detection 163

TABLES

Tables Page

1 Original and Augmented Data Size for Lung Sounds 22

2 Dataset File Distribution for Original Heart Sounds 22

3 Dataset File Distribution and Augmented for Heart Sounds 23

4 Dataset File Distribution and Augmented for UrbanSounds8k 23

5 Literature Survey for Environmental Sound Classification 62

6 State-of-the-Art Accuracy in Urban Sound Classification (Image & Audio Approaches) 63

7 FDA-NET Model Architecture and Parameters 83

8 Urban Sound and Environmental Sound Datasets 83

9 Model Comparison for Fusion Based Environmental Classification 89

10 FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 93

11 FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 93

12 FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 94

13 FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 104

14 FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 104

15 FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset 105

16 FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 105

17 FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 106

18 FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 106

19 FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 106

20 FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 107

21 FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset 107

22 FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 107

23 FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 108

24 FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 108

25 Audio Classification Benchmark Dataset 109

26 Audio and Image Input (Original and Augmented) 109

27 FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 109

28 FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 110

29 FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset 110

30 Literature Survey for Lung Sound Classification 129

31 Literature Survey for Heart Sound Classification 130

32 FDC-1 Model Hyper-parameters 139

33 FDC-2 Model Hyper parameters 140

34 FDC-3 Model Hyper-Parameters 141

35 FDC-FS Model Hyper-Parameters (1) 142

36 FDC-FS Model Hyper-Parameters (2) (Continued) 143

37 Lung Sound Dataset: Original and Augmented Data 143

38 Heart Sound Dataset: Original and Augmented Data 143

39 Lung & Heart Condition Detection with Original Data (Testing Accuracy: Weighted Average) 146

40 Lung & Heart Condition Detection with Augmented Data (Testing Accuracy: Weighted Average) 146

41 Class Wise Accuracy for Lung Dataset 148

42 Class Wise Accuracy for Heart Dataset 153

43 State-of-the-art Lung Classification Models 156

44 State-of-the-art Heart Classification Models 157

45 Overall Results for Multimodal Deep Acoustics 167

46 Overall Results for Network Fusion Model (FDA-NET) 167

47 Average Runtime for Environmental Sound Classification 167

48 Overall Results for Feature-Fusion Model (FDC-FS) 167

49 Average Runtime for Lung and Heart Disease Classification 167

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor, *Dr.Yugyung Lee*, for convincingly guiding and encouraging me to be the best researcher. This dissertation would not have been completed if it hadn't been for her constant guidance. I want to express my deep and sincere appreciation to my committee members *Dr. Farid Nait-Abdesselam* (co-discipline advisor), *Dr. Ghulam M. Chaudhry*, *Dr. Ahmed Hassan*, and *Dr. Md. Yusuf Sarwar Uddin*.

I would like to express my gratitude to the School of Computing and Engineering and the School of Graduate Study for providing me with the opportunity to complete my study at the University of Missouri-Kansas City in the United States.

I would like to thank my husband, *Dr. Sayed Khushal Shah*, for his unwavering support during my studies, his immense love and sincerity throughout. I would also like to thank my sons *Hanzalah Hilal Shah*, and *Hadi Shah* who joined us in my Ph.D. journey, giving me unlimited happiness and pleasure.

I am forever indebted to my father *Tariq Jan Khan*, and mother *Romina Khanzada* for giving me the opportunities and experiences to make me who I am. I would like to thank them for supporting me throughout my life and leaving me free in all my decisions. I would also like to thank my brothers *Toheed Jan*, and *Momin Khan* for their immense love, support, and prayers.

CHAPTER 1

INTRODUCTION

The advanced technologies are essential to achieving the improvement of lifestyle and health care. Machine learning is one of the most promising techniques used for analytic and making cities healthy. Moreover, deep learning is a branch derived from machine learning. It is known for allowing the computational models, which consist of several layers of processing used to learn the data representations over multiple levels of abstractions. It has attracted a lot of attention due to its high performance in prediction and classification. These learning techniques are among the fastest-growing fields nowadays in the area of audio classification.

Audio classification plays an essential role in the environmental and medical domain. This system monitors and detects acoustics and sounds in the environment and improves the lifestyle. People suffer from noise problem in the environment, which affects their daily routine work efficiency. On the other hand, healthcare expenditures make the life of many people stressful nowadays. The growing expenses of health care that are increasing with time in the United States are almost double the time as it is happening in most developing countries around the globe.

More specifically, an extensive investigation in a partnership among researchers, health care providers, and patients is integral to bringing precise and customized treatment strategies in taking care of various diseases. Due to this fact, using deep learning for advanced analytic helps in reducing costs and improving the health care system. These classifiers outperform humans due to the ability to ignore noise and memory issues. For example, a traditional approach uses a stethoscope to detect different sounds from the human body through which disease is detected. However, it takes much time to analyze the condition manually, and there may involve a high risk of missing data by physicians. Current work approaches for classification of audio and diseases through audio signals are based on data obtained from costly machines, and no cost-effective real-time based approach is addressed via deep learning. The manual process or obtaining scans has significant drawbacks, i.e., expensive and error-prone.

The overall view of our proposed method is shown in Figure 1. A brief abstract of each chapter is as follows:

Chapter 2: Deep Acoustic Model

Recent advances in deep learning (DL) have improved the state-of-the-art results of the data-driven approaches and applications in a wide range of domains. However, building robust classifiers with diverse datasets is one of the most significant challenges to deep learning researchers. With the advent of datasets, deep learning technologies for audio classification have recently received a lot of attention. There is increasing attention for audio classification research that aims to support various emerging applications. One of challenges in the research is finding the data that is publicly available and cleaning the data that are not recorded properly and cannot be accepted if it is given as an input to a class. Further, deep learning relies on large amounts of data. Due to limited amount of publicly available data, the researchers are facing difficulties in adding their input in this field.

Contributions

- Designing and implementing the advanced normalization and augmentation techniques for sound data
- Audio-based approach to design and implement an integrated deep learning network model.
- Evaluation on benchmark lung sound, heart sounds, and speech emotion, and environmental data.

Chapter 3: Multimodality for Deep Acoustics

Deep learning technologies have received significant attention for large-scale image classification in the area of computer vision. The characteristics of sounds for classification are too complicated to understand the hidden patterns of data. The image-based sound classification was introduced to effectively captures the diverse patterns in the dataset. With the continuous development and innovation of algorithms for image processing in many fields such as medical and environmental, deep learning has become an important research direction. Specifically, distortion, fracture, and noise of audio data are the primary obstacles affecting the accuracy of environmental audio classification. Advanced technologies are essential to achieving the improvement of lifestyle and health care.

Contributions

- Generate images from audio feature original and augmented data.

- Construction of an integrated deep learning network.
- Multimodality based approach to network model for audio and generated image data
- Low-cost Application of multimodality to human disease classification
- Evaluation using two heart sound, lung sound and environmental sound data.

Chapter 4: Deep Network-based Fusion for Deep Acoustics Learning

The environmental sound classification (ESC) is an important area of research that needs more dynamic and be adapted to possible changes in the environment, facilitate the adoption of enhancement techniques with innovative and effective solutions. While there is an increasing amount of audio data available to environmental audio classification, we still face significant challenges to conduct accurately deep learning in the environmental audio domain. These challenges may occur in particular due to the various field and categories, e.g., UrbanSound8k, ESC-10 and ESC-50 etc. Specifically, deep learning plays an essential role in audio classification over traditional machine learning nowadays. Convolutional neural network is growing most in audio classification using Environmental audio data. The data for audio classification is too complicated to understand multiple characteristics and latent patterns of data. As new technologies like image-based audio classification were introduced, more diverse fusion approaches.

Contributions

- Designing and Construction of models to be trained on image and audio data.

- Fusion-based approach with single feature to deep acoustic classification.
- Evaluation for audio-based fusion approach using three benchmark datasets for classification.
- Evaluation for image-based fusion approach using three benchmark datasets for classification.

Chapter 5: Feature-based Fusion Learning for Deep Acoustics

There have been significant recent advances in deep learning and the potential of the deep learning model for various medical applications. Recently, there has been increasing attention for the classification of human body sounds for clinical conditions in the medical domain. Advanced technologies are essential to achieving the improvement of lifestyle and health care. More specifically, an extensive investigation in a partnership among researchers, health care providers, and patients is integral to bringing precise and customized treatment strategies in taking care of various diseases. This chapter aim to propose a feature-based fusion model transferred from different feature-based convolutional neural network models to classify lung and heart disease.

Contributions

- Multi-Feature approach using audio data.
- Designing and Construction of multi-models to be trained on generated images.
- Feature based deep network fusion to deep learning classification.

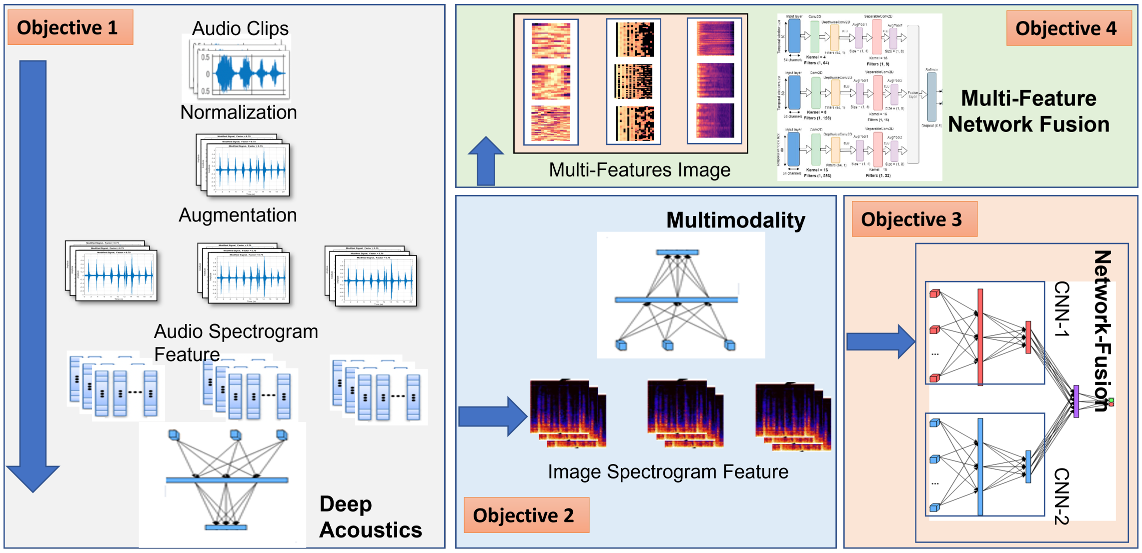


Figure 1: Overall Workflow for Deep Acoustic Analytics

- Evaluation for image-based fusion approach using two benchmark datasets for classification.

CHAPTER 2

DEEP ACOUSTIC MODEL

Deep learning (DL) improvements have enhanced the state-of-the-art results of data-driven methodologies and applications in a variety of disciplines. One of the most promising techniques for audio classification is machine learning (ML). Building strong classifiers with a variety of datasets, on the other hand, is one of the most difficult tasks for deep learning researchers. Deep learning approaches for audio classification have recently gotten a lot of attention due to the availability of datasets. Audio classification research is growing rapidly as it intends to support a variety of new applications, such as environmental monitoring, health care, and smart cities [1, 2, 3, 4].

ML approaches have been used to detect the type of urban noise in acoustic environments in some cases. These machine learning approaches are currently the fastest expanding areas in the field of audio classification [5, 6, 7]. DeepEar is an audio sensing model that can also be used in acoustic contexts to classify sounds for ambient sceneries, emotion recognition, stress detection, and speaker identification [8]. Deep Learning techniques based on Convolutional Neural Networks (CNN) was used for environmental sound classification [9].

Finding publically available data and cleaning data that has been improperly recorded and cannot be accepted as an input to a class are among of the research's obstacles. Data normalization, also known as feature scaling, is the process of rescaling or standardizing data according to data standards [10]. The first step to perform classification in the sound domain is data preprocessing. For normalizing the sounds, the low values and high values are rescaled in such a way that all the sound clips have an average and similar values based on clipping the peaks. Because of directly recording audio from one source, the audio samples may have some noise coming from the any other sounds/source that exist in the environment. Applying normalization minimizes those factors and allows us to recognize and classify the sounds easily via deep learning.

Large volumes of data are required for deep learning. Researchers are having difficulty contributing to this topic due to the restricted number of publicly available data. To deal with the issue. We presented a solution called as it Data Augmentation in the deep learning domain. Augmentation of data [11, 12, 13] is a technique that adds various aspects to the original audio samples, such as slowing down or speeding up the recording, which may be useful in assessing the intended subject's audio and removing unnecessary noise from the sound clip. Changing the pitch of a sound can also enable the model recognize the audible sounds that it is interested in.

To address these issues, we offer our model, which is based on a popular deep learning network called the Convolutional Neural Network (CNN). For an effective sound classification, we propose various advance preprocessing approaches such as normalization and augmentation. The spectrogram features generated from the audio dataset are used to classify the data. Noise in the audio samples, which is caused by environmental influence, causes the classification results to vary.

The contributions of this chapter can be summarized as follows:

- The key contribution of our work is an effective way of building an integrated classifier with advanced pre-processing techniques.

- Our methodology and model ”Deep Acoustic Model” is generic enough to be used for multiple dataset in multiple domain.

- The data augmentation techniques were applied to generate more diverse training data. Besides, a normalization approach was proposed for normalizing noise signals of vast environmental and human body sound datasets.

- We extracted spectrogram features and labels of the annotated sound samples and used them as an input to our 2D Convolutional Neural Network (CNN) model.

- A comprehensive evaluation of the proposed model has been conducted using the four benchmark datasets mainly related to the environmental and health.

- Our work aims to assess the degree of accuracy acceptable in the multiple domain specially in medical field by utilizing deep learning to available data.

- Our preliminary results shows that our model outperforms state-of-the-art results in terms of learning time and accuracy. DA model improved performance compared to the existing models. [14, 15]

2.1 Related Work

There are many existing techniques available that demonstrate the usability of audio classification using diverse datasets. The section below has all work that has been done for the audio classification with diverse datsets.

2.1.1 Deep Learning for Audio Classification

The analysis of environmental sound in [16], which is close to our work, has been performed where the authors proposed Convolutional Neural Network (CNN) to classify environmental sounds. The architecture consists of two convolutional rectified layer unit by applying max pooling, two fully connected hidden layers, and a softmax output layer. The datasets used for classification were environmental science and the UrbanSound8k dataset [17]. Random time delays and pitch shifting were used to enhance the data. Mel Spectrograms were extracted from all audio files, re-sampled, and normalized with varied window sizes using the librosa implementation. The classifier, one of the datasets, and features utilized were comparable to those used in ourÂ research; however, we developed an integrated network classification model for sound classification that included enhanced normalization and augmentation approaches. Multiple datasets from various domains, such as the environment and human body noises, are used to test our model. We also compared our findings to those of other researchers.

Salamon and Bello extended the work and presented the data augmentation technique in [4] for environmental sound classification using Deep Convolutional Neural Network. The deformation of audio was performed through time stretching, pitch shifting, dynamic range compression, and background noise. Davis et al. [18] compared and tested the UrbanSound8K data based on augmentation. Each augmentation strategy is applied to the convolutional neural network separately, and the results are compared. We utilized data augmentation on a normalized method based on two separate parameters, and the results were significantly improved.

The main limitation of the research described in [19] is insufficient training of data due to unavailability. For effective treatment, we have addressed the solution for low and unclean data.Another limitation of the work in [20] shows an imbalance dataset used for heart signals classification, and for each segment, the authors assumed only one type of disease. Lim et al. [21] proposed a solution for emotion detection using an emotional speech database. Convolutional neural networks, long short term memory, and time distributed convolutional neural networks were among the models utilized. Our method is based on the audio approach and focuses on deep learning classification. Because deep learning requires a vast quantity of data for sufficient training to achieve high accuracy acceptable in the medical area, the method also employs advanced normalization and augmentation techniques for efficient training.

2.1.2 Audio Feature Based Classification

The review in [22] mentioned several feature extraction and classification techniques for obstructive pulmonary diseases such as COPD and asthma. The process involves several traditional and deep learning classification techniques such as K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Deep Neural Network (DNN), and Convolutional Neural Network (CNN) and feature extraction through signals such as Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT), spectrograms, and wavelet transform.

Zhang et al. in [23] proposed a three-step classification for environmental sound classification. The first stage is to show a convolutional filter, which improves in the detection of energy modulation patterns and the attention method by focusing on the relevant channel of the filters. The next phase is to introduce temporal and channel attention. This enhances the overall performance of the Convolutional Neural Network. Finally, they employed data augmentation in the third stage to minimize overfitting due to the lack of data.

Dalal et al. [14] has compared four methods of machine learning approaches for the purpose of lung sound classification using lungs dataset. They have extracted spectrograms, MFCC and LBP features which were given as input to the classifier. Support vector machine, k-nearest neighbor, gaussian mixture model, and convolutional neural network are among the techniques they used. CNN beat all other classifiers in their tests, according to the researchers. This, however, is dependent on the batch size and epoch count.

Rocha et al. [24] developed algorithms to detect the sounds for the clinical and non-clinical tests that checked the environmental and chest placement of stethoscope. They generated a database of lung sounds with 920 recordings divided into several groups (i.e.,COPD, Healthy etc). The challenge's second aim was to extract features and classify the sounds based on their nature (Wheezes, Crackles or both). MFCC, spectral characteristics, energy, entropy, and wavelet coefficients were all utilized. They also looked into the viability of various machine learning algorithms including support vector machines and artificial neural networks. However, because to a lack of data, they were restricted in many ways. We used the same dataset and retrieved spectrogram characteristics in our research. The audio samples contained a lot of noise due to the environmental recordings, so we used normalization and data augmentation techniques to clean them up.

Although previous works were based on convolutional neural networks and data augmentation techniques, where the authors used traditional techniques to classify audio using environmental, lung and heart datasets. The benefit of our DA model is to improve accuracy performance on the model which lacks some training in the audio-based techniques and retrain the classes with low classification accuracy for diverse datasets. We will explain our approach in detail as we further proceed with our upcoming sections.

2.2 Deep Audio Analytics

2.2.1 Pre-processing

For advanced classification, preprocessing is performed in two steps, i.e., Data Normalization [25] and Data Augmentation [11].

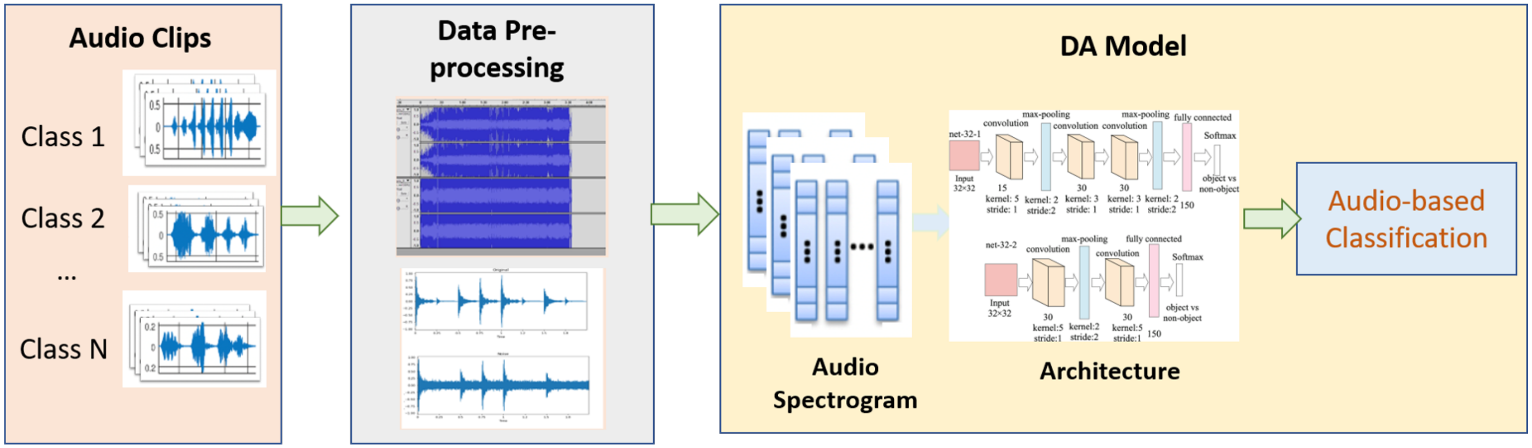


Figure 2: An overview of Proposed Deep Acoustic Model

2.2.1.1 Data Normalization

We have evaluated several different types of normalization techniques. We have shortlisted the three best ones, which outperformed others through the evaluation. These three best normalization techniques are (1) Root Mean Square, (2) European Broadcasting Union Standard R128 Normalization, and (3) Peak Normalization. Each normalization technique has been applied to the dataset before the training.

Root Mean Square Normalization The amplitude level in the Root Mean Square (RMS) Normalization uses the average of a signal amplitude rather than the arithmetic mean of a signal received. When using traditional calculations, such as taking the arithmetic mean of a signal, there can be an issue because the signal amplitude can contain both positive and negative values, which can offset each other and result in a zero amplitude. In this case, taking into account RMS amplitude can be advantageous. The RMS level is useful for determining signal strength based on amplitude, regardless of whether the signal is positive or negative. For a given signal, $x = x_1, x_2, \ldots, x_n$, the RMS value, x_{rms} is:

$$x_{rms} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2)} \tag{2.1}$$

Only by determining the scaling factor that can execute the linear gain change will we be able to normalize the signal amplitude. It is possible to scale a signal that has an amplitude greater than 1 or less than zero 0 decibels (db). We'll rearrange the preceding RMS level calculation as indicated in Equation 2.2, where R has a linear scale, to apply the linear gain adjustment.

$$
\begin{aligned}
R &= \sqrt{\frac{1}{n}[(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2]} \\
R^2 &= \frac{1}{n}[(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2] \\
nR^2 &= [(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2] \\
a^2 &= \frac{nR^2}{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2} \\
a &= \sqrt{\frac{nR^2}{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2}}
\end{aligned} \tag{2.2}
$$

Peak Normalization Peak normalization analyzes the peak signal level in decibels relative to full scale (dBFS) and increases the loudness of the signal to the point where the output is 0 dB maximum. The signal has large loudness peaks as a result of this feature. Some signals stay quiet even after peak normalization, and the quality of those signals cannot be improved further. The above 0 procedure can scale the amplitude of all incoming audio signals to the point where the signal's greatest amplitude equals 1. The output signal based on the above scale can be calculated analytically as

$$
out = \frac{1}{max(abs(in))}.in \tag{2.3}
$$

European Broadcasting Union Standard R128 Normalization

European Broadcasting Union Standard R128 Normalization focused on measuring the average loudness of a program in the normalization of audio signals. This normalization is commonly applied on media channels where we have heavy amount of broadcast material. The R128 performs normalization based on the psycho-acoustic model.

This model performs normalization on the perception of a signal where the human audibility is considered. This model commonly approaches the loudness of unit scale and is considered better than RMS and peak in theory [26].

2.2.1.2 Data Augmentation

For data augmentation it is always important to select the deformation patterns in such a way that their original labels are maintained and the data is augmented. After data augmentation of the original audio samples, the augmented data is given as an input to the neural network model. We have experimented different types of data augmentation and concluded to experiment our results in three different ways such as time stretching, pitch shifting, and dynamic range compression [27]. The three data augmentation techniques are discussed below.

Time Stretching This is a data augmentation technique where the speed of the audio sample is changed and is increased or decreased by some factors [28]. For our experimentation technique, we used four types of audio samples speed, {i.e., 0.5, 0.7, 1.2 and 1.5} along with the original files, which has the speed of 1 and keeping the pitch and other factors the same as original audio sample files.

Pitch Shifting In this technique of data augmentation, the pitch of the audio samples are either decreased or increased by 4 values (semitones) [18]. The duration of the audio samples is kept constant similar to the original audio samples i.e.,4 - 10 seconds. The value changed in semitones ranged between -2, -1, 1, 2.

Dynamic Range Compression This technique compresses the dynamic range of the audio sample by four parameters [29]. Among the four parameters, three are taken from Dolby E Standard and 1 is taken from ice cast radio live streaming server. Those are music standard, film standard, speech, and radio.

2.2.2 Classification Model

Convolutional neural network is gaining importance in the field of deep learning and growing very fast. It is being considered as one of the best network models for recognizing the audio features in the health domain more accurately than any other network. CNN is also considered as the leading model in image, text and other computer vision related fields. We have implemented the 2D CNN model using Keras framework along with tensor board for loss and accuracy graph generations. Keras architecture is supported both by CPU and GPU.

A convolutional neural network has two main components: a feature extractor and a classifier. The feature extractor takes the spectrogram features from the audio signal and sends them to a classifier for classification. The classifier is made up of various convolutional and pooling layers, which are then activated. It also has fully connected layers as well as some units that are hidden.

The 2D CNN architecture is shown in Figure 3 is composed of 5 layers. The convolutional layers are the first three, followed by two fully connected layers, which are enclosed by the max pool layer. We utilized a window size of 23 ms and a hop size of 23 ms to extract spectrogram features, and we kept the extraction to 3 seconds to make every piece of the sound sample acceptable. The input from the sound clips is reshaped, and $X \in R^{128x128}$ shape is provided to the classifier.

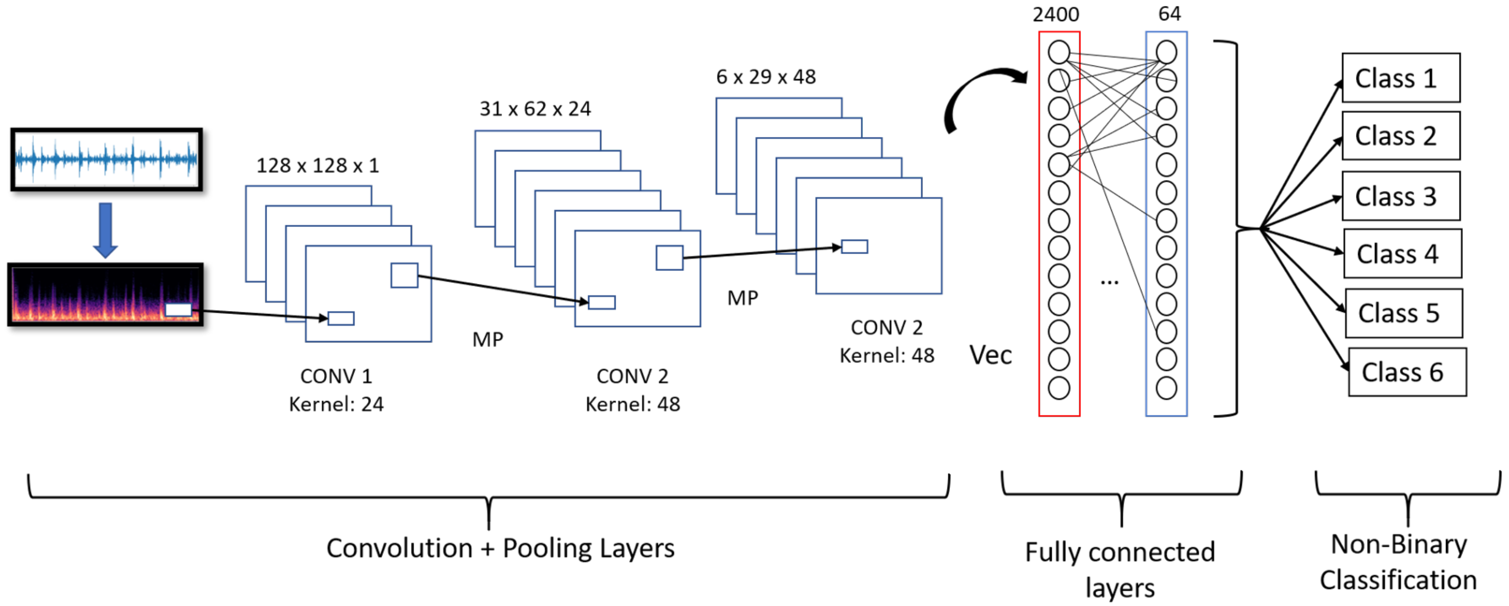
The reshaped features are fed into the first layer in the form of spectrograms with 24 filters. It has the dimensions of [24x1x5x5]. This layer's stride is [4x2], and the activation function is ReLU. The second layer contains 48 filters in the shape [48x24x5x5], as well as a [4x2] stride max-pooling layer and ReLU as an activation layer. The third layer uses 48 filters with a receptive field of [5x5], yielding a shape of [48x48x5x5], with ReLU activation without pooling. Finally, the fourth layer contains 64 hidden units, yielding shapes [2000x64] with ReLU activation and [64x10] with softmax activation. Due to the localized patterns, we proposed a [5x5] tiny receptive layer in the top layer.

2.3 Experimental Works and Results

The Deep Acoustic (DA) model is mainly based on the classification of sounds along with data normalization and data augmentation techniques.

2.3.1 Dataset

Selection of dataset is a real challenge for classification in deep learning. We have chosen four diverse dataset to evaluate our model efficiency. The description for each dataset is in next section.



*MP = Max-Pooling

Figure 3: Classification Model for Deep Acoustics

2.3.1.1 Dataset for Lung Sounds

Working in health domain requires us to closely evaluate the dataset and come up with the accuracy, which is almost near to what is expected to be evaluated by real physicians. We came across datasets, which were not feasible for our experimentation and faced a lot of difficulties. The best dataset that we came across were offered by R.A.L.E [30] a Canadian Lab for medical instructions but it was not available publicly and it has 70 audio samples, which is not very useful for deep learning. Finally, we came across a dataset, which was made public for the challenge hosted in 2017 by International conference on Bio-medical Health Informatics for lungs disease classification [24]. The dataset is composed of total 5.5 hours of recording, which are further divided into recording samples of 126 patients. The recordings are distributed into Asthma, Chronic Obstructive Pulmonary Disease (COPD), Healthy people, Upper Respiratory Tract Infection (URTI), Lower Respiratory Tract Infection (LRTI) and Pneumonia disease. We followed the annotated text files provided by the dataset authors and separated all the data according to the disease category. Hence, the data according to the categories obtained are given in Table 1.

2.3.1.2 Dataset for Heart Sounds

We used heart sound on the publicly available dataset. The dataset [31] was released for a challenge. The dataset consists of different categories, i.e., Murmur, Noisy Murmur, Normal, Noisy Normal, extrasystole, and unlabelled test. Each of the recorded

Table 1: Original and Augmented Data Size for Lung Sounds

ID	Name of Disease	Data Size	Augmented Data Size
1	Asthma	1	13
2	Bronchiectasis	29	377
3	COPD	785	10205
4	Health	35	455
5	LRTI	2	26
6	Pneumonia	37	481
7	URTI	31	403
Total		**920**	**11,960**

sounds is of varying length between 1 to 30 seconds. Some of the audio files have extreme noise in it. The dataset was rearranged into six categories based on types and noises. The distribution of the files is given in Table 2. We have also applied normalization and augmentation techniques to increase and improve the number of files to train our model better. The file distribution for original, normalized, and augmented is shown in Table 3.

Table 2: Dataset File Distribution for Original Heart Sounds

S.No.	Category	No. of Files
1	Extra Systole	46
2	Normal	200
3	Noisy Normal	120
4	Murmur	66
5	Noisy Murmur	29
6	Unlabelled Test	195
	Total	**656**

2.3.1.3 Dataset for UrbanSound8k

We have considered UrbanSound8K [32] dataset, which consists of 8732 labeled audio samples distributed in ten labeled categories such as dog bark, children playing, car

Table 3: Dataset File Distribution and Augmented for Heart Sounds

Dataset	No. of files
Original	656
Peak Normalized	656
RMS Normalized	656
Original Augmented	5904
Peak Normalized Augmented	5904
RMS Normalized Augmented	5904

Table 4: Dataset File Distribution and Augmented for UrbanSounds8k

Dataset	No. of files
Original	8732
Normalized	8732
Augmented	113516

horn, street music, gunshot, air conditioner, siren, engine idling, jackhammer, and drilling. Each audio sample is of 3 second long. The audio recordings are based on environmental sounds which are recorded in an open environment. These audio contains a different type of noise and distortion that are recorded with different environmental effects. The file distribution for original, normalized, and augmented is shown in Table 4.

2.3.2 Experimental Setup

The experimental setup is important since we can observe how the experimentation was performed and what was the expected outcome based on our analysis. First, we took the data in its original form, extracted the spectrogram features and provided the features as an input to our network model, which has reported us accuracy. After getting the accuracy the main aim was to improve the accuracy. For the purpose of accuracy, we have considered normalization to be applied on audio samples rather than the vector normalization. Furthermore, we have analyzed that there could more room for improving the accuracy, which could be realistic enough to be considered especially for health domain. The only problem at this stage was availability of data. Deep learning needs large amount of data to recognize and report a model accurately. For this purpose we used different types of data augmentation techniques. Data augmentation was applied directly on the original dataset and also on the data that was normalized. During the experimentation of the model we kept training of data at 70% and 30% for testing, the batch size was kept at 32 and the number of epochs was fixed at 100 to avoid any over-fitting and under-fitting issue that may report a wrong accuracy.

2.3.3 Preliminary Results

2.3.3.1 Results for Lung Sounds

It was observed during our experimentation stage that the highest accuracy achieved by the existing research is 97% which is dependent on GPU usage and memory consumption. We have used 2D CNN with normalization and augmentation techniques for classifying the models on the available data. Although the data was not enough for proper training and testing, we were able to achieve good results with augmentation techniques. Our model is experimented for 2D CNN classification network on the original dataset, which reported an accuracy of approximately 83%. Further, we have applied the three types of normalization i.e.,Peak, RMS and EBU, and obtained an accuracy of 86%, 87% and 88% respectively. The data augmentation is considered as the trend making technique

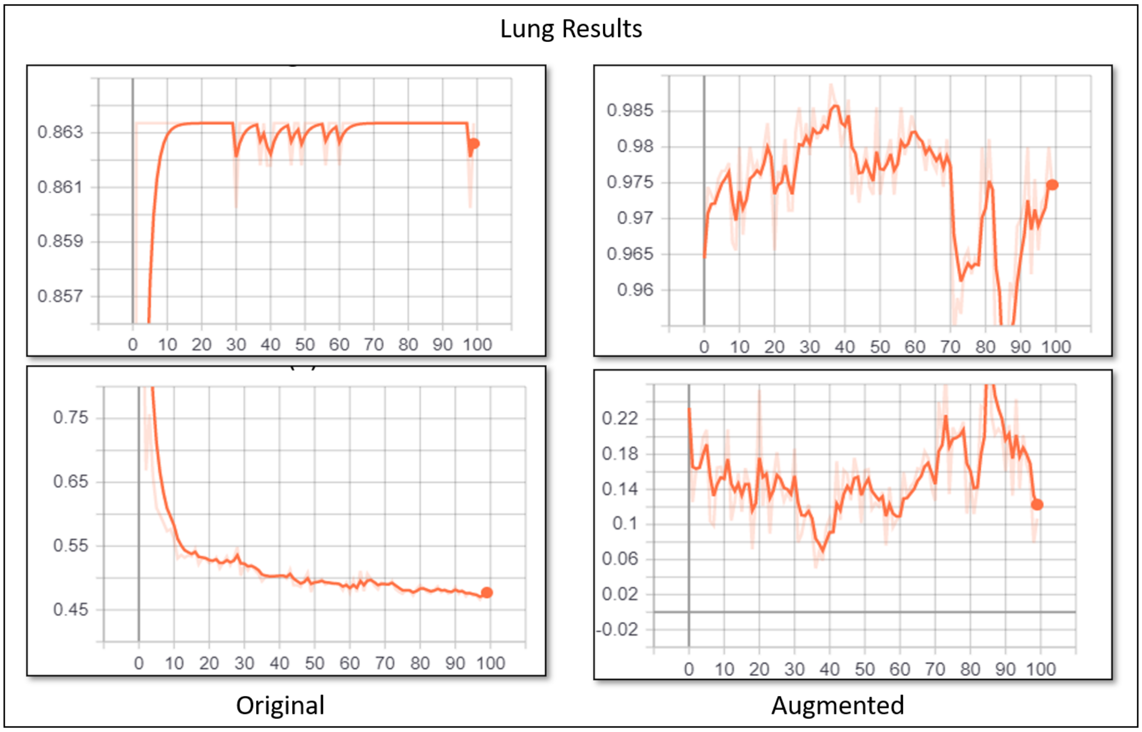


Figure 4: Classification Accuracy for Lung Sounds

in deep learning for small datasets.

Among all other techniques we have considered the three most common types of augmentation techniques, such as time stretch, pitch shifting and dynamic range compression, which outperformed among other techniques with signal and audio data. The original dataset that consists of 920 audio samples were augmented and eventually 11960 audio samples were generated. The accuracy reported from the 2D CNN for the original data augmentation was 93%. We have also applied three augmentation techniques on normalized data and the highest accuracy achieved was 97%.

2.3.3.2 Results for Heart Sounds

We have used a publicly available dataset for heart disease classification. Initially, the data was not in the correct arrangement. We have re-arranged the data into six different folders based on various categories. We adopted a strategy to execute our model on original data, augmented data, and normalized data. Then we applied the normalization technique to remove any type of out liars in the data. It was observed that audio improved better. Finally, we considered applying the augmentation technique to increase the number of files with different filters. In this way, our model has more training data available. Based on augmented the normalized data, it has increased the number of files by almost 8X times. The audio training and graphs can be seen in Figure 5

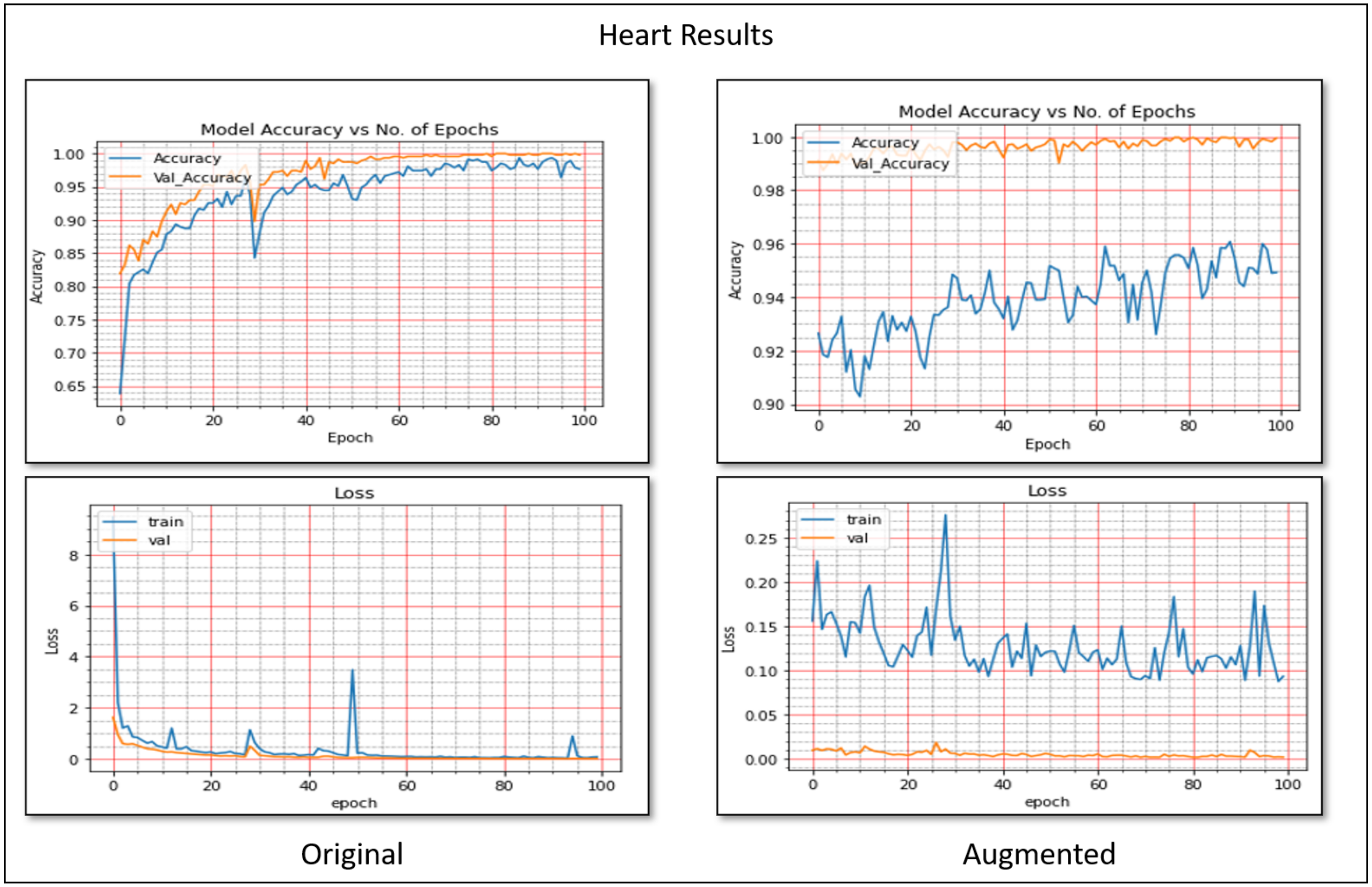


Figure 5: Classification Accuracy for Heart Sounds

2.3.3.3 Results for UrbanSound8k

We started with the classification of urban sound using a publicly available dataset ”UrbanSound8K”. We extracted the spectrogram features and used the 2D CNN model and came up with an accuracy of approximately 74 - 75%. To remove noise and obtain structured data, we applied normalization techniques such as RMS, Peak, and EBU. Among the three methods, we got the best results for EBU, i.e., approximately 88% accuracy. To increase data for more efficient training, we have used augmentation techniques such as Time scaling, Pitch shifting, Dynamic range compression. Finally, after augmentation, our achieved accuracy was 95%. Figure 6 shows the learning performance of original, and normalized augmented data. We can see from the figure that augmentation has increased the accuracy of the system from a remarkable quantity.

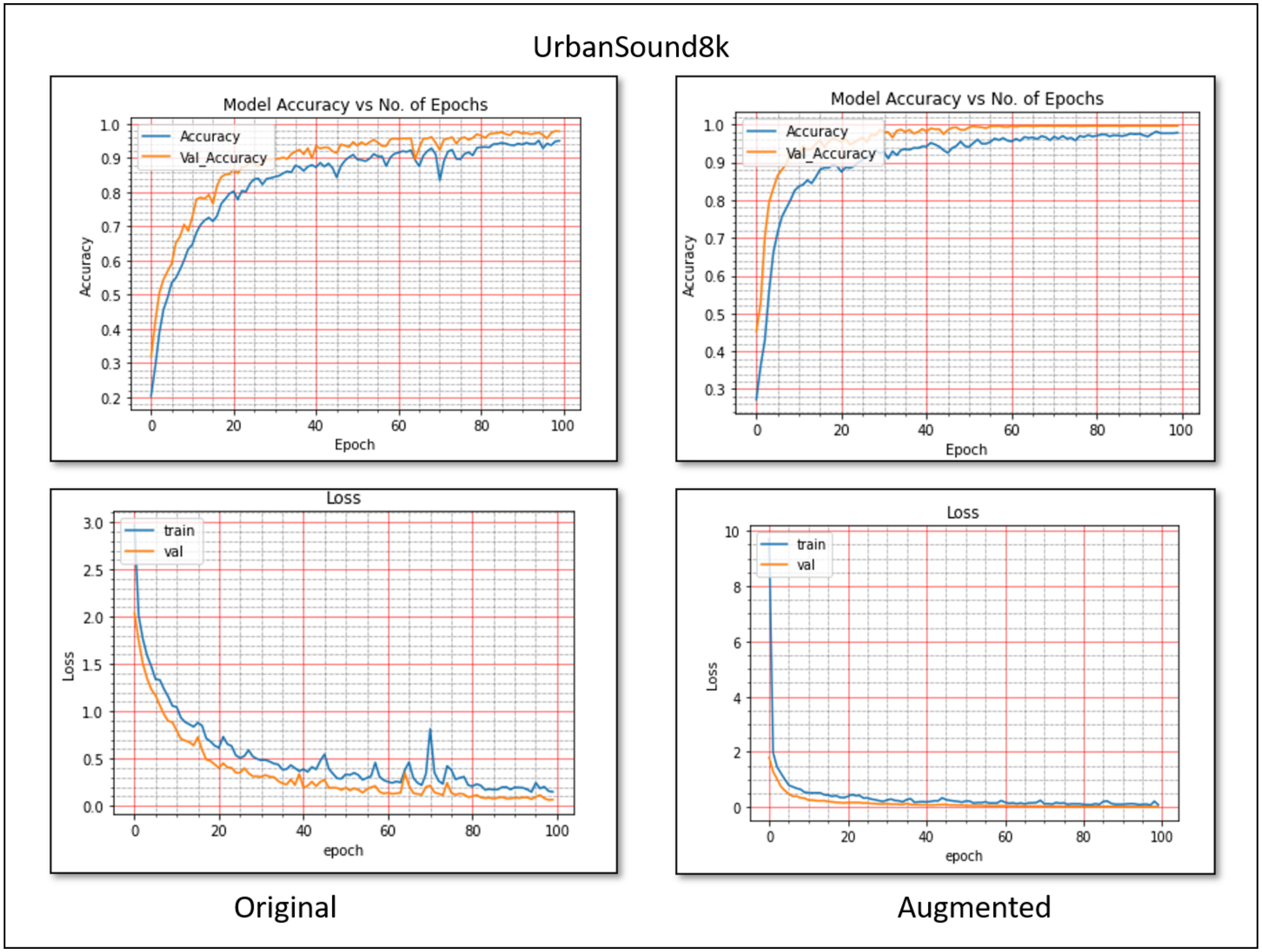


Figure 6: Classification Accuracy for UrbanSound8k

2.3.4 Conclusion

In this chapter, we propose an integrated deep learning model named as Deep Acoustic (DA) model for deep acoustic analytics. The model is based on data normalization, data augmentation and CNN. The proposed deep acoustic analytics techniques have been applied in the fields of environment and healthcare, i.e., Lung and Heart condition detections, urban and environmental sounds.

CHAPTER 3

MULTIMODALITY FOR DEEP ACOUSTICS

3.1 Introduction

Deep learning technologies have received significant attention for large-scale image classification in the area of computer vision [33]. The characteristics of sounds for classification are too complicated to understand the hidden patterns of data. The image-based sound classification was introduced to effectively captures the diverse patterns in the dataset [34]. With the continuous development and innovation of algorithms for image processing in many fields such as medical and environmental, deep learning has become an important research direction [35]. For example, in the medical field, detecting human body sound can be challenging due to the presence of internal or external noise, as well as ear sensitivity in determining whether a sound is normal or abnormal, which can alter diagnosis results.

As a result, a created automated algorithm for identifying diseases through sounds provides an efficient way for clinical diagnosis, reducing subjectivity on human body sound. Similarly, there is an increasing amount of audio data available to environmental audio classification, we still face significant challenges to conduct accurately deep learning in the environmental audio domain. These challenges may occur in particular due to the various field and categories. Specifically, distortion, fracture, and noise of audio data are the primary obstacles affecting the accuracy of environmental audio classification.

Advanced technologies are essential to achieving the improvement of lifestyle and health care.

As critical data is required for a useful deep learning model, publicly available data is insufficient and uncleaned in specific domains, mainly for audio classification where the length and size of continuously recorded sounds are challenging. The normalization techniques play a vital role in normalizing the sound's peak values and removing the wanted noise from the signals without affecting the type of diseases. To address the insufficient data, a technique known in deep learning is *Data Augmentation* [36]. This technique is useful for increasing the quantity of training data. It is considered a better method to tackle common data problem issues by adding some extra features to data files.

The main contribution of this chapter is the proposal of a multimodal classification approach for audio and image classification based on a spectrogram feature. The spectrogram is a textural representation of time, phase, and frequency of sounds in the form of an image. The multi-modal classification is useful for better learning and identification of different sound classes through images. This basically reflect the patterns that are not visible using sound. The techniques of image processing are applied to normalized spectrogram feature extracted from different types sound.

The objectives and contributions of this chapter can be summarized as follows:

- We propose an integrated Multimodal Deep Acoustic (MDA) model with advanced data pre-processing techniques for high performance.
- The pre-processing is conducted using advanced data normalization and data augmentation techniques.

- The three benchmark datasets i.e., lung, heart and environmental sounds are normalized by removing the unwanted noise and adjusting the signal's peak values.

- we have used advanced augmentation techniques to generate the data without affecting the category of datasets.

- After preprocessing, we developed the spectrograms from sound datasets and used these spectrograms features and images for audio and image classification respectively.

- The proposed method increases the visualization of different types of individual datasets and passes the images to our designed network model.

- Our preliminary results shows that our model outperforms state-of-the-art results in terms of learning time and accuracy. MDA model improved performance compared to the existing models.

3.2 Related Work

3.2.1 Deep Learning for Audio-Visual Classification

Chen et al. [37] proposed a novel solution for lung sounds classification by using a publicly available dataset. The dataset was divided into three categories, i.e., wheezes, crackles and normal. They proposed a detection method using optimized S-transformed (OST) and deep residual networks (ResNets). They performed preprocessing on the audio samples by using OST, which rescaled the features for ResNets.

Rupesh et al. [22] have reviewed several features extraction and classification techniques for pulmonary obstructive diseases such as COPD and asthma. In their review, the feature extraction used were FFT, STFT, spectrograms and wavelet transform.

Bozkurt et al. [38] focused on segmentation and time-frequency components for the CNN-based designs. The Mel-spectrogram and MFCC features were extracted from heart sound data using the PhysioNet dataset [39]. They also performed the data augmentation by changing the sampling rate with a random value in range.

Jannat et al. [40] proposed a solution for emotion detection using audio and video features separately and by fusing them. They used the RAVDESS dataset to train their model for audio emotion detection and BP4D+ multimodal emotion corpus for training the video part of their research. They used the inception V3 convolutional neural network in their research.

3.2.2 Visual Feature based Classification

Demir et al. [41] proposed a solution for environmental sound classification using spectrogram extraction. They considered deep features for classification problems and trained their network model, an end-to-end system trained on spectrogram images rather than real audios. Feature vectors are combined using a concatenation of the fully connected layers. Finally, for testing the model's efficiency, they have provided a feature set as an input to K nearest neighbor algorithm using ensemble voting classifier to report an accuracy.

Bian et al. [42] proposed the multiple networks of the Densely connected convolutional network (DenseNet) and Residual Neural Network (ResNet). They extracted spectrograms and converted them to grayscale images using the cv2 library in python. They further clipped the audio signal into small sub-signals and applied ensemble voting to better accuracy for biased cases where the precision was dropped during the extraction phase. They evaluated their work using Support Vector Machine and Convolutional Neural Network. Their experimentation received better results for DenseNet for the music audio tagging problem using FMA-small and GTZAN datasets.

Toffa et al. [43] presented an audio classification model using a lightweight convolutional neural network-based method based on a texture feature local binary pattern (LBP) with audio features, e.g., MFCC, GFCC. Their work with convolutional neural network-based methods using LBP was better than classical machine learning algorithms, e.g., SVM, random forest, and KNN, using audio features. This work is very similar to our work in terms of the lightweight CNN-based model and the evaluation conducted with both visual and audio feature-based classifiers.

The benefit of MDA model is to improve accuracy performance on the models, which lacks some training in the audio-based techniques and retrain the classes with low classification accuracy. We will explain our approach in detail as we further proceed with our upcoming sections.

3.3 Multimodal Acoustic Method

The diagram depicts the overall picture of our proposed method in Figure 7. The suggested method uses a dynamic structure to extract features and parameters, minimize redundant and inexpressive data, address the issue of limited data necessary for deep learning, and identify classes using signal and image processing. There are three stages to our model: 1) Sound pre-processing with data normalization and augmentation, 2) spectrogram feature extraction for multi-modality 3) The Model of Classification. Below are the specifics for each level.

3.3.1 Data Normalization

We already selected the top three normalization strategies for sound datasets, namely lung and heart, following many evaluations. The methods were tested on the training dataset, and they outperformed the other normalization methods. They are classified as 1) Root Mean Square 2) Peak 3) European broadcast Union Standard R128 (EBU). The summary for each normalization is given below:

3.3.1.1 Root Mean Square Normalization

The RMS level is useful for identifying the signal strength based on the amplitude regardless of the positive or negative values of the signal, where it does not work as the arithmetic mean of a signal received. The signal amplitude normalization can only be possible if we can figure out the scaling factor that can perform the linear gain change. There is a possibility to scale a signal with an amplitude that is higher than 1 or less than zero 0 decibels (dB).

For applying the linear gain change to a given signal, $x = x_1, x_2, \ldots, x_n$, the RMS value, x_{rms} is shown in Equation 3.1, where R has a linear scale.

$$\begin{aligned} R &= \sqrt{\frac{1}{n}[(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2]} \\ a &= \sqrt{\frac{nR^2}{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2}} \end{aligned} \tag{3.1}$$

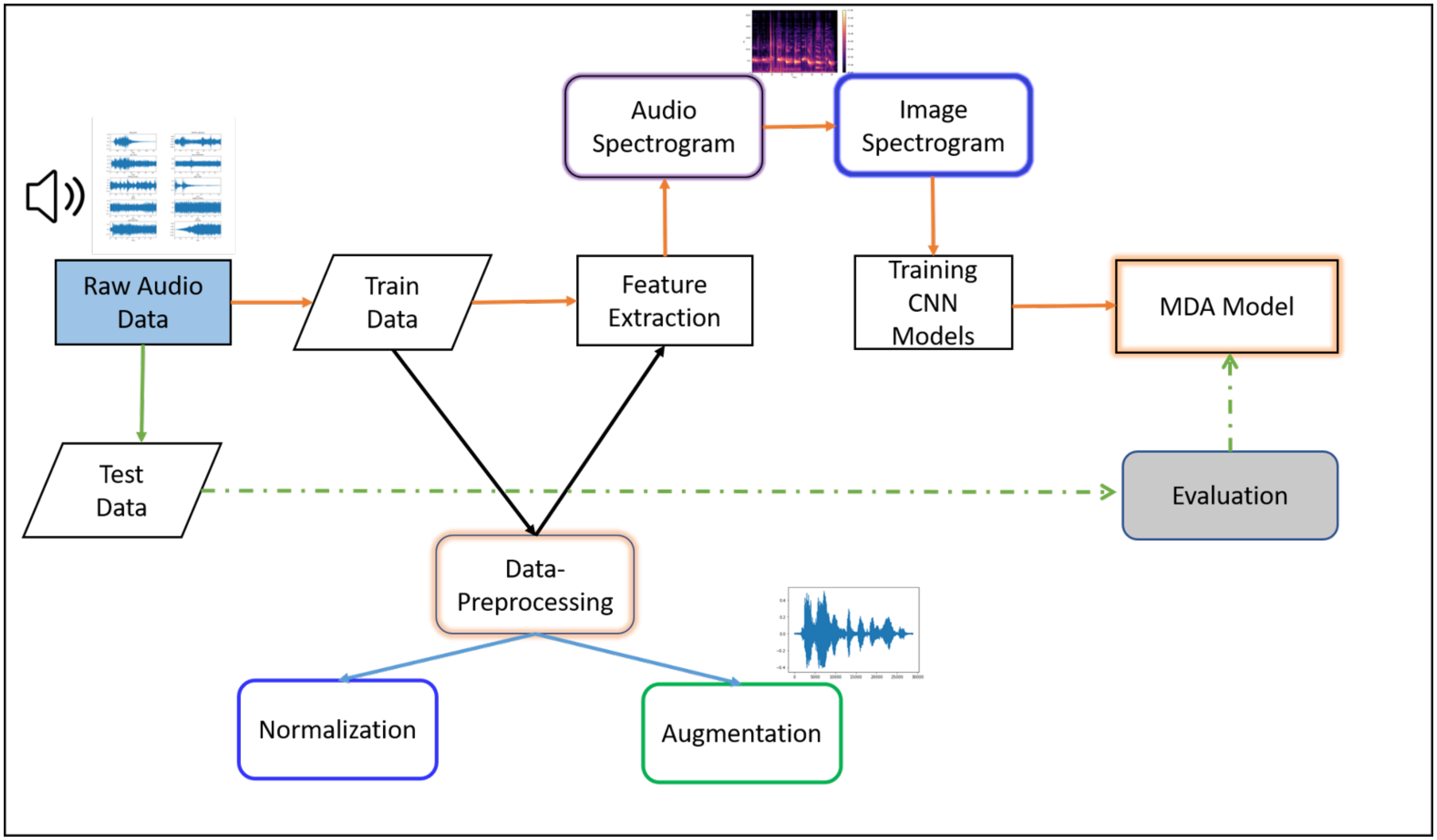


Figure 7: Workflow for Proposed Multi-modal Classification Model

3.3.1.2 Peak Normalization

In peak normalization, the signal has high volume peaks as it amplifies the volume of the signal in such a manner that the output gets 0 dB maximum. Even after peak normalization, some of the signals remain quiet, and the quality of those signals cannot be improved further. The above 0 processes can scale the amplitude of all input audio signals in such a way that the highest magnitude of the signals has a value of 1.

3.3.1.3 European Broadcasting Union Standard R128 Normalization

This technique focuses on measuring the average loudness of a program for the normalization of audio signals. This model performs normalization on the perception of a signal where the human audibility is considered. It commonly approaches the loudness of the unit scale and is considered better than RMS and Peak in theory.

3.3.2 Data Augmentation

We used data augmentation to develop lung sound data and tested with three of the best techniques for lungs sound data, including time stretching, pitch shifting, and Dynamic Range Compression.

3.3.2.1 Pitch Shifting

The pitch of the audio samples is either decreased or increased by four values in this data augmentation technique (semitones) [28]. We assume that with the pitch shifting factor a_{shift}, the artificial training data generated is $Naug$ times larger than the original lung sound data. The audio samples' duration is kept constant similar to the original audio samples. For our experimentation, the value changed in semitones were in the interval $[-a_{Shif}, a_{shif}]$ for each signal. The value has changed in semitones ranged of (-2, -1, 1, 2).

3.3.2.2 Time Stretching

Like the pitch shifting, the lungs' data signals are stretched horizontally along the time axis by a scaling factor $a_{stre} > 0$. We sampled the values in the interval $[1, a_{stre}]$ if $a_{stre}\geq$ 1 or $[a_{stre}, 1]$ along with original files while keeping the pitch and other factors same. The four audio speed is (0.5, 0.7, 1.2, 1.5) along with original files.

3.3.2.3 Dynamic Range Compression

The audio sample is compressed using this method based on its dynamic range. This can be accomplished by either boosting the sample or lowering the level of loud sounds. Three of the four parameters come from Dolby E Standard, while the fourth comes from the ice cast radio live streaming server.

3.3.2.4 Spectrogram Generation

The spectrogram is a visual representation of a signal in the time-frequency domain. These are generated by the application of the short-time Fourier transform(STFT) [44]. According to the theorem, a single Fourier analysis may not see a nonstationary signal's spectrum variation. Hence, the spectrogram considers the stationary signal by computing the Fourier transform of the segmented signal into slices. Hence the spectrogram is also called STFT, which can be calculated as:

$$STFT_x^f(t,f) = \int_{\infty}^{\infty} [x(t)w(t-\tau)e^{-j2\pi ft}dt \tag{3.2}$$

where x(t) is time-domain signal, τ is the time localization of STFT and $w(t-\tau)$ is a window function to cut and filter the signal. The length of the window function must be selected and adjusted according to the signal's length because it affects the time and frequency resolution [45]. We have transformed the spectrogram into a grayscale image, where we used the image processing methods to extract the information.

Scaling Process The scaling process is applied to the spectrogram to expand the values range between 0-255 because the range of the spectrogram is usually wide. The method of scaling is done in a linear manner, which can be expressed as follow:

$$S(m,n) = \frac{|Spec(m,n)|}{max|Spec|} \times 255 \tag{3.3}$$

where Spec(m, n) is the value of the spectrogram and S(m, n) is the expanded value from a spectrogram.

3.3.3 Classification Model

Convolutional neural networks have become a major trend in deep learning, with applications in music, image recognition, computer vision, and a variety of other domains. Convolution, pooling, and fully connected layers comprise CNN. We investigated the power of CNN for lung disease classification in this paper. With Keras' implementation, we created a 2D convolutional neural network.

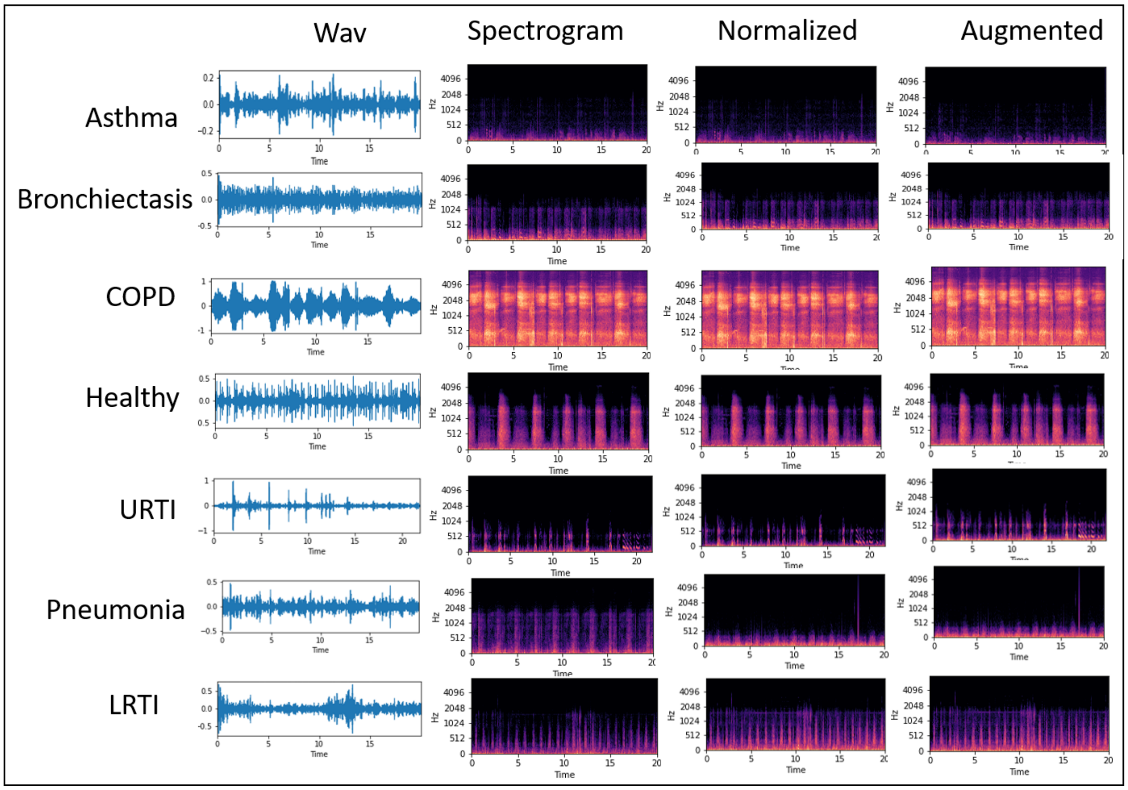


Figure 8: STFT Obtained from Original Wav files a) Spectrogram Obtained from Original Lung Sound Data, b) Spectrogram Obtained from Normalized Lung sound data, c) Spectrogram Generated from Augmented Lung Sound Data

Our 2D CNN architecture is composed of 5 layers. The first three are the convolutional layers enclosed by the max pool layer, followed by two fully connected layers. We extracted Librosa features for Mel spectrograms because noise data spectrograms are considered the best to differentiate between the type of sounds. During the extraction of features, we have used window size and hop size of 23 ms. As the sound clips vary for different categories, that is why we kept the extraction to 3 seconds for each datasets to make every bit of the sound clip usable. We have reshaped the input taken from the sound clips to $X \in R^{128x128}$ shape. Further, we have sent these reshaped features to the classifier for classification of sounds and images.

3.4 Experimental Works and Results

For experimentation and evaluation, we are using lung and heart sound on the publicly available dataset. The description for each dataset is given below:

3.4.1 Dataset for Lung Sounds

The lack of publicly available data for training is a key disadvantage of lung disease classification. We used the dataset, which is the only publicly available dataset known as the Respiratory dataset, for classification. [24]. This research used acoustic recordings from 126 patients in Portugal and Greece. Healthy, Asthma, Chronic Obstructive Pulmonary Disease (COPD), Bronchiectasis, URTI, LRTI, and Pneumonia are among the diseases represented in the recordings. In audio format, the dataset contains the recording index, patient number, position on the chest, and instrument used to diagnose diseases. In addition, spectrograms were created and sorted in the same order as the original categories to create an image dataset. For the same task and sounds, images were created to qualify the differences between an audio and visual dataset. The audio files for the lungs included both clean and noisy audios.

3.4.2 Dataset for Heart Sounds

For experimentation and evaluation, we are using heart sound on the publicly available dataset. The dataset [31] was released for a challenge. The dataset consists of different categories, i.e., Murmur, Noisy Murmur, Normal, Noisy Normal, extrasystole, and unlabelled test. Each of the recorded sounds is of varying length between 1 to 30 seconds. Some of the audio files have extreme noise in it. The dataset was rearranged into six categories based on types and noises. We have also applied normalization and augmentation techniques to increase and improve the number of files to train our model better. The distribution of files is kept same as shown in Chapter 2.

3.4.3 Experimental Setup

We used advanced techniques to develop our model for data classification. We used an NVIDIA GPU graphic card with four 11GB GPU slots and a 1080i resolution. The RAM capacity is 16 GB. To assess the efficiency of our approach, we used publicly available sound datasets for the lungs and heart. The experiment was carried out using two separate approaches: audio feature-based technique and audio to visual spectrogram-based approach. We used the same model to classify the outcomes from two different cases stated previously. We calculated the classification accuracy by comparing the findings from both situations. We also compared our training and testing accuracy scores.

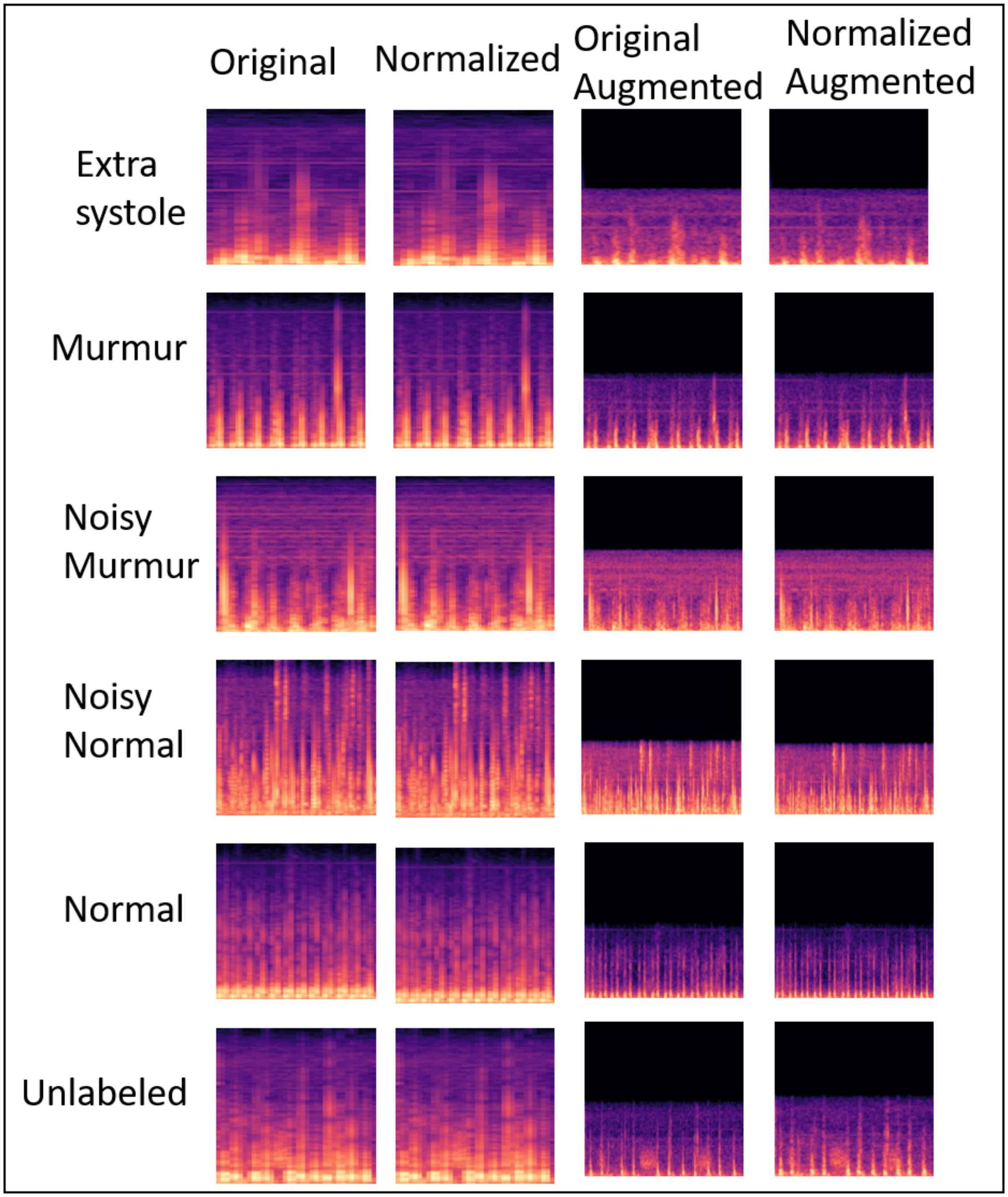


Figure 9: Spectrogram Visualization of Heart Sounds

We used Early-Stopping in our code to avoid unfitting and overfitting and to get accurate results for the network we developed.

3.4.4 Feature Extraction

We have used the librosa library [29] to extract the spectrograms from the audio files. STFT is used to divide a continuous signal into segments for spectrograms. Each audio file was in the wav format, and the lengths of the files varied. Using the librosa input duration capability, we reduced the audio files to 3 seconds for our model. For audio classification, we extracted the features from the spectrogram and placed them in a NumPy array. In contrast, we used the CV2 package to convert the visuals into grayscale images for spectrogram image classification [46] for better prediction of results. The images are then directly entered into a CNN model in shape (128,128,1) after being resized to the correct shape for our CNN model.

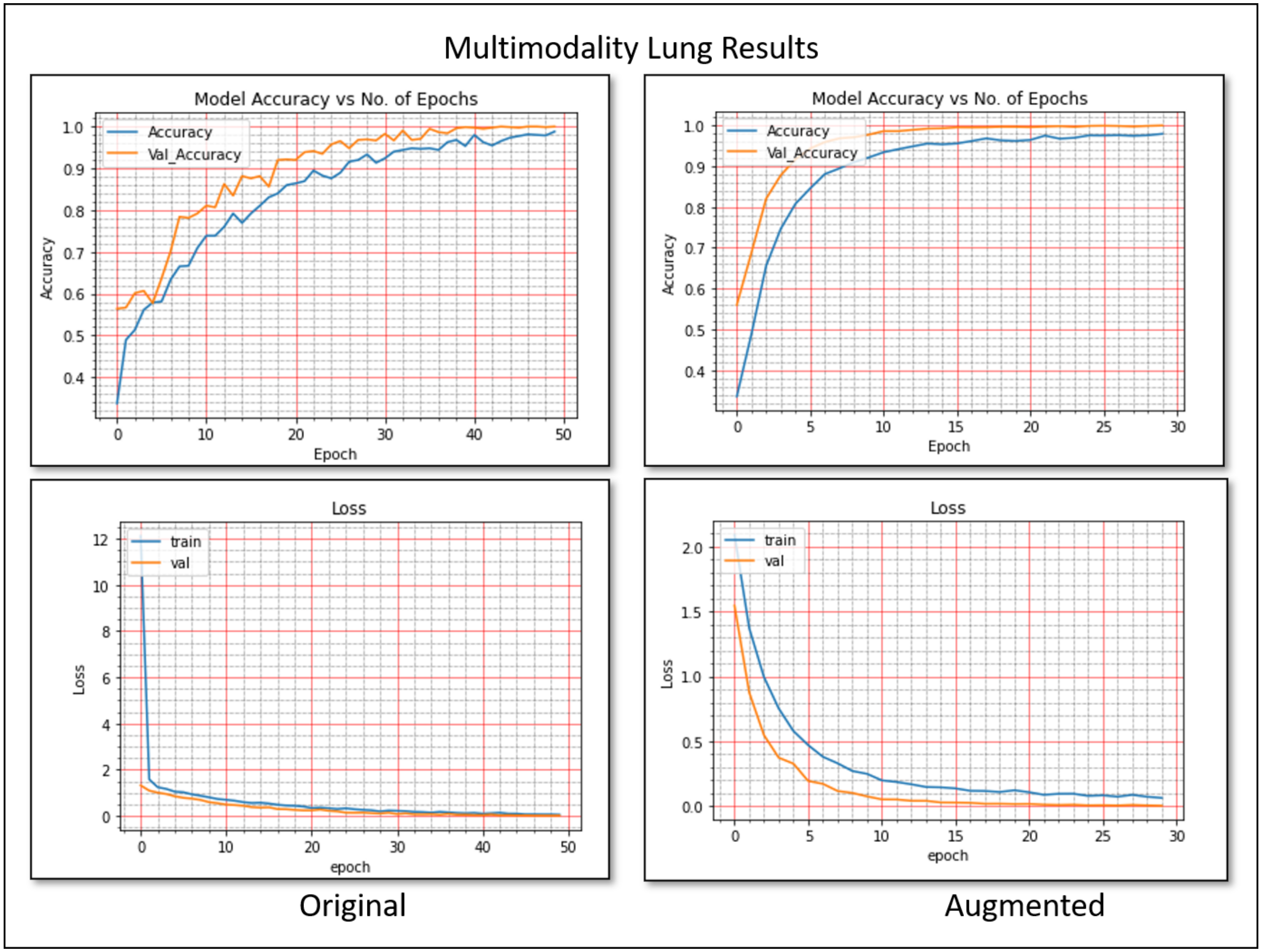


Figure 10: Model Training Accuracy and Loss for Image-based Lung Sounds

3.4.4.1 Results for Lung Sounds

We continued our research in this publication by testing the multi-modality of our network model in a different domain, namely image-based classification. It was determined during the audio extraction process that spectrograms may be stored to a NumPy array or directly as spectrogram images. This time, we chose images to be used in the classification of lung sounds. The overall workflow of our image classification system is shown in Figure 7. Images are a more visual form of a signal that can be easily interpreted by clinicians, whereas audio classification cannot. We did this by extracting visual features from audio files that were 72x72 pixels in size. This ratio was retained to ensure that the images were of the highest quality and that the images used for classification were of the highest quality. To assess the accuracy utilizing multi-modality, the images were downsized to 128x128x1 to fit in our network model. The accuracy testing results we obtained were highly reliable. For the original form of the dataset in images format, the maximum testing accuracy given was around 84%. We wanted to improve the performance of our network model. We used data normalization, and the findings show that due to the nature of the images, the accuracy reduced by a small amount. We transformed the images to grayscale images during the spectrogram generation stage, and when normalizing was applied, the image quality declined even further, resulting in slightly reduced accuracy. Finally, we did augmentation after the audio classification procedure. The augmentation improves accuracy by a right margin that was dropped when normalization was used since the image quality was decreased.

Augmentation, on the other hand, allowed the network model to train the images in numerous ways. Image classification achieves a final and highest accuracy of 95%. The training of the model was done incredibly successfully, as shown in Figure 10, however the accuracy reduced by a tiny difference due to the loss of image quality. Due to the model overfitting issue, the accuracy was also reduced in the early phases.

3.4.4.2 Results for Heart Sounds

Our results are divided into two categories, i.e., Audio classification and spectrogram image classification. After extraction of the files' corresponding features, it is fed into the convolutional neural network for classification. We have used a publicly available dataset for heart disease classification. Initially, the data was not in the correct arrangement. We have re-arranged the data into six different folders based on various categories. We adopted a strategy to execute our model on original data, augmented data, and normalized data. It should also be noted that we have used the above techniques for feature-based techniques and image-based techniques. The audio feature-based approach seems to perform the image-based approach due to the number of images and the image's quality. Then we applied the normalization technique to remove any type of out liars in the data. It was observed that audio improved even better as compared to the image. Finally, we considered applying the augmentation technique to increase the number of files with different filters. The image performed drastically well due to clean data, much-improved filters, and a high training amount.

One important key that we can see from the training and loss figure is the model overfitting. The image-based approach shows the network's performance is better than audio and takes much less time than audio. We considered batch size to be 64 and the number of epochs to be 100. While using Early Stopping, the image-based approach performs better. The image's better accuracy was reported as 96%, while 93% was reported for audio.

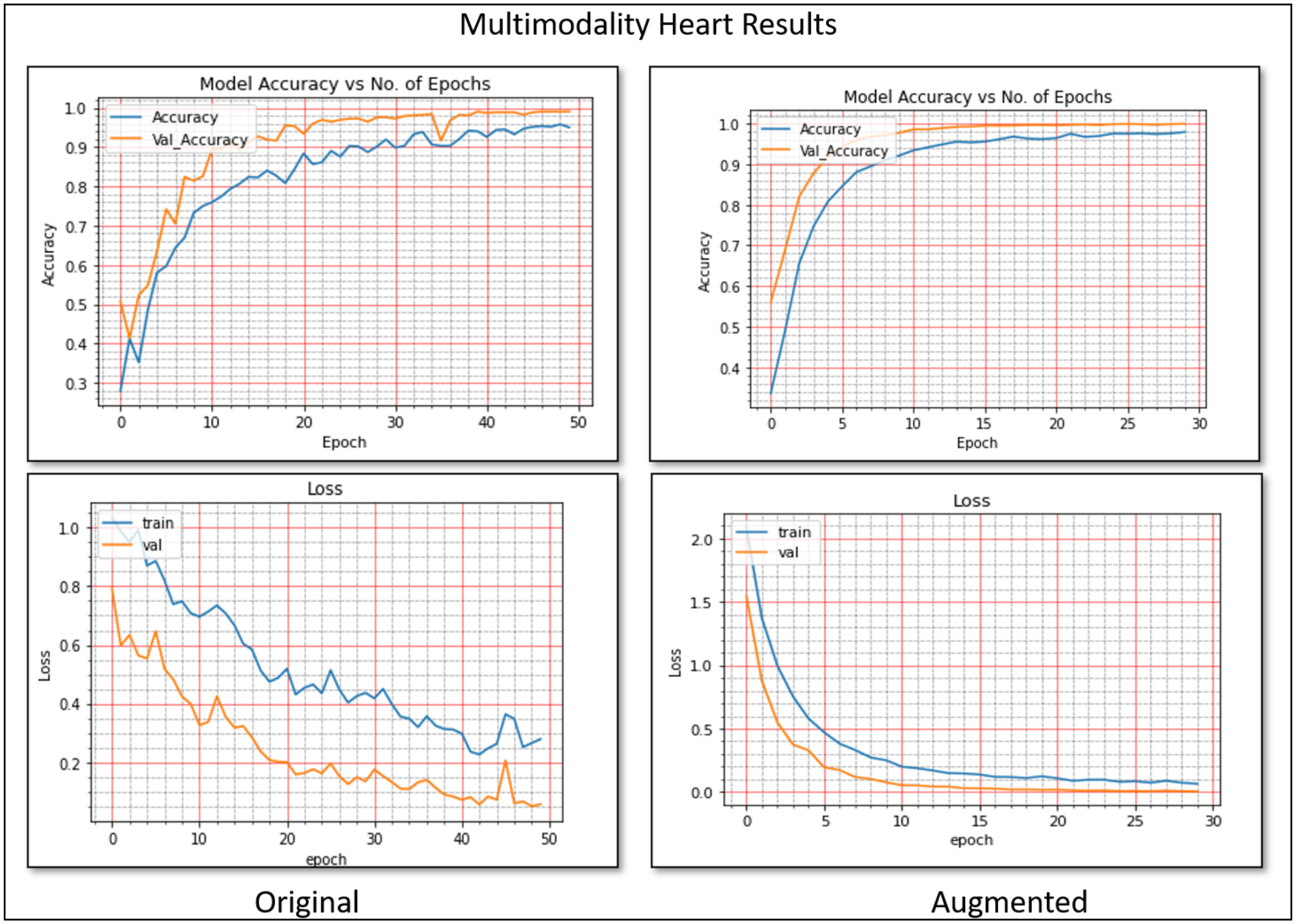


Figure 11: Model Training Accuracy and Loss for Image-based Heart Sounds

3.4.5 Comparison of audio and image techniques

Overall, we can state that, when comparing audio and image, audio can report acceptable accuracy while using more memory and reporting time. The image can complete the same task in a relatively short time and with 80% less memory than audio, and it can access the correctness quickly. The image may be visualized and is thought to be simple to utilize by health practitioners.

3.4.6 Conclusion

We introduced the Multimodal Deep Acoustic Classification (MDA) model for high-performance classification in this chapter, which is coupled with advanced data normalization and data augmentation approaches. We experimented with our multi-modality model, which proved to be quite effective.

CHAPTER 4

DEEP NETWORK-BASED FUSION FOR DEEP ACOUSTICS LEARNING

4.1 Introduction

There is increasing attention for audio classification research that aims to support various emerging applications, including environmental monitoring, health care, and smart city The environmental sound classification (ESC) is an important area of research that needs more dynamic and be adapted to possible changes in the environment, facilitate the adoption of enhancement techniques with innovative and effective solutions.

While there is an increasing amount of audio data available to environmental audio classification, we still face significant challenges to conduct accurately deep learning in the environmental audio domain. These challenges may occur in particular due to the various field and categories, e.g., UrbanSound8k [17], ESC-10 and ESC-50 [16], animal sound [47], noises, the massive volume of data [48]. Specifically, distortion, fracture, and noise of audio data are the primary obstacles affecting the accuracy of environmental audio classification. There have been many advancements in the correct detection of audio sounds; the audio data depending on how the features are extracted from the audio, how modality to represent the audio data, and how much noise is present in raw audio [49].

With the rapid increase of environmental datasets, extracting relevant data to create an adequate classification of environmental sounds is a crucial challenge [50]. Recently, substantial advances in environmental sound classification have been achieved.

Audio classification modeling based on deep neural network architectures has gained remarkable success with massive environmental sound data [51, 52]. Some alternative approaches have been introduced, focusing on extracting acoustic features and generating corresponding audio-visual datasets in audio classification [53, 54, 55].

Specifically, deep learning plays an essential role in audio classification over traditional machine learning nowadays. Convolutional neural network is growing most in audio classification [56, 9] using Environmental audio data [17]. Sounds are to be classified by distinguishing between different types of distinct patterns across time-frequency domain features like spectrograms [57]. In audio classification domains, noise in sound data may affect the performance of classifiers. Normalization techniques with feature scaling [58] was introduced to eliminate noises in the environmental sounds. Also, data augmentation is typically used to produce more audio samples with various extended features.

The data for audio classification is too complicated to understand multiple characteristics and latent patterns of data. Ensemble learning [59] or data fusion techniques are introduced for effective classification. Recent works are suggesting that ensemble or fusion techniques are needed to improve the performance of audio classification. Piczak et al. [9] observed different performance of CNN models, e.g., poor performance in the context of various short-scale temporal structures (drilling, jackhammer, engine idling) while excellent for one containing specific classes (playing children, air conditioner, car horn). They suggested that ensemble averaging different approaches, such as convolutional and non-convolutional, may effectively capture the diverse patterns present in the dataset.

As new technologies like image-based audio classification were introduced, more diverse fusion approaches, e.g., feature-level fusion [60, 61, 23, 43], layer-based fusion [62], modality-based fusion [63], methodology-based fusion [41], and network-level fusion [64, 59], were introduced. Among various fusion techniques for audio classification, the most popular are those based on the features. However, not much work has been conducted on the general fusion approach, such as network-level fusion, which can be applicable to more diverse datasets or applications.

In this chapter, we propose the Fusion-based Learning for Deep Acoustic Classification model for environmental audio classification. Our contributions can be summarized as follows:

- Design an effective fusion model is designed based on the model fusion weights learned from two models for various audio datasets, optimized by the architectural model fusion techniques.

- Our methodology and architecture is generic enough to be used in a dual-modal presentation of audio features and spectrogram images for environmental audio classification.

- To apply the advance data augmentation techniques to combine time stretching and pitch shifting to generate more diverse training data. Besides, a normalization approach of Root Mean Square (RMS) and peak value detection was proposed for normalizing noise signals of vast environmental sound datasets.

- The experiments with the fusion model show the effectiveness and robustness in the environmental sound classification compared to state-of-the-art environmental audio classification on UrbanSound8k [17], ESC-10, and ESC-50 [16]. Our aim is to show that the model is significantly smaller than existing ones, especially from the state-of-the-art environmental audio classification.

4.2 Related Work

There are many existing techniques available that demonstrate the usability of audio classification using diverse datasets. The section below has all work that has been done for the environmental sound classification and fusion techniques. The researchers have achieved some good results based on high computation power. In contrast, we focused on designing 2D convolutional neural network models with low computational requirements to arrive at a fusion model design by achieving high accuracy from datasets' diverse nature. The literature survey for environmental sound classification is shown in Table 5

4.2.1 Deep Learning for Audio Classification

The analysis of environmental sound in [16], which is close to our work, has been performed where the authors proposed Convolutional Neural Network (CNN) to classify environmental sounds. The architecture consists of two convolutional rectified layer unit by applying max pooling, two fully connected hidden layers, and a softmax output layer. The datasets used for classification were environmental science [16] and the UrbanSound8k dataset [17]. The data was augmented through random time delays and pitch shifting. Using librosa implementation, Mel Spectrograms were extracted from all audio files, re-sampled, and normalized with different window sizes. The classifier, data set, and features used were similar to our experiments; however, we have created our network classification model with fusion techniques for environmental sound classification. We have also compared our experiments with some other similar work.

4.2.2 Audio Feature Based Classification

Abdoli et al. [51] proposed a solution for the classification of UrbanSound8K where they have proposed an end to end solution for the 1D convolutional neural network. The proposed solution is feasible for audio of any length due to splitting and frames overlapping. They used parameter-based comparison with available state-of-the-art researches; the highest mean accuracy reported by other researchers was 78%. Their proposed solution achieved approximately 89% and incredibly accuracy, performing better in low utilization of computation power.

Zohaib et al. [66] propose a solution for the environmental sounds classification using a deep convolutional neural network. They have used three feature extraction techniques, i.e., Mel Frequency Cepstral Coefficient, Mel spectrogram, and Log-Mel. Furthermore, they used a Convolutional neural network of 5 layers in two different formats, such as max-pooling (Model1) and without max-pooling (Model 2). They have also used data augmentation to avoid overfitting. The highest accuracy they got for ESC-10, ESC-50 & UrbanSound8K is from Log-Mel features of approximately 94%, 89%, and 95% using Model 2 along with data augmentation.

Zhang et al. in [23] proposed a three-step classification for environmental sound classification. The first step is to present a convolutional filter, which helps detect energy modulation patterns and the attention mechanism, focusing on the filters' relevant channel. The step involves introducing temporal attention and channel attention. This improves the performance of the Convolutional Neural Network as a whole. Finally, in the third step, to avoid overfitting due to the low amount of data, they have used data augmentation. Keep the above in view; they have applied these techniques to ESC-10 and ESC-50, and DCASE16 datasets. The highest accuracy reported for ESC-10 and ESC-50 was approximately 94% and 86% and 88%, respectively.

4.2.3 Image Based Audio Classification

Demir et al. [41] proposed a solution for environmental sound classification using spectrogram extraction. They considered deep features for classification problems and trained their network model, an end-to-end system trained on spectrogram images rather than real audios. Feature vectors are combined using a concatenation of the fully connected layers. Finally, for testing the model's efficiency, they have provided a feature set as an input to K nearest neighbor algorithm using ensemble voting classifier to report an accuracy. Their fusion technique is based on CNN and classical machine learning, while our fusion is based on two different CNN networks. The absolute accuracy was reported for two datasets, i.e., DCASE-2017, Urbansound8K, with the respective accuracy of approximately 96% and 86%.

Mushtaq et al. [65] proposed a solution for data augmentation, which is applied directly to the audio clips rather than applying to images. They have presented their models built with the 7 and 9 layer CNN architecture. They have further experimented with their models for transfer learning, freezing, and unfreezing layers. In their work, the highest accuracy was reported by ResNet-152 for ESC-10 and UrbandSound8K as 99.04% and 99.49%, respectively. For ESC-10, DenseNet-161 performed better and reported an accuracy of 97.57%.

4.2.4 Ensemble & Fusion for Audio Classification

Palanisamy et al. [64] proposed a solution for sound classification using both weight-based and ensemble models. They also built an image model with spectrograms using transfer learning from a model pre-trained with ImageNet. There was a significant difference between spectrogram and ImageNet, but the transfer learning approach allowed them to obtain reasonable classification accuracies for image-based audio classification. During the experimentation, it was observed that weight-based model performance was better than randomly selected model performance. The best accuracy for the model-based approach was 91%. While using an ensemble approach, the highest accuracy achieved was approximately 93%. The dataset selected for experimentation is GTZAN, UrbanSound8K, and ESC-50.

Choi et al. [68] proposed Convolutional Recurrent Neural network (CRNN) for music classification. The model-based fusion was proposed as the last convolutional layers were replaced by Recurrent Neural Network (RNN), and both the classifiers are used for feature extraction and summarization, respectively. Computational controlled experiments were performed by changing the parameters of the networks.

A multi-label Recurrent Neural Network was proposed by Parascandolo et al. [69] in the shape of bi-directional long short term memory (BLSTM) recurrent neural network for polyphonic sound event detection in real-life recordings. The model was tested for real-life recordings, including 61 classes from different contexts such as beach, basketball game, inside a bus, hallway, car, shop, restaurant, office, street, and stadium with events. The spectral features, such as Mel spectrograms, were extracted and presented in a sequence of frames. The dataset was augmented through time-stretching, subframe time-shifting, and blocks mixing techniques and mapped to the model. The accuracy of the augmented data was improved overall for the bi-directional model, which is 64.7%.

Alsouda et al. [70] have implemented an Internet of things based solution for the detection of different noises. They have used Raspberry Pi Zero for the recording and classification. They have used a dataset of 3000 sound samples for seven categories: car horn, jackhammer, etc. Furthermore, they have used Support Vector Machine, Random Forest, K-Nearest Neighbor, and Bagging Aggregation for the classification. They achieved accuracy in the range of 88% to 94%, where each algorithm's parameters were kept standard.

Although previous works were based on convolutional neural networks and ensemble learning, where the authors used traditional techniques to classify UrbanSound8K, ESC-50, and ESC-10 audio datasets, some authors used ensemble learning for combining two or more different models for better classification using a voting system of ensemble learning. We will demonstrate a fusion-based model applied to a diverse range of techniques, such as an audio-based approach or an image-based approach. Our proposed model, FDA-NET, is mostly related to audio and image-based classification, and finally, we use fusion-based learning on the audio datasets. The benefit of FDA-NET is to improve accuracy performance on the models, which lacks some training in the audio-based techniques and retrain the classes with low classification accuracy. We will explain our approach in detail as we further proceed with our upcoming sections.

Table 5: Literature Survey for Environmental Sound Classification

Publication	Network	Feature	Approach	Limitation
Mushtaq et al. [65]	ResNet-152 DenseNet-161	Spectrogram	A visual approach with 7- and 9-layer CNN and transfer learning	Large number of trainable parameters and high consumption of computational power
Toffa et al. [43]	CNN	Local Binary pattern	An audio classification model using light weight CNN model	Low performance in terms of accuracy
Demir et al. [41]	CNN+SVM	Short time Fourier transformation	A fusion based approach using CNN and classical machine learning model trained on spectrogram images	Low performance in terms of accuracy
Mushtaq et al. [66]	CNN	Log-Mel Mel Spectrogram MFCC	A visual approach and Construction of CNN model and augmentation techniques	High number of augmentation and high consumption of computational power
Palanisamy et al. [64]	DenseNet	Spectrogram	An ensemble approach for the classification of environmental sounds using images. The data was augmented	Low accuracy with high number of trainable parameters
Boddapati et al. [67]	GoogLeNet	Spectrogram	A convolutional deep neural network approach to classify environmental sounds s images	High number of trainable parameters and low performance
Luz et al. [61]	CNN	Handcraft	An ensemble approach using CNN to classify environmental sounds	Low accuracy and handcrafted feature extraction require more computational power
Zhang et al. [23]	DCNN	Spectrogram Gammatone Spectrogram	A fusion based approach using deep convolutional neural network on augmented audio data	Low performance in terms of accuracy
Li et al. [60]	DS-CNN	Log-mel	An ensemble DS-CNN model by combining log-mel feature and raw waveform data using Dempster-Shafer (DS) evidence theory	Their fusion approach is a feature-level ensemble that cannot be applied to a diverse range of techniques such as audio-based or image-based approaches
Piczak et al. [9]	CNN	Librosa	A CNN based approach for different performance for different classes	Low performance in terms of accuracy
Piczak et al. [16]	kNN + RF		A baseline approach for environmental sound classification	Low performance in terms of accuracy

Table 6: State-of-the-Art Accuracy in Urban Sound Classification (Image & Audio Approaches)

Publication	**Arch.**	**Network**	**Par#**	**Mode**	**Aug.**	**Feature**	**US-8K**	**ESC-10**	**ESC-50**
FDA-NET (Ours)	Fusion	DCNN	**3.9M**	Visual	✓	SP	98.53%	97.5%	96.1%
Mushtaq et al. [65]		ResNet-152	60.19M	Visual	✓	SP	**99.49%**	**99.04%**	97.30%
Mushtaq et al. [65]		DenseNet-161	12.30M	Visual	✓	SP	99.46%	98.87%	**97.57%**
Toffa et al. [43]		CNN		Visual		LBP	-	88.5%	64.6%
Mushtaq et al. [66]		DCNN	3.9M	Visual	✓	LMS	95.3%	89.2%	94.9%
Demir et al. [41]	Fusion	CNN+SVM		Visual		STFT	78.14%	94.8%	81.4%
Palanisamy et al. [64]	Ensemble	DenseNet	20M	Visual	✓	SP	87.42%	-	92.89%
Boddapati et al. [67]		GoogLeNet	6.4M	Visual		SP	93%	91%	73%
FDA-NET (Ours)		DCNN	**3.9M**	Audio		SP	**96.96%**	**93.47%**	**93.1%**
Luz et al. [61]	Ensemble	CNN		Audio		HC	96.16%	86.2%	-
Zhang et al. [23]	Fusion	DCNN		Audio	✓	SP	-	94.2%	86.5%
Zhang et al. [71]	Fusion	DCNN		Audio	✓	GT	83.7%	83.9%	91.7%
Li et al. [60]	Ensemble			Audio			91%	90.2%	81.1%
Piczak et al. [9]		CNN		Audio			73.7%	64.9%	80.5%
Piczak et al. [16]		k-NN + RF		Audio				72.7%	44.3%

Arch.: Architecture; Par#: Number of Trainable Parameters Aug.: Augmentation; US-8K: UrbanSound8K; LMS: Log-Mel spectrogram; STFT: Sort Time Fourier Transform; SP: Spectrogram; GT: Gammatone Spectrogram; HC: Handcraft

4.3 Methodology

The conceptual frameworks of the fusion model, the FDA-NET model, aim to take advantage of the strengths of the individual networks Figure 12.

We introduced three essential components to improve acoustic classification: (1) Preprocessing with data normalization and augmentation, (2) multimodal feature extraction and representation, and (3) fusion of convolutional neural networks. First, the data normalization and augmentation handle data issues, such as noise, imbalance, feature selection. Second, the multimodal feature extraction and representation focuses on extracting the spectrogram features from the audio signal and transforming them into a visual representation. Third, the fusion techniques with convolutional neural networks were proposed to build multimodal classifiers to learn and infer acoustics classification either in acoustic or visual mode.

4.3.1 Audio and Image Feature Selection

Audio feature and feature images do not always carry the same information. Hence, they can not be predicted in the same way. If one is talking in a visual context, it is known that sound is transparent while the image is considered as non-transparent (opaque) [72]. With the pixel formation of a snapshot, it can always be noted that whatever is the position of an object, it belongs to the same category. However, audio cannot quickly identify where it belongs due to its observed frequency in spectrogram features. The audio spectrogram depends on the magnitude of the frequency on how the object or the combination of objects has produced in a complex manner. Therefore, in audio, it's challenging to identify the simultaneous sounds that are represented in the spectrogram features.

When a neural network uses images for classification, it considers the sharing weights that are available from the images in two dimensions, i.e., X and Y-axis. The image will carry the same information if it is stretched or repositioned regardless of the position and presence. However, in audio, the two-dimensional data represents the frequency and time. Suppose the audio is moved horizontally or vertically; the meaning of the audio can changes. That being said, if the pitch of the sound increases, the voice of a male can switch to a child, or a gunshot can change to an explosion. Even the spatial features of the sound can change if we increase or decrease the pitch. Therefore, the intentionally introduced invariance can change the meaning, and the neural network will not perform well as it should perform on the trained and augmented data.

In the image format, the extracted pixels can assume that the image belongs to the same class, while in audio, this is not true. Audio is composed of periodic sounds that are composed of frequency and harmonics. The harmonics are spaced apart from each other according to their nature. The combination of these harmonics usually determines the sound's timbre. Suppose we assume from the physical point of view. In that case, the audio that is played to the audience will examine the nature of the sound only, while images typically contain a lot of parallel static information. The image classification is done based on the brightness, among other features.

4.3.2 Data Normalization

Our study evaluated several normalization techniques and identified the two best ones: Root Mean Square and Peak Normalization.

4.3.2.1 Root Mean Square Normalization

The Root Mean Square (RMS) Normalization converts digitalized signals to the average amplitude of the signal by computing the square root of the mean squared amplitudes. Suppose we use traditional calculations by taking the arithmetic mean of a signal. In that case, there can be an issue as the signal amplitude can have some positive values and negative values, which can be offset by each other and result in a zero amplitude. In such a situation, considering RMS amplitude can be beneficial. The RMS level determines the signal strength using the amplitude regardless of its positive or negative values.

For a given signal, $x = x_1, x_2, \ldots, x_n$, the RMS value x_{rms} is to be computed. The signal amplitude normalization can be used only when the scaling factor is applicable to perform the linear gain change. It is possible to scale a signal with an amplitude higher than one or less than zero decibels (dB). As the linear gain change is applied, the RMS level formula is defined as shown in Equation 4.1, where R has a linear scale.

$$\begin{aligned}
R &= \sqrt{\frac{1}{n}[(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2]} \\
R^2 &= \frac{1}{n}[(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2] \\
nR^2 &= [(ax_1)^2 + (ax_2)^2 + \ldots + (ax_n)^2] \\
a^2 &= \frac{nR^2}{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2} \\
a &= \sqrt{\frac{nR^2}{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2}}
\end{aligned} \tag{4.1}$$

4.3.2.2 Peak Normalization

The peak normalization analyzes the peak signal level in decibels relative to full scale (dBFS). For the normalization, the signal's volume is amplified to get 0 dB maximum as the output. Thus, the signal has high volume peaks. Some signals remained the same without sufficient improvement of signal quality even after the peak normalization. The above 0 processes can scale the amplitude of input signals to get 1 as the highest amplitude. The output signal based on the above scaling can be mathematically formulated as follows:

$$out = \frac{1}{max(abs(in))}.in \tag{4.2}$$

4.3.3 Data Augmentation

Deep learning relies on a large amount of data for more accurate classification. For better variety, we have experimented with several augmentation techniques, out of which we have selected the best two methods, i.e., Time Stretching and Pitch Shifting. We have experimented with these techniques using the three datasets.

Time Streching In this technique, the audio sample's speed is changed and is increased or decreased a scaling factor $a_{stre} > 0$ [73]. For our experimentation technique, we used four types of audio samples speed, i.e., {0.5, 0.7, 1.2, and 1.5} along with the original clips, which has the speed of 1 while keeping the pitch and other factors the same as original audio sample clips.

Pitch Shifting In this data augmentation technique, the audio samples' pitch is either decreased or increased by four values (semitones) [28]. We assume that with the pitch shifting factor a_{shift}, the artificial training data generated is $Naug$ times larger than the original sounds. The audio samples' duration is kept constant similar to the original audio samples, i.e., 3 - 5 seconds. For our experimentation, the value changed in semitones were in the interval $[-a_{Shif}, a_{shif}]$ for each signal. The value changed in semitones ranged between {-2, -1, 1, 2}.

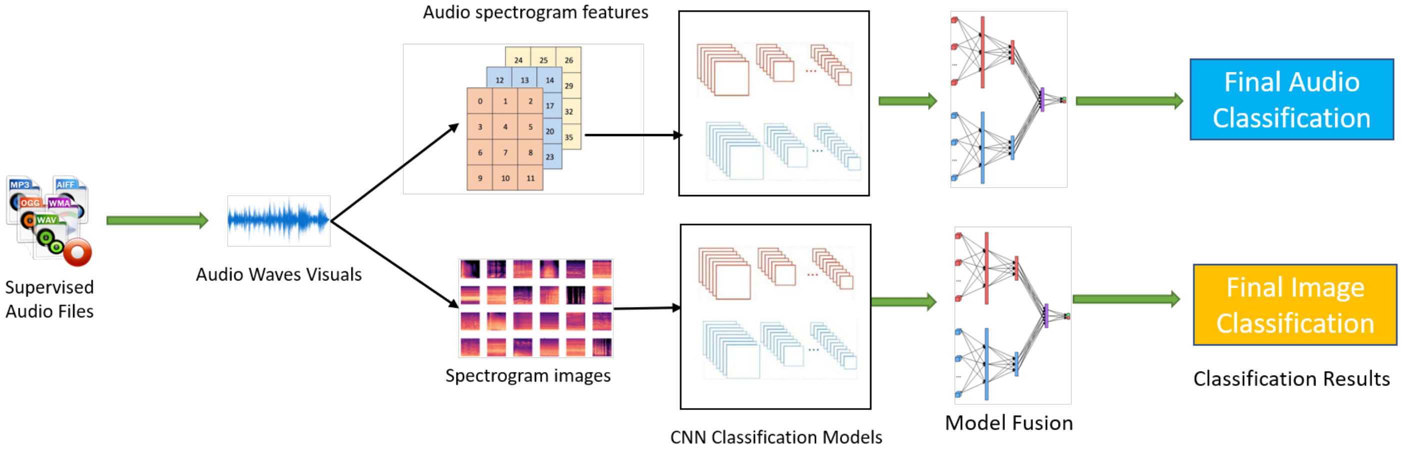


Figure 12: Overall Architecture for our Proposed FDA-NET Model

4.3.4 Spectrogram Generation

The spectrogram is a visual representation of a signal in the time-frequency domain. These are generated by applying the short-time Fourier transform (STFT) [44]. According to the theorem, a single Fourier analysis may not see a non-stationary signal's spectrum variation. To overcome this issue, the spectrogram considers the stationary signal by computing the Fourier transform of the segmented signal into slices. Hence the spectrogram is also called STFT, which can be calculated as:

$$STFT_x^f(t,f) = \int_{\infty}^{\infty} [x(t)w(t-\tau)e^{-j2\pi ft}dt \tag{4.3}$$

where x(t) is time-domain signal, τ is the time localization of STFT and $w(t-\tau)$ is a window function to cut and filter the signal. The length of the window function must be selected and adjusted according to the signal's length because it affects the time and frequency resolution [45]. We have transformed the spectrogram into a grayscale image, where we used the image processing methods to extract the information shown in Figure 13, Figure 14 and Figure 15. The spectrogram feature extraction is shown in a colored format for ESC-10, ESC-50, and UrbanSound8K.

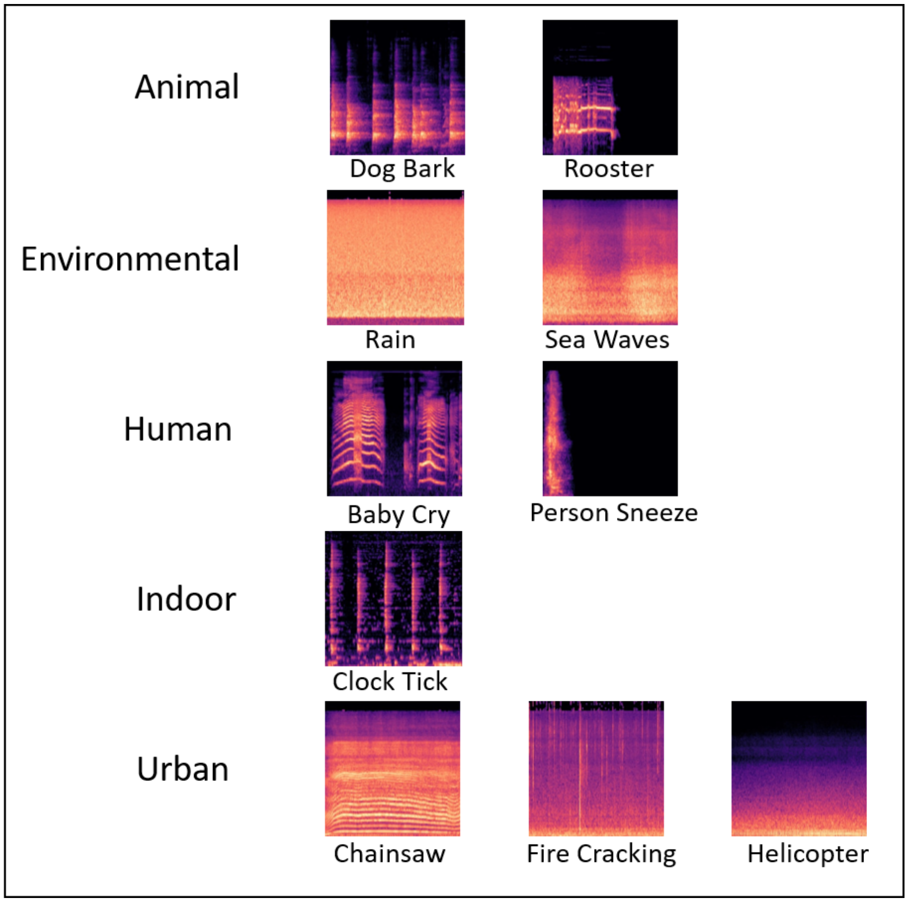


Figure 13: Spectrogram for ESC-10 five categories

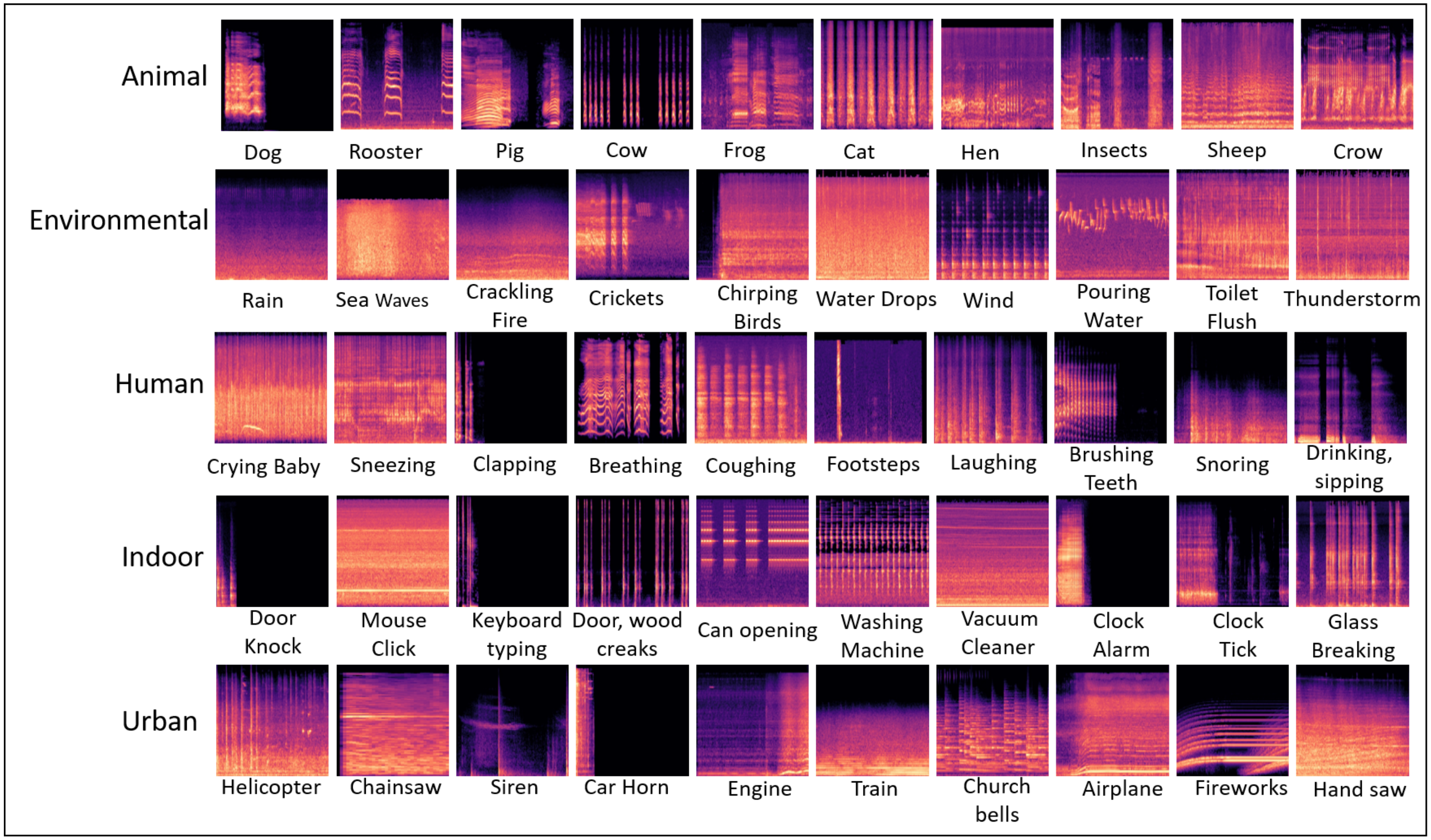


Figure 14: Spectrogram for ESC-50 five categories

Scaling Process The scaling process is applied to the spectrogram to expand the values range between 0-255 because the spectrogram range is usually wide. The process of scaling is done in a linear manner, which can be expressed as follow:

$$S(m,n) = \frac{|Spec(m,n)|}{max|Spec|} \times 255 \tag{4.4}$$

where Spec(m, n) is the spectrogram's value, and S(m, n) is the expanded value from a spectrogram.

4.3.5 Feature Extraction

- **Audio Feature Extraction**: We have used librosa features [29] to extract the spectrograms from the audio files arranged in different folders. Short-Time Fourier Transform (STFT) is used to cut down the continuous signal into parts. Spectrograms operations such as inverse STFT and instantaneous frequency spectrogram (IFS) are used for down streaming analysis of features [74]. For the spectrogram features extracted from the UrbanSound8K dataset, each audio clip in ESC-50 was in an OGG format and wav for UrbanSound8K, where each audio file's length was ¡=5 seconds. We have cut down the audio files to 3 seconds for our model using the librosa functionality of input duration. The spectrogram extracted features [75] from audio files are stored in a NumPy array for input into a CNN model in shape (128 x 128).

- **Image Feature Extraction**: For extracting features [76] from the image, we have generated spectrogram images from audio files directly in a shape of (128 x 128).

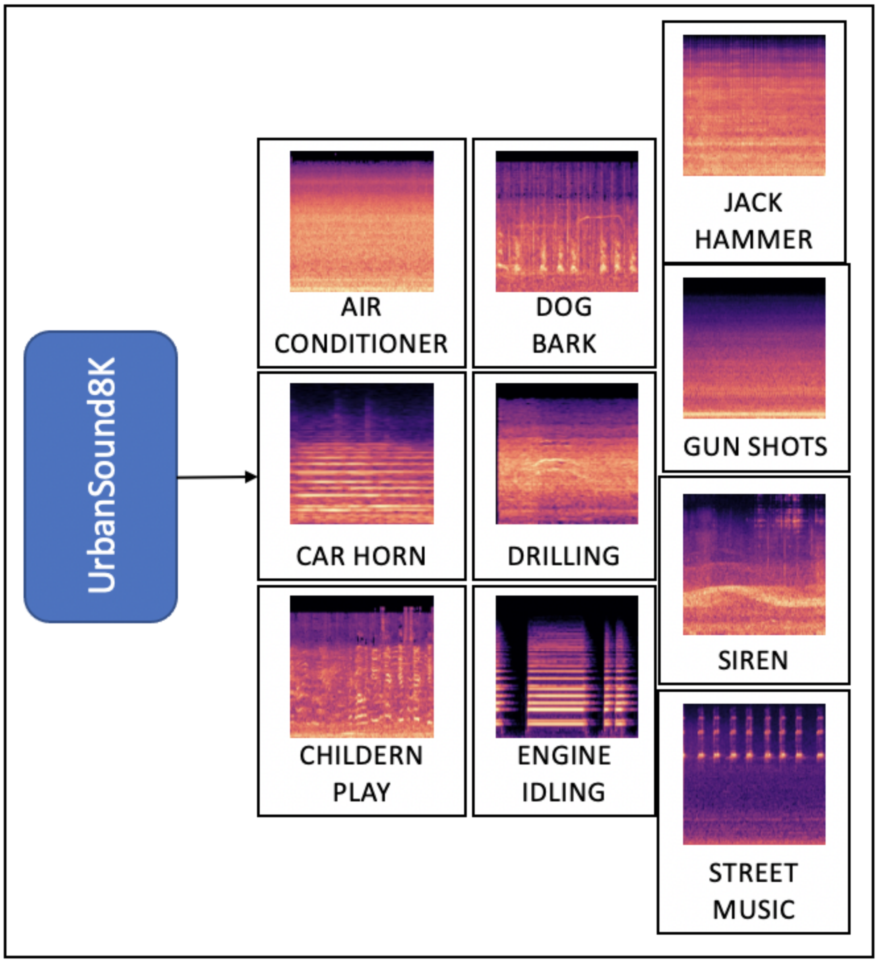


Figure 15: Spectrogram for UrbanSound8k

After the spectrograms are generated, we convert all spectrogram into a grayscale image using the CV2 library to predict results. We save the image into a NumPy array and the class label using the append feature for each file. This way, we can input our image into a CNN model in a shape of (128 x 128 x 1) after resizing it to the desired shape that is required by our customized 2D convolution neural network models, i.e., FDA-NET-1, FDA-NET-2, and FDA-NET-3.

4.3.6 Audio Classification Models

We have explored the power of CNN for environmental sound classification. We have designed a 2D convolutional neural network for both audio and image-based approaches. The innovation is by applying model fusion, i.e., 1) FDA-NET-1 architecture has three convolutional and two hidden layers, which focuses on having a kernel regularizer to avoid overfitting issues. 2) FDA-NET-2 is made of two convolutional layers, along with two hidden layers. 3) FDA-NET-3 is based on the fusion of both models, which were trained separately. The fusion was conducted on their final layer to append both model's parameters. As shown in the result section, the converged model achieved higher accuracy than individual models.

More specifically, in the design of multi-modal models with CNNs, the two essential components include (1) the feature extractor retrieves the spectrogram features from the audio signals and represent them as acoustic feature vectors (for the audio-based approach) and transform the spectrogram features into images and represent them as visual feature vectors (for the image-based approach), (2) The classifiers (FDA-NET-1, 2, 3), which consists of multiple convolutional and pooling layers, activation and fully connected layers with several hidden units, classify either the acoustic feature vectors or the visual feature vectors into their appropriate categories.

There are two primary components of a convolutional neural network, i.e., feature extractor and classifier. The feature extractor extracts the spectrogram features from the audio signal and passes them to a classifier to classify the spectrogram features into their appropriate categories. The classifier consists of different convolutional and pooling layers, followed by activation. It also holds fully connected layers with some hidden units.

The mathematical form of the convolutional layers is given in Equation 4.5 and 4.6

$$[x^{l}_{i,j,k} = \sum_{a}\sum_{b}\sum_{c} w^{(l-1,f)}_{i,j,k} y^{(l-1)}_{i+a,j+b,k+c} + bias^{f}] \tag{4.5}$$

$$[y^{l}_{i,j,k} = \sigma(x^{(l)}_{i,j,k})] \tag{4.6}$$

The output layer is represented by $y^{l}_{i,j,k}$ where as the 3-dimensional input tensor is denoted by i, j, k. The weights for filters are denoted by $w^{(l)}_{i,j,k}$ and $\sigma(x^{(l)}_{i,j,k})$ describes the sigmoid function for linear activation. The fully connected is layer that is represented by equation 4.7 and 4.8.

$$[x^{(l)}_{i} \sum_{j} w^{l-1}_{i,j} y^{l-1}_{j} + bias^{l-1}_{j}] \tag{4.7}$$

$$[y^{l}_{i,j,k} = \sigma(x^{l}_{i,j,k})] \tag{4.8}$$

The 2D CNN architecture is composed of different layers. The initial layers are the convolutional layers enclosed by the max pool layer, followed by fully connected layers. During the extraction of spectrogram features, we have used window size and hop size of 23 ms as the sound clips vary between 3 to 5 seconds, so that we kept the extraction to 3 seconds to make every bit of the sound clip usable. The input from the sound clips is reshaped, and $X \in R^{128x128}$ shape is provided to the classifier.

- **FDA-NET-1**: There are a total of 5 layers of convolutional, hidden, and fully connected layers. The first layer in the architecture takes the reshaped features as an input in spectrograms with 24 filters. It takes the shape of [24x1x5x5]. The stride in this layer is [4x2] with ReLU as the activation function. The second layer has 48 filters of the shape [48x24x5x5] with [4x2] stride max-pooling layer and using ReLU as the activation layer. The third layer also takes 48 filters with a receptive field [5x5], resulting in shape [48x48x5x5], and the activation is ReLU without pooling. Finally, the fourth layer has 64 hidden units resulting in shape [2000x64] with ReLU activation and [64x10] with softmax activation. We considered a [5x5] small receptive layer in the top layer due to the localized patterns. The convolutional layers in this model go through a kernel regularizer to avoid overfitting.

  This model, FDA-NET-1, performs better than the second model, FDA-NET-2, due to Max pooling's larger size, which is known for downsampling the features and selecting the next layer's essential elements. Secondly, better performance is due to striding, which is a concept that considers data compression, i.e., the data block considered for (2 x 2) in the first layers and (4 x 2) in the next layers is taken as an input. It moves 1 unit ahead and provides the output volume. Hence, reducing the number of parameters from 4 and 8 units in the top two layers to 1 and 1 unit, respectively. The drastic decrease of parameters in FDA-NET-1 is due to larger Max Pooling and Strides.

- **FDA-NET-2**:This model has a total of 4 layers. The layers are composed of two convolutional layers and two dense layers. The model takes spectrogram input in the shape of $X \in R^{128x128}$. The number of filters that the first takes are 32. So, the shape of the first layers is [32x3x3], followed by ”ReLU” activation and Maxpooling2D in the size of [2x2]. Similarly, the second layer takes 64 filters in the form of [64x3x3], followed by ReLU activation and pool size [2x2]. After the convolutional layer, we flattened the output vector for the best resolution and added a dropout of factor 0.5 for independent learning. The padding was kept as ”valid” for a well-structured output.

- **FDA-NET-3**: This model is a fusion of FDA-NET-1 and FDA-NET-2 that is concatenated before the final dense layer, depending on the number of classes in the training setâthe last layer of this model uses ”Softmax” activation due to the multiple classification problem. The network was trained with more extensive training parameters with FDA-NET-1 and FDA-NET-2, and their loss and accuracy were computed as the training matrix. As per the working of this network model, features were selected from FDA-NET-1 and FDA-NET-2, and the network was trained with the features, hence, having better parameters, strides, and output data. Due to FDA-NET-3, it can be seen that in the results that it performed better in both cases, i.e., audios and images.

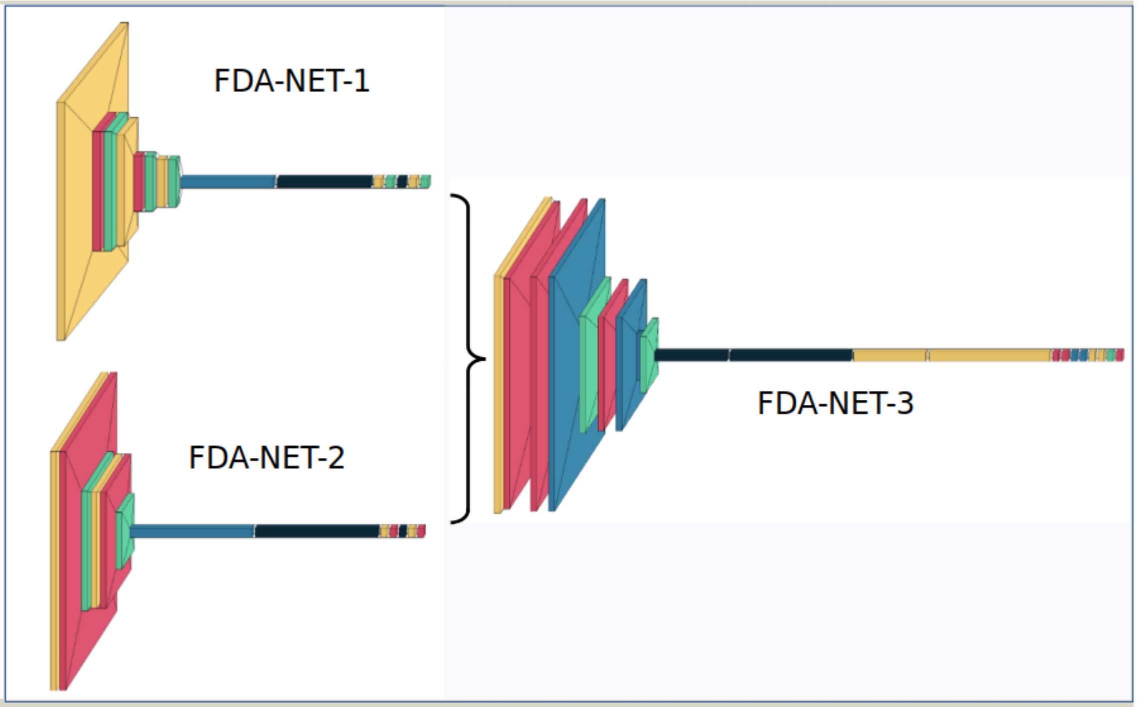


Figure 16: Architectures: (a) FDA-NET-1, (b) FDA-NET-2, (c) FDA-NET-3

4.4 Results and Evaluation

In this section, we present the experiments' results to verify the proposed model's effectiveness, FDA-NET. For comprehensive experiments, we have considered the main matrix for overall dataset accuracy, loss, validation loss, and the total number of trainable parameters for the three proposed models (FDA-NET-1, FDA-NET-2, FDA-NET-3). We have conducted experiments with three proposed network models for both audio and image-based approaches. We present the results and evaluation for different models and strategies.

For the development and experiment of the proposed models, we used the NVIDIA GeForce Â® GTX 1080 Ti, which is packed with 11 Gbps GDDR5X memory, and a 11

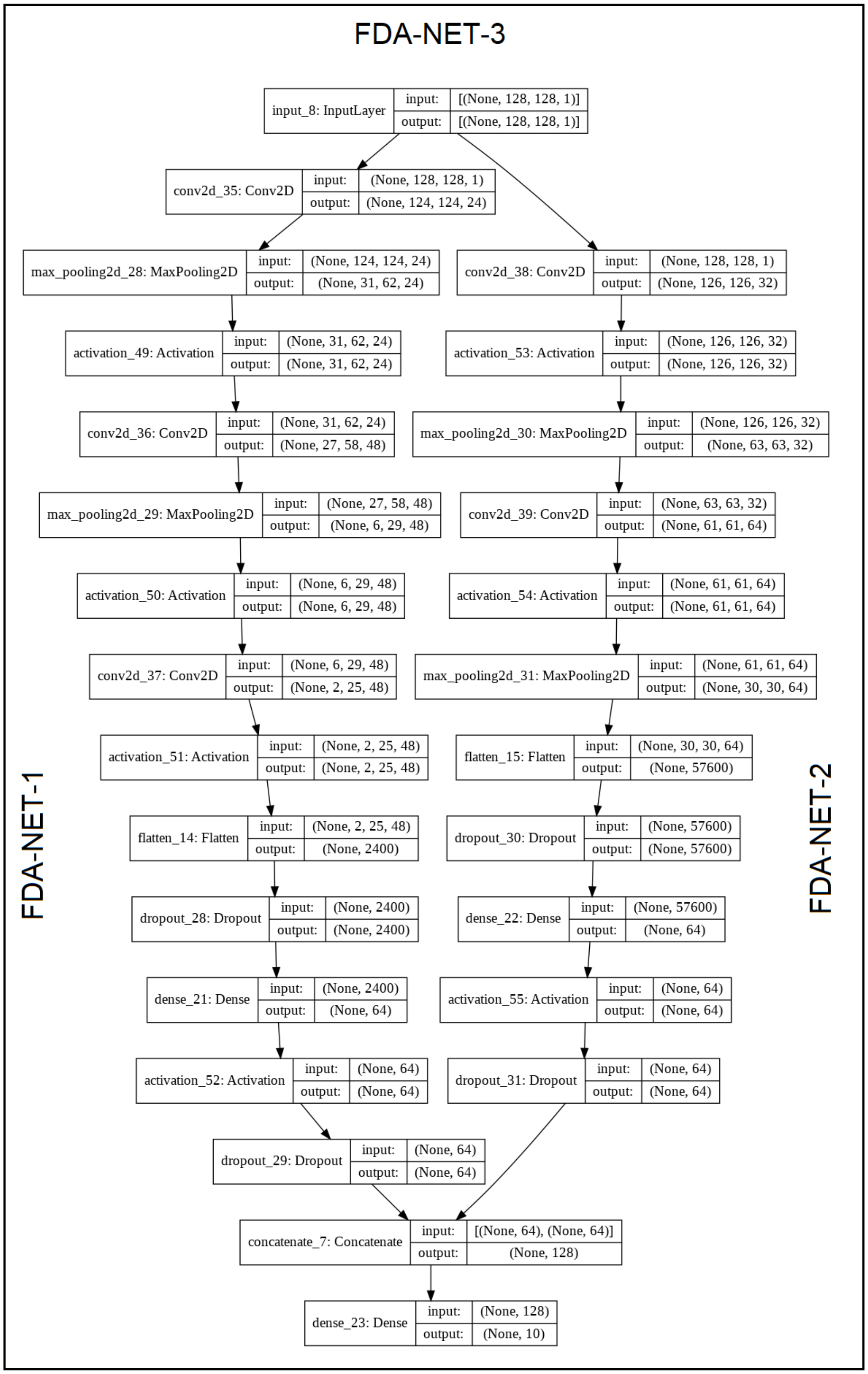

FDA-NET-3
FDA-NET-1
FDA-NET-2
input_8: InputLayer | input: | [(None, 128, 128, 1)] | output: | [(None, 128, 128, 1)]
conv2d_35: Conv2D | input: | (None, 128, 128, 1) | output: | (None, 124, 124, 24)
max_pooling2d_28: MaxPooling2D | input: | (None, 124, 124, 24) | output: | (None, 31, 62, 24)
activation_49: Activation | input: | (None, 31, 62, 24) | output: | (None, 31, 62, 24)
conv2d_36: Conv2D | input: | (None, 31, 62, 24) | output: | (None, 27, 58, 48)
max_pooling2d_29: MaxPooling2D | input: | (None, 27, 58, 48) | output: | (None, 6, 29, 48)
activation_50: Activation | input: | (None, 6, 29, 48) | output: | (None, 6, 29, 48)
conv2d_37: Conv2D | input: | (None, 6, 29, 48) | output: | (None, 2, 25, 48)
activation_51: Activation | input: | (None, 2, 25, 48) | output: | (None, 2, 25, 48)
flatten_14: Flatten | input: | (None, 2, 25, 48) | output: | (None, 2400)
dropout_28: Dropout | input: | (None, 2400) | output: | (None, 2400)
dense_21: Dense | input: | (None, 2400) | output: | (None, 64)
activation_52: Activation | input: | (None, 64) | output: | (None, 64)
dropout_29: Dropout | input: | (None, 64) | output: | (None, 64)
conv2d_38: Conv2D | input: | (None, 128, 128, 1) | output: | (None, 126, 126, 32)
activation_53: Activation | input: | (None, 126, 126, 32) | output: | (None, 126, 126, 32)
max_pooling2d_30: MaxPooling2D | input: | (None, 126, 126, 32) | output: | (None, 63, 63, 32)
conv2d_39: Conv2D | input: | (None, 63, 63, 32) | output: | (None, 61, 61, 64)
activation_54: Activation | input: | (None, 61, 61, 64) | output: | (None, 61, 61, 64)
max_pooling2d_31: MaxPooling2D | input: | (None, 61, 61, 64) | output: | (None, 30, 30, 64)
flatten_15: Flatten | input: | (None, 30, 30, 64) | output: | (None, 57600)
dropout_30: Dropout | input: | (None, 57600) | output: | (None, 57600)
dense_22: Dense | input: | (None, 57600) | output: | (None, 64)
activation_55: Activation | input: | (None, 64) | output: | (None, 64)
dropout_31: Dropout | input: | (None, 64) | output: | (None, 64)
concatenate_7: Concatenate | input: | [(None, 64), (None, 64)] | output: | (None, 128)
dense_23: Dense | input: | (None, 128) | output: | (None, 10)

Figure 17: Fusion Network: (a) FDA-NET-1, (b) FDA-NET-2, (c) FDA-NET-3

GB frame buffer. The experimentations were performed in both Audio (feature) and Image (audio to image spectrogram) based approaches using the three benchmark datasets UrbanSound8K [17], ESC-10 and ESC-50 [16]. For these three datasets, we kept our training to 80%, testing to 20%, and further split the training data into 80% training and 20% validation. The FED-NET models (two DCNN models and one fusion model) were built using the audio and image-based approaches. For the training, the number of epochs was observed at 50 and batch size at 128. All of the models were kept at the same parameters for UrbanSound8K, ESC-10, and ESC-50. We reported the accuracy performance for the proposed models in terms of learning and loss curve graphs, testing accuracy, confusion matrix, precision, recall, and F1-score.

The evaluation metrics were defined based on the Confusion Matrix in which TP is True Positive, FN is False Negative, TN is True Negative, and FP is False Positive. The following metrics were deduced from the Confusion Matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.9}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.10}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.11}$$

$$F1Score = \frac{2xPrecisionxRecall}{Precision + Recall} \tag{4.12}$$

4.4.1 Datasets

4.4.1.1 Original Audio Dataset

We have conducted extensive experiments over three benchmarking datasets of UrbanSound8K [17], ESC-10 and ESC-50 [16].

- **UrbandSound8K Dataset [17]:** This dataset serves as a popular benchmark for environmental sound classification. This dataset consists of 8732 environmental sound clips and is divided into 10-fold cross-validation. Each of the clips is 3 - 4 seconds long. The audio clips included in this dataset are labeled with different categories.

- **ESC-50 Dataset [16] :** This dataset consists of 50 classes divided into 5-fold cross-validation. All audio sounds are divided into Animals, Environment, Human, Indoor, and Urban. There is a total of 2000 sound clips. Each folder consists of 40 clips. Each clip is 5 seconds long. This dataset is also labeled and available for download publicly.

- **ESC-10 Dataset [16] :** This is a publicly available dataset. There is a total of 400 audio files overall divided into ten folders. Each folder consists of 40 sound samples with a sampling rate of 44100 HZ. The ten folders are divided into 5-fold cross-validation. Each sound clip is 5 seconds long on average. The ESC-10 dataset is a subset of ESC-50 dataset.

Table 7: FDA-NET Model Architecture and Parameters

Model	Convolutional Layers	Total Layers	Trainable Parameters
FDA-NET-1	3	5	241,434
FDA-NET-2	2	4	3,705,930
FDA-NET-3	5	9	3,947,354

Table 8: Urban Sound and Environmental Sound Datasets

	Animals	**Nature**	**Human**	**Indoor**	**Outdoor**
UrbanSound8K	Dog bark (DB)		Children playing (CP)	Air conditioner (AC), Drilling (DR), Jackhammer (JH)	Car horn (CH), Engine idling (EI), Gun shot (GS), Siren (SI), Street music (SM)
ESC-10	Dog	Rain	Crying baby	Clock tick	Helicopter
	Rooster	Sea waves	Sneezing		Chainsaw
		Crackling fire			
ESC-50	Dog	Rain	Crying baby	Door knock	Helicopter
	Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
	Pig	Crackling fire	Clapping	Keyboard typing	Siren
	Cow	Crickets	Breathing	Door, wood creaks	Car horn
	Frog	Chirping birds	Coughing	Can opening	Engine
	Cat	Water drops	Footsteps	Washing machine	Train
	Hen	Wind	Laughing	Vacuum cleaner	Church bells
	Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
	Sheep	Toilet flush	Snoring	Clock tick	Fireworks
	Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

Human: Human, non-speech sounds; Nature: Natural sounds capes & water sounds;
Indoor: Interior/domestic sounds; Outdoor: Exterior/urban noises

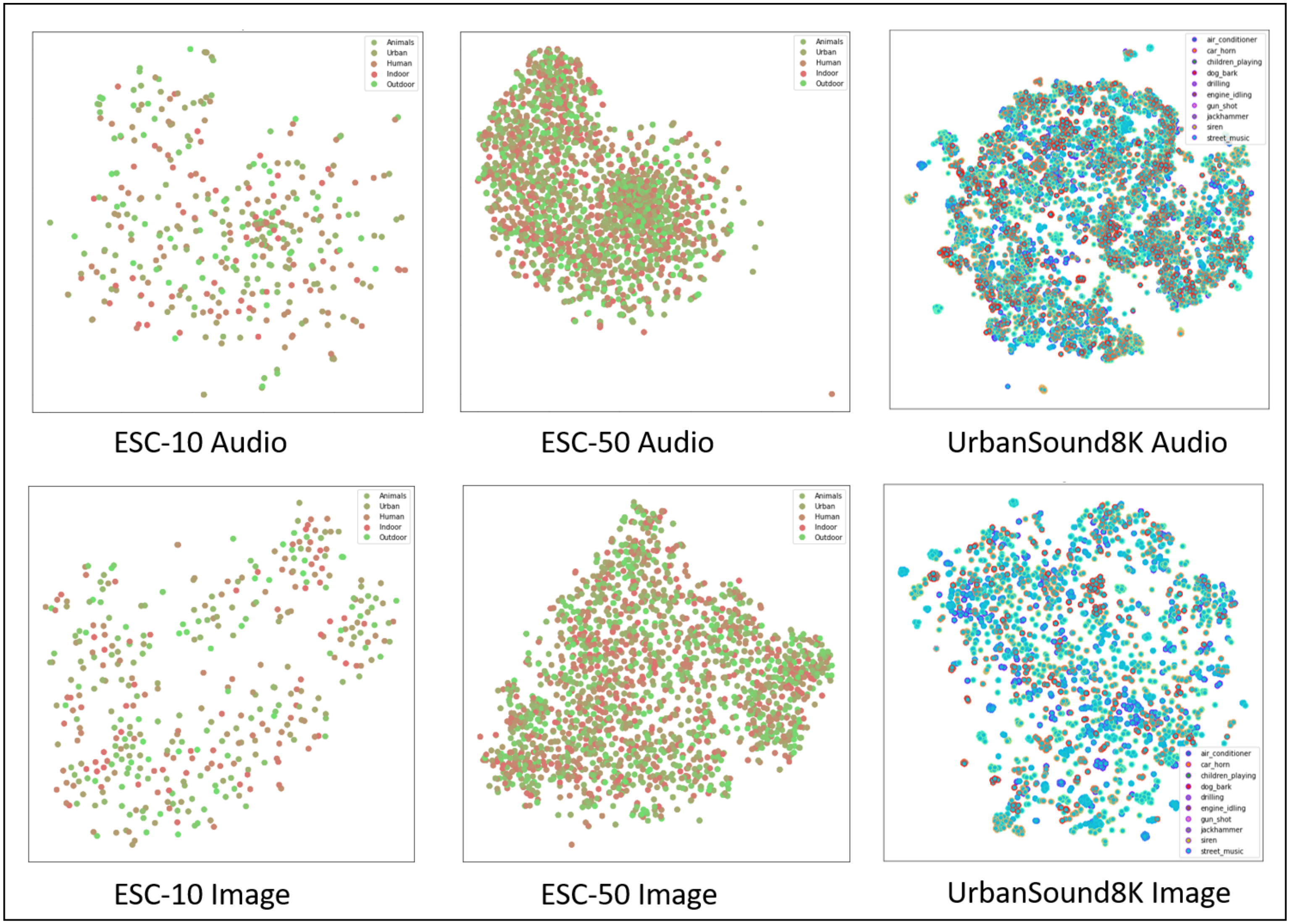


Figure 18: Workflow of Fusion-based Audio Classification

4.4.1.2 Audio and Image Feature Dataset

For the audio clips of the three benchmark datasets, i.e., ESC-10, ESC-50, and Urban Sound 8K in an OGG and Wav format where each audio file was less than and equal to 5 seconds long, ranging between 32000 HZ to 44100 HZ. We generated the audio feature data with the same frame length (30ms) considering a 3s overlapping window. The spectrogram extracted features [75] from these audio files were generated with librosa, and the audio feature vectors (128 x 128) were used in building the CNN models. The numbers of the audio data for the three different benchmark datasets are shown in Table 25.

For the audio feature datasets, the spectrograms in the audio feature vector of size 128 x 128 were converted into the grayscale images of size 128 x 128. As it can be seen in Table 26, the number of images is the same as the number of the audio sample. These images are used to build the 2D convolution neural network models (FDA-NET-v1, FDA-NET-v2, FDA-NET-v3).

4.4.1.3 Augmented Dataset

The spectrograms from the original dataset were generated and saved as an image using the librosa library. Each spectrogram image's dimensions were kept at 128 x 128, which is the FDA-NET network's input shape. The spectrogram images should be clear and large enough for proper image-based classification. However, the nature of spectrograms, i.e., flipping the image of a spectrogram would change the meaning of a visual area may degrade the performance of classification. In this paper, we used audio augmentation rather than using image augmentation. We have applied two different types of augmentation techniques for improving accuracy, i.e., time stretch and pitch shift for different scales. The augmentation helped provide synthesized data for improved training. Besides, we used data normalization for each audio clip. Thus, we applied audio-based augmentation files that were converted into spectrogram images.

For the UrbanSound8K dataset, t-SNE shows the nearest neighbors' visual representation of the original data and the augmented data in Figure 18. The benchmark dataset for UrbanSound8K, ESC-10, and ESC-50 consists of 8732, 400, and 2000 files. After applying data augmentation, the number of files raised to 78588, 3600, and 18000 for UrbanSound8K, ESC-10, and ESC-50, respectively. Data augmentation was used to overcome the issue of insufficient data in building deep learning models. However, the original and augmented datasets were split into training, validation, and testing, and the validation and testing sets were not used for training. There are some limitations of our data augmentation approach. First, it is related to the problem of memory shortage and computational cost. Due to a considerable volume of audio data, it was not easy to perform practical memory management. It was required to process the data locally so that we could not conduct data augmentation technique during model execution [65]. Second, we need to set consistent conditions to allow for fair performance comparison with the state-of-the-artwork research.

4.4.1.4 Data Split for Training and Testing

In training with the data, the number of epochs was kept as 50, while the batch size was 128. However, for experimenting with data augmentation, we kept the batch size was 128, and the number of epochs was kept the same as for testing the original dataset. For training and testing our models, we have split the dataset into 80% for training while 20% for testing. The evaluation was performed on the original and augmented datasets. For example, in the initial UrbanSound8K, the total samples are 8,732, out of which 6,985 were provided for training, and 1,747 were provided for testing. In the same fashion, the entire files for the augmented dataset are 78,588, out of which 62,871 were designated for training, and 15,717 were for testing. The same strategy was applied to ESC-10 and ESC-50 datasets. For data validation, the training samples are further split into 80% training and 20% validation.

4.4.2 Overall Evaluation

We have reported the validation and testing accuracy (Figure 20 and Table 9), the validation of deep learning models (refer to the learning performance (Figure 21, and the validation loss (Figure 22)). In addition, we have evaluated our 18 models (three different benchmark datasets, three different FDA-NET, with augmented data, audio-based or image-based approach) against Precision, Recall, and F1-Score.

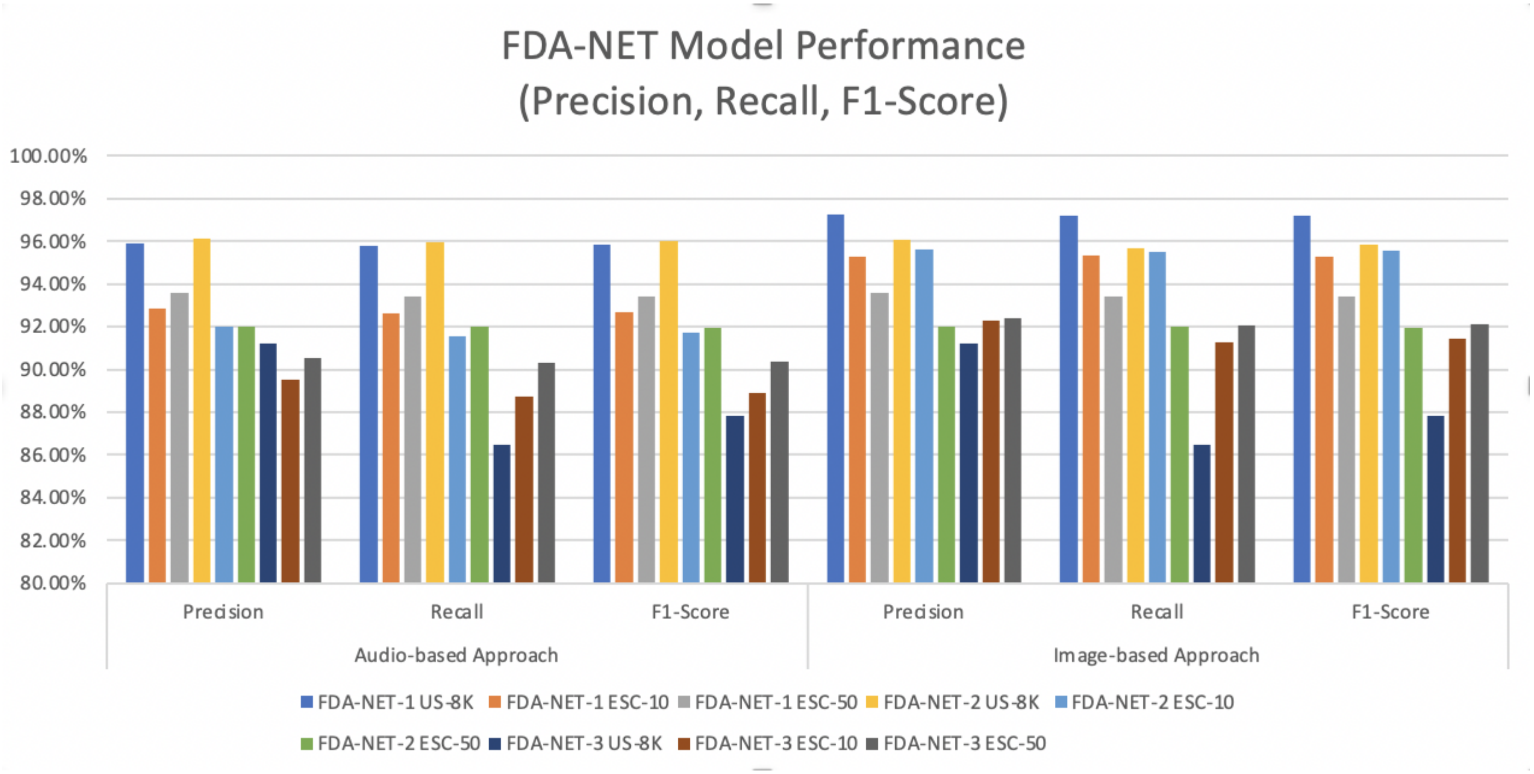


Figure 19: FDA-NET Model Comparison (Precision, Recall, F1-Score) for Overall Models for Environmental Classification

Table 9: Model Comparison for Fusion Based Environmental Classification

Data	Approach	**FDA-NET-1**			**FDA-NET-2**			**FDA-NET-3**		
		US-8K	ESC-10	ESC-50	US-8K	ESC-10	ESC-50	US-8K	ESC-10	ESC-50
Original	Audio	89.83%	90%	85.75%	89.63%	89%	84%	**96.19%**	**91.25%**	**87.25%**
	Image	95.8%	93.75%	86.7%	90%	88%	86.5%	**97%**	**96.67%**	**92.75%**
Augmented	Audio	95.82%	92.63%	92%	95.7%	91.8%	91%	**96.96%**	**93.47%**	**93.1%**
	Image	96.9%	95.27%	92.11%	96%	95.5%	90.67%	**98.53%**	**97.5%**	**96.1%**

US-8K: UrbanSound8K (10 Classes), ESC-10: Environmental Sound Classification (10 Classes), ESC-50 (50 Classes)

4.4.3 Results on UrbanSound8K Dataset

The t-SNE visualization for the UrbanSound8K dataset is shown for both audio and image datasets in Figure 18. The image visualization is shown a more distinctive distribution compared to the audio feature visualization. Furthermore, 78588 audio clips were generated by two common types of data augmentation techniques, in two main types of augmentation such as time stretch and pitch shift, in 4 different variations, i.e., For time stretch, the augmentation was performed for reducing it to the factor of 0.5, 0.7, 1.2 1.5 and including one sample of the original files. Similarly, pitch shifting was applied for raising and dropping the pitch in 4 different ways. We have split the datasets of 8732 (original) and 78588 (augmented) clips in 80% training and 20% testing. These datasets have enough audio clips, and the training performed on this dataset can be considered good enough for testing.

4.4.3.1 Audio-based Classification Results

According to the setting explained previously, we have obtained impressive testing accuracy compared with state-of-the-art accuracy. FDA-NET-a1 for using the audio feature approach reported us testing the accuracy of approximately $\pm 84\,\%$. After data augmentation was applied to the original data, the number of files was generated, as shown in Table 26. The testing accuracy reported from the augmented dataset using FDA-NET-a1 is approximately 95.82%. FDA-NET-a2 reported accuracy of approximately 89.63% for the original dataset and 95.7% for the augmented dataset. Finally, FDA-NET-a3 reported

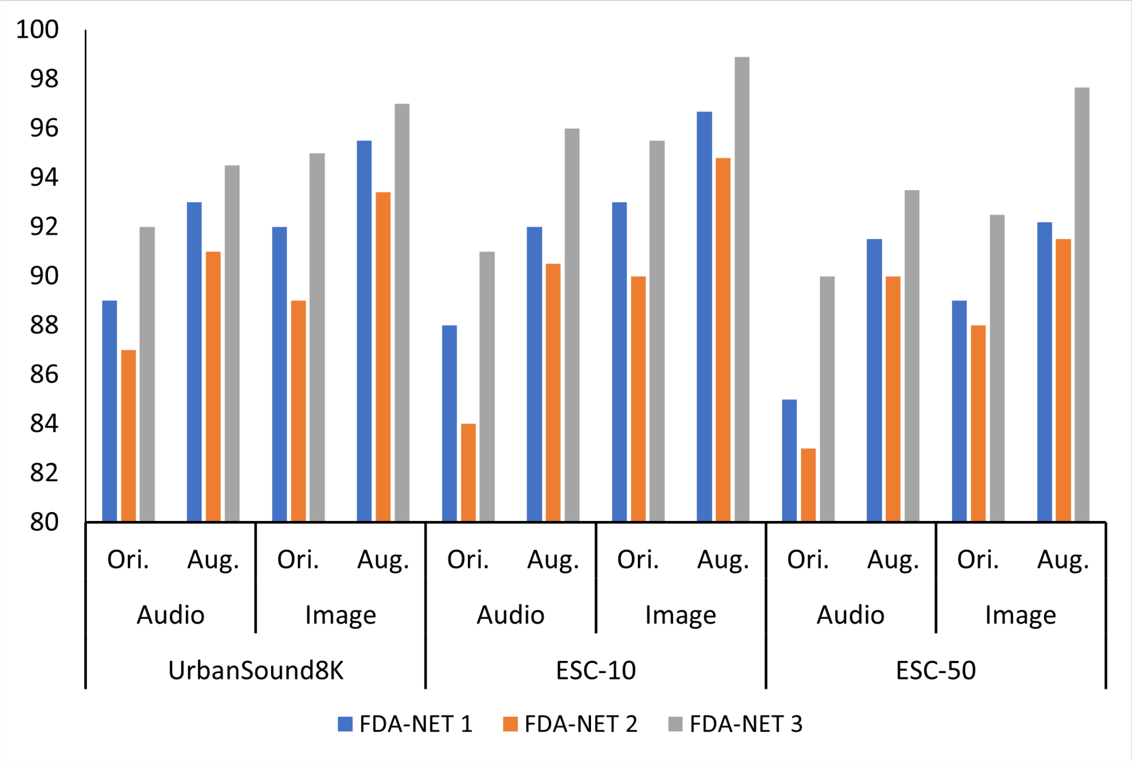


Figure 20: Model Comparison for Fusion Based Environmental Classification

a better-combined testing accuracy of 96.19% for the original dataset, while 98.53% is reported for the augmented dataset. The learning performance is shown in the FDA-NET's model learning curves in Figure 21 and the loss graphs in Figure 22. The class-wise accuracy for the augmented dataset is shown in Table 10, Table 11, and Table 12.

Table 10: FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1464	0	4	0	4	0	2	13	20	24
CH	2	255	0	0	3	0	0	0	1	15
CP	2	0	1451	5	4	5	2	1	48	13
DB	7	0	45	1073	3	2	0	0	13	10
DR	4	4	5	1	1338	1	0	46	2	2
EI	7	1	13	1	1	1556	1	5	4	7
GS	0	0	0	0	0	1	116	0	0	1
JH	2	0	0	0	29	1	0	1471	0	4
SI	1	3	8	1	1	0	0	2	1324	10
SM	4	3	44	2	5	2	0	4	24	1540
Precision	98.06%	95.86%	92.42%	99.08%	96.40%	99.23%	95.87%	95.40%	92.20%	94.71%
Recall	95.62%	93.07%	94.90%	93.06%	95.37%	97.49%	98.31%	97.61%	98.07%	94.59%
F1-Score	96.83%	94.44%	93.64%	95.97%	95.88%	98.36%	97.07%	96.49%	95.05%	94.65%

Table 11: FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1500	1	10	0	6	11	1	0	1	13
CH	0	267	0	0	1	1	0	0	2	0
CP	13	0	1523	17	5	8	0	0	31	41
DB	11	0	24	1084	6	5	0	1	12	16
DR	6	0	4	1	1293	1	0	17	0	11
EI	14	1	10	3	5	1524	0	7	13	6
GS	1	0	0	1	3	0	119	0	1	1
JH	8	0	5	0	31	6	0	1377	0	1
SI	12	8	9	8	4	2	0	5	1337	5
SM	7	5	27	3	1	8	0	0	5	1558
Precision	95.42%	94.68%	94.48%	97.05%	95.42%	97.32%	99.17%	97.87%	95.36%	94.31%
Recall	97.21%	98.52%	93.72%	93.53%	97.00%	96.27%	94.44%	96.43%	96.19%	96.53%
F1-Score	96.31%	96.56%	94.10%	95.25%	96.21%	96.79%	96.75%	97.14%	95.77%	95.41%

Table 12: FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1593	0	0	1	0	3	0	3	17	19
CH	0	239	0	0	1	1	0	0	0	8
CP	1	0	1492	17	11	5	0	1	27	36
DB	105	0	334	724	2	4	0	0	16	4
DR	6	3	0	31	1244	1	0	12	2	3
EI	8	0	4	5	13	1546	1	3	2	2
GS	14	0	4	34	1	0	55	0	0	0
JH	2	0	3	1	89	15	0	1329	1	1
SI	0	1	61	97	0	17	0	3	1256	9
SM	16	17	32	16	10	1	0	0	11	1447
Precision	91.29%	91.92%	77.31%	78.19%	90.74%	97.05%	98.21%	98.37%	94.29%	94.64%
Recall	97.37%	95.98%	93.90%	60.89%	95.55%	97.60%	50.93%	92.23%	86.98%	93.35%
F1-Score	94.23%	93.91%	84.80%	68.46%	93.08%	97.32%	67.07%	95.20%	90.49%	93.99%

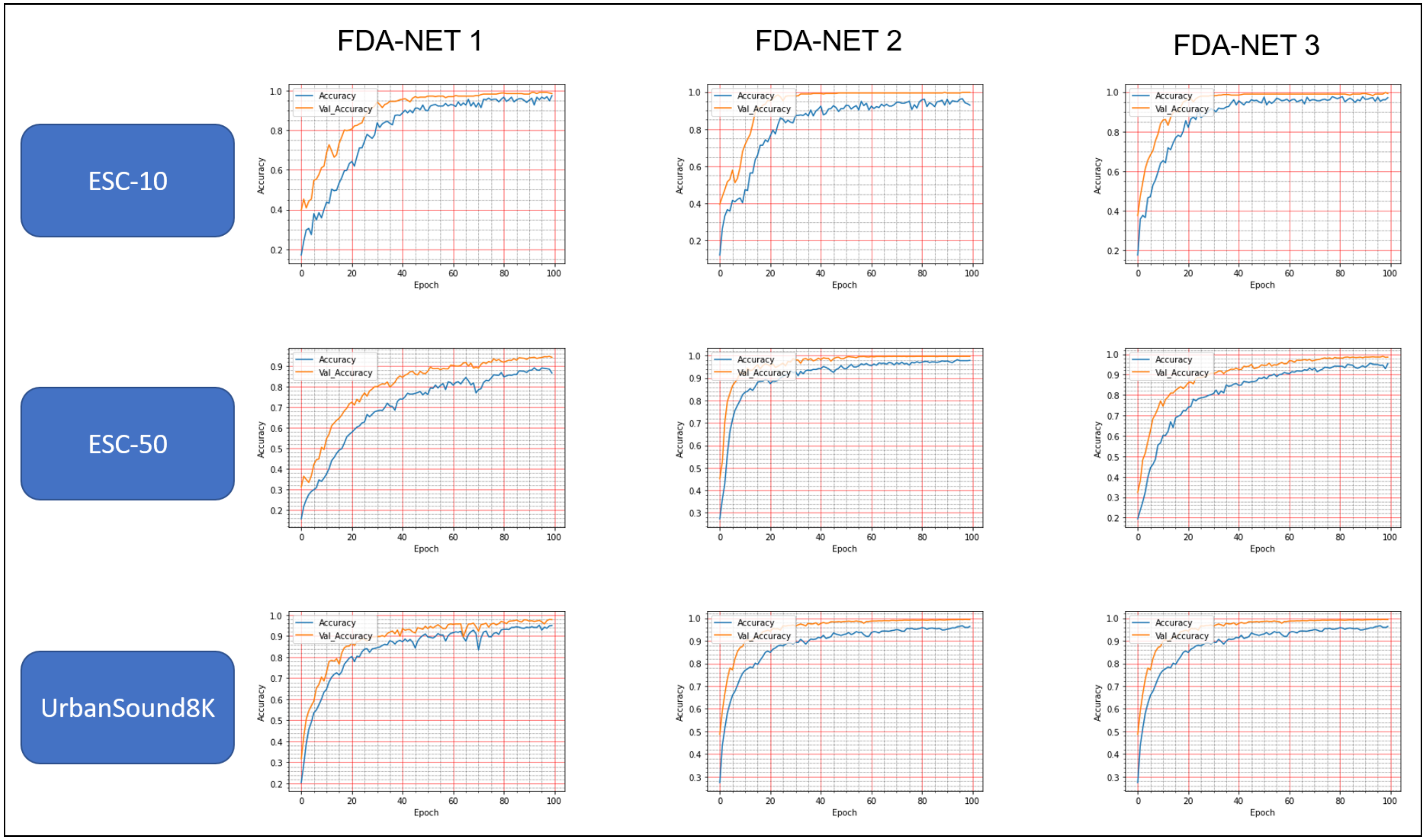


Figure 21: FDA-NET Audio Model Accuracy and Validation Accuracy for FDA-NET-1, FDA-NET-2 & FDA-NET-3 for Original Datasets

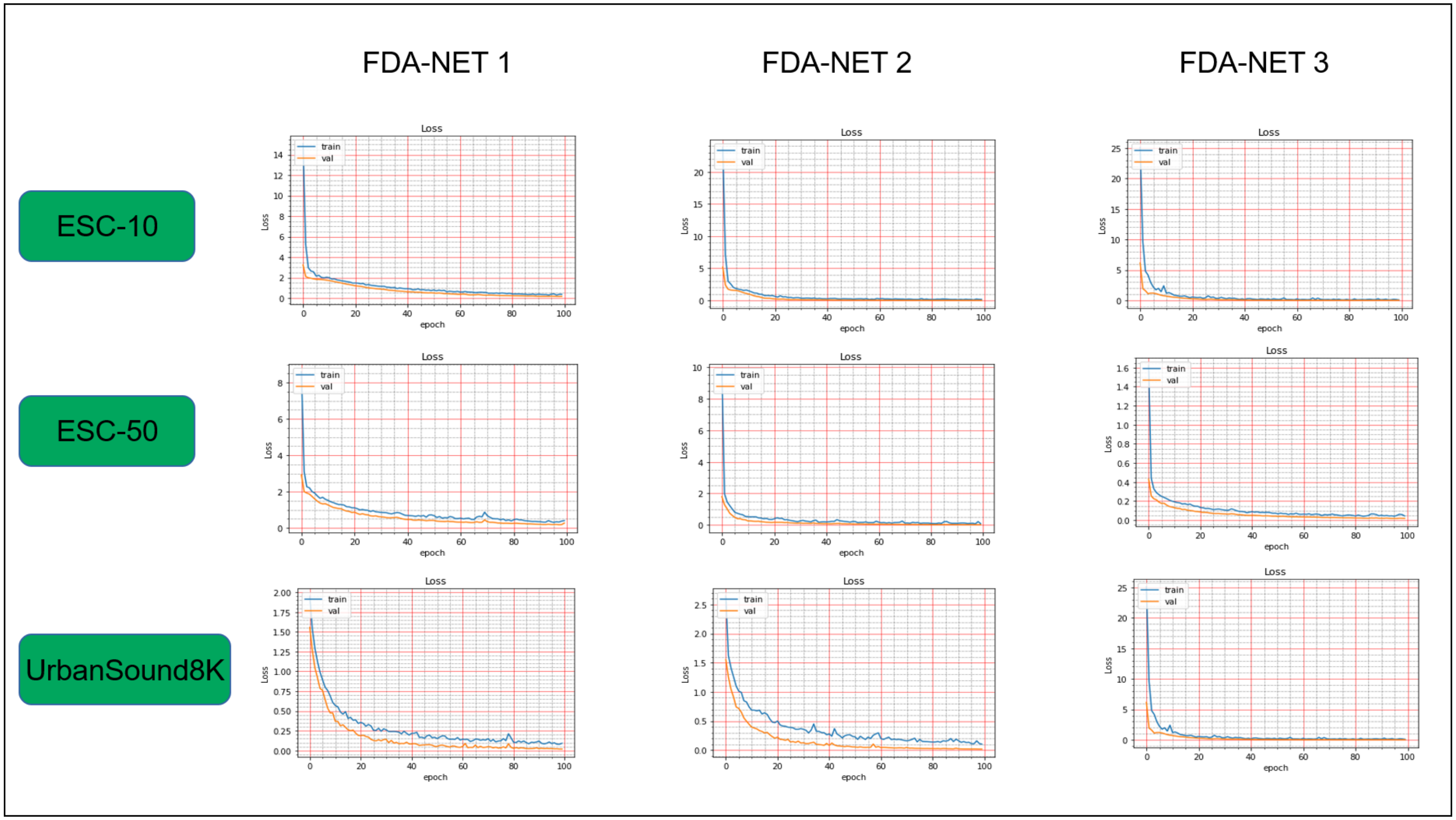


Figure 22: FDA-NET Audio Model Loss vs Validation Loss for FDA-NET-1, FDA-NET-2 & FDA-NET-3 for Original Datasets

4.4.3.2 Image-based Classification Results

The proposed model FDA-NET for the image-based approach is represented by FDA-NET-v1, FDA-NET-v2, and FDA-NET-v3. For the image-based approach, FDA-NETs have the same settings as for the audio-based approach. The image-based approach was more efficient than audio, consuming less time in training the classifiers. Also, image-based classifiers' overall testing accuracy has outperformed the audio-based approach, more importantly, better than most of the state-of-the-art accuracy performance. The accuracy for UrbanSound8K on the image-based approach is shown in Table 9. The image-based results are shown in Figure 20. For the image-based approach with the original UrbanSond8K dataset, the testing accuracy reported is 95.8% for FDA-NET-v1, while for the augmented dataset, it is reported as 96.9% for FDA-NET-v1. FDA-NET-v2's accuracy for the image-based approach is 90%, whereas, for the augmented dataset, it is 96%. However, for FDA-NET-v3, the image-based approach's accuracy is 97%, and 98.53% is the testing accuracy based on the augmented dataset. The overall performance evaluation including Confusion Matrix, Precision, Recall and F1-Score is shown in Table 13, Table 14, and Table 15.

4.4.4 Results on ESC-10 Dataset

The dataset is composed of 400 audio clips and images and 3600 augmented audio clips and images, and each of them is arranged in 5 folders.

4.4.4.1 Audio-based Classification Results

We conducted experiments with the FDA-NET models based on audio features for environmental audio classification, and the results are reported in the performance test accuracy. Table 9 shows the testing accuracy for ESC-10. The learning and loss curves of the models are shown in Figure 21 and Figure 22. FDA-NET-a1, FDA-NET-a2 & FDA-NET-a3 with ESC-10 dataset for the original 400 audio samples reported a testing accuracy of 90%, 89%, 91.25% and for the augmented data a testing accuracy of 92.63%, 91.8%, and 93.47%, respectively. The testing accuracy comparison of ESC-10 for the three models is shown in Figure 20. Table 16 - Table 18 show the augmented dataset performance report.

4.4.4.2 Image-based Classification Results

We now present the classification results for the proposed classifier architecture, FDA-NET-1, FDA-NET-2, and FDA-NET-3, on the image dataset derived from the ESC-10 dataset. The 400 spectrogram images (as shown t-SNE 2D dimensional space in Figure 18) and the 3600 spectrogram images generated from the augmented audio files are reported in Table 26. The testing accuracy for FDA-NET-v1 was reported as 93.75% for the original dataset, while 95.27% was reported for the augmented dataset. The class-wise accuracy for ESC-10 is shown in Table 19 - Table 21. FDA-NET-v1 is better than FDA-NET-v2. However, FDA-NET-v2 for image-approach reports a testing accuracy of approximately 88% and 95.5% for the original and augmented datasets, respectively.

FDA-NET-v3 has better training due to clear features and efficient network, therefore reporting a better accuracy than the FDA-NET-v1 & FDA-NET-v2. The reported accuracy for the original and augmented dataset based on FDA-NET-v3 are 96.67% and 97.5%, respectively. We have obtained the highest classification results among all the tested cases. The overall performance evaluation for ESC-10 augmented dataset's report is shown in Table 19 - Table 21.

4.4.5 Results on ESC-50 Dataset

This section describes the experiments we conducted to assess the impact of the proposed FDA-NET models for environmental audio classification on the ESC-50 dataset. The ESC-50 class categorization visualization is shown in Figure 14. The t-SNE visualization in Figure 18 shows clear distinction of different classes. However, the image-based distribution of t-SNE is better than audio's distribution. Specifically, we conducted the audio and image-based approaches for 2000 original audio clips and 18000 augmented audio clips. The setting for data and networks is similar to ESC-10 Dataset Evaluation.

4.4.5.1 Audio-based Classification Results

The evaluation matrix was generated and confirmed against the prediction evaluation. The accuracy was reported for testing and assessment in terms of validation and prediction. FDA-NET-1 testing accuracy reported for ESC-50 in the original format is 85.75%, while the testing accuracy for ESC-50 in augmented form is 92%. FDA-NET-2 obtained testing accuracy of 84% and 91% for the original and augmented datasets, respectively. However, the highest accuracy for ESC-50 is obtained by FDA-NET-3 is

87.25% and 93.1% for the original and augmented datasets, respectively. The accuracy performance for all models is shown in Table 9. The audio-based performance on ESC-50 dataset in the augmented data are shown in Table 22 - Table 24.

4.4.5.2 Image-based Classification Results

We followed the grayscale feature extraction and arranged the data in the same fashion that was set for audio data. The categories and folders remained the same as for audio. We performed experimentation on three different models of FDA-NET-v1, FDA-NET-v2, and FDA-NET-v3. Images for spectrogram were represented in grayscale as input to the image-based classifiers. The image data for ESC-50 is arranged into 5 folders as the original datasets but in an image format. The performance of the visual design was trained and tested. The visual approach seems to obtain better accuracy due to its efficient visualization in spectrogram format. We converted images to grayscale and further normalized the input by dividing the vectors by 255. The audio normalization and further vector normalization allow us to obtain more precise, accurate results. It is because removing the blam areas from the image may be influenced by accuracy performance. The testing accuracy reported is 86.7%, 86.5% & 92.75% for FDA-NET-v1, FDA-NET-v2 & FDA-NET-v3 for the original image dataset. For the augmented image dataset, the same models reported the testing accuracy of 92.11%, 90.67% & 96.1%. The accuracy performance of the image-based approach is shown in Figure 20.

4.4.6 Comparative Evaluation

We now evaluate the FDA-NET model and compare it with other state-of-the-art algorithms. We achieved the best overall audio classification results as compared to seven other image-based approaches [65, 43, 66, 41, 64, 67] and six other works in an audio-based approach [61, 23, 71, 60, 9, 16] in terms of both classification accuracy and computational efficiency. For a more comprehensive evaluation, we compared the accuracy and network parameter characteristics of our approach with those of both the image and audio-based approaches in the environmental audio classification in Table 6.

Our comparison is based on architecture, the network model, the number of parameters, data augmentation, and features. The comparison is made for all three available datasets, i.e., UrbanSound8K, ESC-10 & ESC-50. The comparison is made with the most recent papers in 2020 until the baseline approach, where visual classification started. The FDA-NET's image-based approach is comparable or superior to those for the current state-of-the-art research. Mushtaq et al. [66] proposed a transfer learning solution for ResNet and DenseNet; they obtained the highest mean accuracy 99.49%, 99.04% and 97.30% for UrbanSound8K, ESC-10, and ESC-50, respectively. Our accuracy is slightly low with a difference of 0.96%, 1.54% & 1.47% for UrbanSound8K, ESC-10 and ESC-50. The accuracy difference is due to their vast network models that can take a lot of memory and time consumption, while our models take way less time in terms of parameters. Their best model has 60.19 Million trainable parameters for their ResNet model and 12.30 Million for the DenseNet model, while FDA-NET has only 3.9 Million trainable parameters. Another critical thing to notice here is that the model performance reported by Mushtaq et al. [65] was the maximum validation accuracy while we reported the testing accuracy. Validation accuracy is typically higher than test accuracy. The parameters of FDA-NET are shown in Table 7.

We also compared our audio-based approach with the state-of-the-art research in audio-based classification, as shown in Table 6. FDA-NET obtained the state-of-the-art accuracy of 96.96%, 93.47%, and 93.1% for UrbanSound8K, ESC-10, and ESC-50, respectively. For UrbanSound8K data, Piczak et al. [16] proposed a baseline approach, while Luz et al. [61] achieved the highest accuracy of 96.16% based on an ensemble approach. Their performance with the UrbanSound8K dataset is comparable to FDA-NET (96.96%). However, their handcrafted feature extraction method requires more computational power than our approach with the librosa spectrogram generation.

4.5 Conclusion

We have worked on three different datasets using a deep learning approach. We have proposed an enhanced Fusion technique on two individual models, i.e., we have combined FDA-NET-1 and FDA-NET-2 to produce FDA-NET-3. FDA-NET-3 is based on the fusion of the other two models. Furthermore, we have considered an audio feature for classification. We further generated spectrogram images for the above datasets in the same format. FDA-NET-1 is five layers of CNN architecture, while FDA-NET-2 is a four layers architecture. We have also discussed how FDA-NET-1 has fewer parameters and better results.

The performance with the audio and image data was influenced by data issues such as noise or imbalance. However, we have considered applying two well-known data augmentation techniques for improving our results, i.e., time stretch and pitch shift. For augmentation of images, we did not consider image augmentation. Instead, we generated spectrograms of augmented audio samples. The presented technique of the data augmentation and normalization provided excellent results. Our validation studies provide fundamental design strategies for improving the classification performance handling the data issues.

The setup explained that the visual model performed better and outperformed all state-of-the-art research in UrbanSound8k, and ESC-10 datasets. ESC-50 performed a little lower because FDA-NET has fewer parameters in training, and the dataset is composed of 50 categories. Hence, saving computation consumption and time. The highest accuracy FDA-NET obtained for the visual model are 98.53%, 97.5%, and 96.1% for UrbanSound8K, ESC-10, and ESC-50, respectively.

Table 13: FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1723	1	3	4	1	40	1	13	2	2
CH	3	698	1	2	2	3	1	0	1	7
CP	0	0	1742	8	1	5	0	0	17	4
DB	5	3	5	1845	1	0	0	0	1	3
DR	8	3	2	3	1651	41	6	76	4	2
EI	27	0	6	3	0	1696	0	20	4	2
GS	0	0	0	2	2	0	698	0	0	0
JH	6	0	0	0	10	36	0	1750	0	0
SI	1	0	2	1	0	9	0	0	1677	1
SM	4	0	20	7	0	7	1	3	9	1771
Precision	96.96%	99.01%	97.81%	98.40%	98.98%	92.32%	98.73%	93.98%	97.78%	98.83%
Recall	96.26%	97.62%	98.03%	99.03%	91.93%	96.47%	99.43%	97.11%	99.17%	97.20%
F1-Score	96.61%	98.31%	97.92%	98.72%	95.32%	94.35%	99.08%	95.52%	98.47%	98.01%

Table 14: FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1500	1	10	0	6	11	1	0	1	13
CH	0	267	0	0	1	1	0	0	2	8
CP	13	0	1523	17	5	8	0	0	31	41
DB	11	0	24	1084	6	5	0	1	12	16
DR	6	0	4	1	1293	1	0	17	0	11
EI	14	1	10	3	5	1524	0	7	13	6
GS	1	0	0	1	3	0	119	0	1	1
JH	8	0	5	0	31	6	0	1377	0	1
SI	12	8	9	8	4	2	0	5	1337	5
SM	7	5	27	3	1	8	0	0	5	1558
Precision	95.42%	94.68%	94.48%	97.05%	95.42%	97.32%	99.17%	97.87%	95.36%	93.86%
Recall	97.21%	95.70%	93.72%	93.53%	97.00%	96.27%	94.44%	96.43%	96.19%	96.53%
F1-Score	96.31%	95.19%	94.10%	95.25%	96.21%	96.79%	96.75%	97.14%	95.77%	95.17%

Table 15: FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on UrbanSound8K Augmented Dataset

Class	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	1593	0	0	1	0	3	0	3	17	19
CH	0	239	0	0	1	1	0	0	0	8
CP	1	0	1492	17	11	5	0	1	27	36
DB	105	0	334	724	2	4	0	0	16	4
DR	6	3	0	31	1244	1	0	12	2	3
EI	8	0	4	5	13	1546	1	3	2	2
GS	14	0	4	34	1	0	55	0	0	0
JH	2	0	3	1	89	15	0	1329	1	1
SI	0	1	61	97	0	17	0	3	1256	9
SM	16	17	32	16	10	1	0	0	11	1447
Precision	91.29%	91.92%	77.31%	78.19%	90.74%	97.05%	98.21%	98.37%	94.29%	94.64%
Recall	97.37%	95.98%	93.90%	60.89%	95.55%	97.60%	50.93%	92.23%	86.98%	93.35%
F1-Score	94.23%	93.91%	84.80%	68.46%	93.08%	97.32%	67.07%	95.20%	90.49%	93.99%

Table 16: FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	133	4	4	1	1
Natural	2	123	5	1	3
Human	1	3	137	3	1
Indoor	0	3	8	147	0
Outdoor	2	0	7	4	127
Precision	96.38%	92.48%	85.09%	94.23%	96.21%
Recall	93.01%	91.79%	94.48%	93.04%	90.71%
F1-Score	94.66%	92.13%	89.54%	93.63%	93.38%

Table 17: FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	134	3	6	4	4
Natural	2	100	5	0	4
Human	0	0	139	4	2
Indoor	3	2	3	168	1
Outdoor	7	0	5	4	120
Precision	91.78%	95.24%	87.97%	93.33%	91.60%
Recall	88.74%	90.09%	95.86%	94.92%	88.24%
F1-Score	90.24%	92.59%	91.75%	94.12%	89.89%

Table 18: FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	150	3	2	4	3
Natural	6	128	1	1	0
Human	9	4	123	5	1
Indoor	3	3	3	139	2
Outdoor	10	3	15	1	101
Precision	84.27%	90.78%	85.42%	92.67%	94.39%
Recall	92.59%	94.12%	86.62%	92.67%	77.69%
F1-Score	88.24%	92.42%	86.01%	92.67%	85.23%

Table 19: FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	148	2	1	2	1
Natural	2	131	0	1	0
Human	4	1	158	4	2
Indoor	3	3	1	136	1
Outdoor	2	2	1	1	113
Precision	93.08%	94.24%	98.14%	94.44%	96.58%
Recall	96.10%	97.76%	93.49%	94.44%	94.96%
F1-Score	94.57%	95.97%	95.76%	94.44%	95.76%

Table 20: FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	153	2	0	2	1
Natural	2	138	1	3	5
Human	0	0	131	0	2
Indoor	5	1	2	122	2
Outdoor	2	1	1	0	144
Precision	94.44%	97.18%	97.04%	96.06%	93.51%
Recall	96.84%	92.62%	98.50%	92.42%	97.30%
F1-Score	95.63%	94.85%	97.76%	94.21%	95.36%

Table 21: FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-10 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	134	0	0	4	0
Natural	4	124	4	1	1
Human	1	3	159	0	0
Indoor	9	0	23	110	1
Outdoor	5	0	4	2	131
Precision	87.58%	97.64%	83.68%	94.02%	98.50%
Recall	97.10%	92.54%	97.55%	76.92%	92.25%
F1-Score	92.10%	95.02%	90.08%	84.62%	95.27%

Table 22: FDA-NET-1 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	739	12	7	2	20
Natural	14	685	6	5	23
Human	11	18	622	10	21
Indoor	11	8	4	661	30
Outdoor	10	8	8	8	657
Precision	94.14%	93.71%	96.14%	96.36%	87.48%
Recall	94.74%	93.45%	91.20%	92.58%	95.08%
F1-Score	94.44%	93.58%	93.60%	94.43%	91.12%

Table 23: FDA-NET-2 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	702	17	32	11	10
Natural	17	639	26	13	17
Human	9	9	646	15	16
Indoor	14	10	15	644	12
Outdoor	11	14	12	9	680
Precision	93.23%	92.74%	88.37%	93.06%	92.52%
Recall	90.93%	89.75%	92.95%	92.66%	93.66%
F1-Score	92.07%	91.22%	90.60%	92.86%	93.09%

Table 24: FDA-NET-3 Audio-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	700	8	19	10	2
Natural	16	663	15	14	10
Human	46	23	574	9	11
Indoor	15	17	18	716	14
Outdoor	25	24	31	16	604
Precision	87.28%	90.20%	87.37%	93.59%	94.23%
Recall	94.72%	92.34%	86.58%	91.79%	86.29%
F1-Score	90.85%	91.26%	86.97%	92.69%	90.08%

Table 25: Audio Classification Benchmark Dataset

Dataset	#Class	#Clip(sec)	#Fold	#File
ESC-10	10	1980	5	400
ESC-50	50	1080	5	2000
UrbanSound8K	10	34920	10	8732

Table 26: Audio and Image Input (Original and Augmented)

	Original Input#		Augmented Input#	
Dataset	Audio	Image	Audio	Image
US-8K	8732	8732	78588	78588
ESC-10	400	400	3600	3600
ESC-50	2000	2000	18000	18000

Table 27: FDA-NET-1 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	739	12	7	2	20
Natural	14	685	6	5	23
Human	11	18	622	10	21
Indoor	11	8	4	661	30
Outdoor	10	8	8	8	657
Precision	94.14%	93.71%	96.14%	96.36%	87.48%
Recall	94.74%	93.45%	91.20%	92.58%	95.08%
F1-Score	94.44%	93.58%	93.60%	94.43%	91.12%

Table 28: FDA-NET-2 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	702	17	32	11	10
Natural	17	639	26	13	17
Human	9	9	646	15	16
Indoor	14	10	15	644	12
Outdoor	11	14	12	9	680
Precision	93.23%	92.74%	88.37%	93.06%	92.52%
Recall	90.93%	89.75%	92.95%	92.66%	93.66%
F1-Score	92.07%	91.22%	90.60%	92.86%	93.09%

Table 29: FDA-NET-3 Image-based Approach: Confusion Matrix and Performance Evaluation on ESC-50 Augmented Dataset

Class	Animal	Natural	Human	Indoor	Outdoor
Animal	788	4	7	11	8
Natural	6	622	13	4	6
Human	18	8	740	10	5
Indoor	38	28	18	571	2
Outdoor	15	36	14	26	602
Precision	91.10%	89.11%	93.43%	91.80%	96.63%
Recall	96.33%	95.55%	94.75%	86.91%	86.87%
F1-Score	93.64%	92.22%	94.09%	89.29%	91.49%

CHAPTER 5

FEATURE-BASED FUSION LEARNING FOR DEEP ACOUSTICS

5.1 Introduction

There have been significant recent advances in deep learning and the potential of the deep learning model for various medical applications. Recently, there has been increasing attention for the classification of human body sounds for clinical conditions in the medical domain [77, 78, 79]. Advanced technologies are essential to achieving the improvement of lifestyle and health care. Some of these applications are ambient assisted living systems [80], fall detection [81], voice disorders [82], and heart condition detection [83]. These systems are useful in the early detection of different types of disease through human body sounds, which ultimately improves healthcare. More specifically, an extensive investigation in a partnership among researchers, health care providers, and patients is integral to bringing precise and customized treatment strategies in taking care of various diseases.

Recently, an electronic stethoscope, similar to the design of standard clinical stethoscopes, was designed to enhance the quality of body sounds through filtering or amplification and then extract features from the sounds for the automatic diagnosis of heart or lung conditions using deep learning algorithms. If a deep learning-based diagnosis of heart or lung disease can increase precision and productivity, we can reduce healthcare costs and improve healthcare quality. Moreover, deep learning is a branch derived from machine learning. It allows the computational models, which consist of several layers of processing used to learn the data representations over multiple levels of abstractions. It has attracted a lot of attention due to its high performance in classification. These learning techniques are among the fastest-growing fields nowadays in the area of audio classification [5]. Some studies reported that the deep learning models outperform humans due to the ability to filter the noise and intensive learning ability [84, 85].

The expenses of health care have been rapidly increased in the United States [86]. Due to the rapid surge of medical care costs, many people cannot afford health care and may have proper medical treatment. A physician can diagnose by using a standard clinical stethoscope by hearing sounds from the human body. An auscultatory method has been applied widely by physicians to examine lung sounds associated with different respiratory symptoms. The auscultatory process has been the easiest way to diagnose patients with respiratory diseases, such as pneumonia, asthma, and bronchiectasis [87]. However, the sound quality is quite a noise or too weak to hear, sometimes due to the complexity of the sound patterns and characteristics. Thus, the manual process takes much time and effort for a physician to detect the condition utilizing a stethoscope accurately [88]. For example, wheezing sounds could not accurately be identified in a series of the pulmonary disease sounds [89].

5.2 Milestone

The contribution of this chapter can be summarized as follows:

- We aim to design a feature-based fusion model transferred from the three feature-based convolutional neural network models to classify lung and heart disease.

- We aim to show that it is more effective to classify heart or lung diseases with images transformed from three different sound features, i.e., Spectrogram, MFCC, and Chromagram.

- Our objective is to apply different types of data augmentation, such as Noise, Pitch-Shift, and Time-Stretch, effectively to the audio dataset for optimal deep learning training and testing performance.

5.3 Related Work

Human body sounds and the signal produced by these sounds play an important role in identifying different types of diseases. The literature survey for lung and heart sound classification is shown in Table 30 and Table 31.

5.3.1 Lung Disease Classification

Rocha et al. [24] developed classification models for the diagnosis of chest conditions by using a stethoscope and for the environmental sounds. First, they created a database of lung sounds, which consisted of 920 samples for different categories (i.e., COPD, Healthy, etc.). The second task of the challenge was to extract the features and classify the sounds according to the nature of sound (Wheezes, Crackles, or both). Finally, they conducted feasibility studies on machine learning algorithms, such as support vector machine (SVM) and artificial neural networks (ANN) using features, such as MFCC, spectral features, energy, entropy, and wavelet coefficients. However, they have not overcome the data issues. Unlike this study, we extracted multiple image features from the same datasets, improved the data issues using data augmentation techniques, and obtained better results.

Several data augmentation techniques have been applied to classify lung diseases using sounds [90, 91, 92, 93]. Dalal et al. [14] explored four machine learning approaches for lung sound classification using lung dataset [30]. This study used data augmentation and extracted Spectrogram, MFCC, and LBP features using multiple machine learning algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), and Convolutional Neural Network (CNN). Among these models, CNN outperformed all other classifiers with an accuracy of approximately 97%. However, their machine utilization was very high by applying almost 1 million or more epochs. On the contrary, we have achieved higher accuracy than the study, with very low machine utilization and only 100 epochs with low parameter consumption.

The review in [22] mentioned several feature extraction and classification techniques for obstructive pulmonary diseases such as COPD and asthma. The process involves several traditional and deep learning classification techniques, such as K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Deep Neural Network (DNN), and Convolutional Neural Network (CNN) and feature extraction through signals such as Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT), Spectrogram, and wavelet transform. For example, the best accuracy for CNN was approximately 95%.

Hai et al. [37] proposed a novel solution for lung sound classification by using a publicly available dataset. The dataset was divided into three categories such as wheezes, crackles, and normal sounds. They proposed a detection method using optimized S-transformed (OST) and deep residual networks (ResNets). They performed preprocessing on the audio samples by using OST for rescaling the features for ResNets and utilized a visual-based approach. Their experimental results showed the best multi-classification model with accuracy (98.7%).

Fateh et al. [94] proposed a pre-trained CNN model to extract deep features by using the ICBHI challenge dataset. The dataset consists of crackles, wheezes, normal, and wheezes plus crackles categories. First, they used the visual approach with spectrogram images generated from lung sounds by extracting the features. Then, they utilized the deep features as the input of the Linear Discriminant Analysis (LDA) classifier using the Random Subspace Ensembles (RSE) method. As a result, their model improved the classification accuracy from 5% compared to the existing methods.

Demir et al. [95] proposed two approaches using CNN for the classification of lung diseases using the ICBHI lung sound dataset. The dataset consists of 4 classes with 6898 recordings. First, they converted the lung sounds to spectrogram images using the short-time Fourier transform (STFT) method. Their first approach is classifying with the SVM based on the features extracted from a pre-trained CNN model. Then, the pre-trained CNN model was fined tuned using the transfer learning method for spectrogram images. The best accuracy for the proposed first and second methods are 65.5% and 63.09%, respectively.

Samiul et al. [96] proposed a lightweight CNN model for detecting respiratory diseases through lung sounds. They designed a hybrid scalogram-based approach using the ICBHI 2017 lung sound dataset by utilizing the empirical mode decomposition (EMD) and the continuous wavelet transform (CWT). As a result, the three-class chronic and the six-class pathological classification accuracy were 98.2% and 98.72%, respectively, with 3M trainable parameters. However, we achieved a better accuracy with lower trainable parameters.

Elmar et al. [97] presented an approach for multi-channel lung sound classification using spectral, temporal, and spatial information. They proposed a convolutional recurrent neural network (CRNN) using spectrogram features to classify lung sounds collected from 16 channel recording devices. Their CRNN model obtained an F1-score of 92% for the binary classification.

Luay et al. [98] proposed homogeneous ensemble learning methods to perform multi-class classification of respiratory diseases. They also used the ICBHI challenge dataset, including 1176 recordings and 308 clinically obtained lung sounds. They used entropy features for machine learning models such as SVM, KNN, and Decision Tree. Among these three models, SVM received the best average accuracy of 98.20%.

5.3.2 Heart Disease Classification

Potes et al. [99] proposed a feature-based ensemble technique for the classification of normal vs. abnormal heart sounds. First, they extracted 124 time-frequency features such as MFCC from the phonocardiogram (PCG) signals. Then, they utilized the combination of AdaBoost classifier and CNN classifier to classify the heart sounds. The overall accuracy achieved was 86%. Zhang et al. [100] proposed a method for heart sound classification using a convolutional neural network (CNN). The spectrogram features were extracted from the cycles of sound signals for different positions of pre-trained CNN, and the classification model was based on a support vector machine (SVM). They reported the precision of 77% and 71% for the two datasets with 4 and 3 classes, respectively.

Bozkurt et al. [38] focused on segmentation and time-frequency components for the CNN-based designs. The Mel-spectrogram and MFCC features were extracted from heart sound data using the PhysioNet dataset [39]. They also performed the data augmentation by changing the sampling rate with a random value in range. The overall accuracy achieved was 81.50%. Shu et al.[101] proposed a novel deep WaveNet model for the classification of heart sounds. They used the dataset, composed of five categories with the 1000 PCG recordings, and obtained the overall highest training accuracy of 97%. Muqing et al. [102] proposed a combined model with a convolutional neural network (CNN) and recurrent neural network (RNN) based on the MFCC features for heart sound classification. Their results for the heart sound classification with pathological or non-pathological categories showed the accuracy of 98% with the PhysioNet database.

Acharya et al. [19] proposed Convolutional Neural Network having nine layers for the classification of heart heartbeat signals, such as non-ectopic, supraventricular ectopic, ventricular ectopic, fusion, and unknown beats. Oh et al. [20] developed hybrid models, i.e., multiple layers of CNN and max-pooling and LSTM as the end layer, to extract the temporal information from the features from the ECG dataset and classify arrhythmia from ECG segments. Rajpurkar et al. [103] developed a 34-layer CNN model for the diagnosis of heart diseases such as arrhythmia with the ECG data, recorded with a single lead heart monitor. However, some issues arise in deep learning modeling with the data, including the limited amount of data for heart conditions, low quality with noise, and significant data variations. We also faced similar problems and addressed them in our work.

5.4 Methodology

We discuss the overall design goals for the FDC network (shown in Figure. 23). The modeling process of FDC includes five stages. First, we apply data augmentation techniques onto the audio data to handle the data issues and improve heart and lung condition detection accuracy. Second, we extract the three types of unique and dominant features inherent from the audio data, i.e., Spectrogram, Mel-frequency cepstral coefficient (MFCC), and Chromagram. Third, we convert the extracted features in the form of the images and generate the feature vectors of the audio images in a color format. Fourth, we feed the image feature vector into the specially designed three convolutional neural network models (FDC-1, FDC-2, FDC-3). Finally, The fusion network model (FDC-FS) fuses these three models to optimize the learning performance for the heart and lung sound datasets.

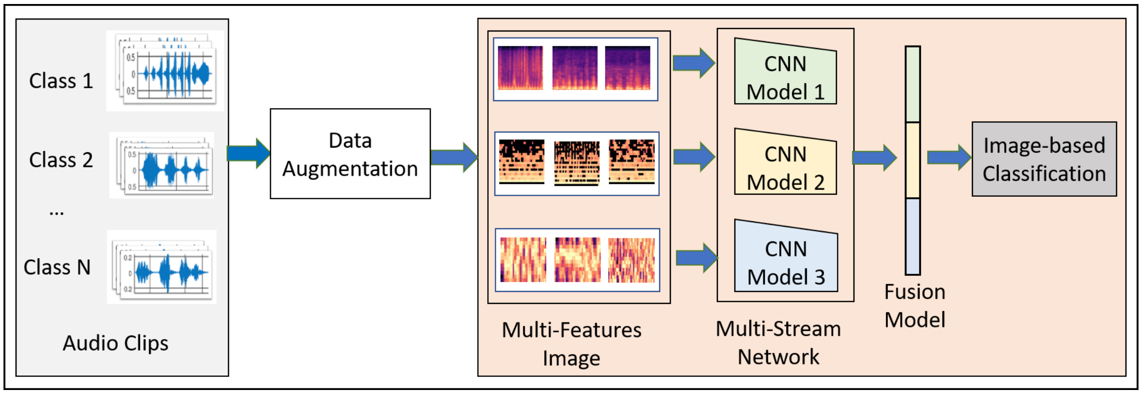


Figure 23: Fusion Based Disease Classification Architecture

5.4.1 Rationale of Design

The rationales of the FDC framework design are as follows: First, it is to enable the effective selection of features from the heart and lung sound data, which are highly noise and unbalanced. Second, it can transform the audio features into consistent and reliable forms, i.e., audio images. FDC supports a multi-modality capability of audio and image in feature extraction, modeling, and inferencing. Third, it is to design the three different network models to effectively learn unique and dominant features. Finally, it is to transfer learning by fusing the three network models into one model to improve the learning performance in lung and heart condition detection.

The differences between the samples of the signals (audio) and images of the same signals are significant, although they are from the same sources. Thus, different modeling techniques are needed to support the multi-modality. For example, if one is talking in a visual context, it means that sound is transparent while the image is considered as non-transparent (opaque) [72]. The pixel formation shows that whatever is the position of an object belongs to the same category. However, audio cannot quickly identify where it belongs due to its observed frequency in spectrogram features. The audio spectrogram depends on the magnitude of the frequency. Therefore, it isn't easy to process the object or its combination in a sequence manner, and it is challenging to identify the simultaneous sounds represented in the spectrogram features.

When a neural network uses images for classification, it considers the sharing weights available from the images in two dimensions, i.e., the X and Y-axis. Thus, the image will carry the same information if it is stretched or repositioned regardless of the position and presence. However, in audio, the two-dimensional data represents the frequency and time. Therefore, if the audio is moved horizontally or vertically, the meaning of the audio can changes. Furthermore, even the spatial features of the sound can change if we increase or decrease the pitch. Therefore, the intentionally introduced invariance can change the meaning, and the neural network will not perform well as it should perform on the trained and augmented data.

In the image format, the extracted pixels can assume that the image belongs to the same class, while in audio, this is not true. Audio is composed of periodic sounds that are composed of frequency and harmonics. The harmonics are spaced apart from each other according to their nature. The combination of these harmonics usually determines the sound's timbre. Suppose we assume from the physical point of view. In that case, the audio played to the audience will examine the type of sound only, while images typically contain a lot of parallel static information. The image classification depends on image features, such as brightness and resolution, among other features.

The design of the fusion-based FDC framework can be justified in terms of the classification effectiveness in terms of three perspectives: (1) Extracting features from the complex audio data of the time and frequent by transformation from audio to images. (2) Designing three specific deep neural networks to optimize their learning performance depending on their unique and dominant features. (3) Optimizing the learning performance through a fusion model by combining the three different models.

5.4.2 Data Augmentation

Deep learning relies on a large amount of data for more accurate classification. Therefore, we have utilized several augmentation techniques for better classification, out of which we have selected the best three methods, including background noise, time-stretching, and pitch shifting. We designed these techniques to address the problems in the lung and heart audio classification.

5.4.2.1 Noise Distortion

We have considered adding random noise to the audio samples to avoid overfitting during training. In this type, noise clips were randomly sampled [104] to be linearly mixed with the input signal represented as y'. α is used to describe random weights along with specific factors that are denoted by U as shown in Equation 5.1.

$$\begin{aligned} Random - Weights &: \alpha \sim_U [0.001, 0.005] \\ Input - Signal &: y' \leftarrow (1 - \alpha) \cdot y + \alpha \cdot y_{\text{noise}} \end{aligned} \tag{5.1}$$

5.4.2.2 Time Stretching

Scaling the audio data horizontally by some stretching factor such as $a\_st > 0$ helps in increasing the size of the data for efficient classification. We have applied time stretch (st) on the audio samples, which were later converted to images. It is to check if the meaning of the data remains the same as image data does not lose the information but changing the position of audio or slowing down the position as we generate the Spectrograms. We have considered four different types of time stretch factor $n \in \{0.5, 0.7, 1.2, 1.5\}$.

5.4.2.3 Pitch Shifting

In this data augmentation technique, the audio samples' pitch is either decreased or increased by four values (semitones) [28]. We assume that with the pitch shifting factor a_s, the artificial training data generated is $Naug$ times larger than the original lung or heart sound data. The duration of audio samples is kept constant like the actual audio samples, i.e., 10 - 90 seconds. For our experimentation, the value changed in semitones were in the interval $[-a_s, a_s]$ for each signal. Factors of pitch shift are $n \in \{-3.5, -2.5, 2.5, 3.5\}$ semitones.

5.4.3 Feature Extraction

We used the two datasets, i.e., lung and heart sound (for six categories for each in the lung dataset and the heart dataset). These datasets consist of sound clips that vary from 10 seconds to 90 seconds. However, to incorporate the consistent data in making

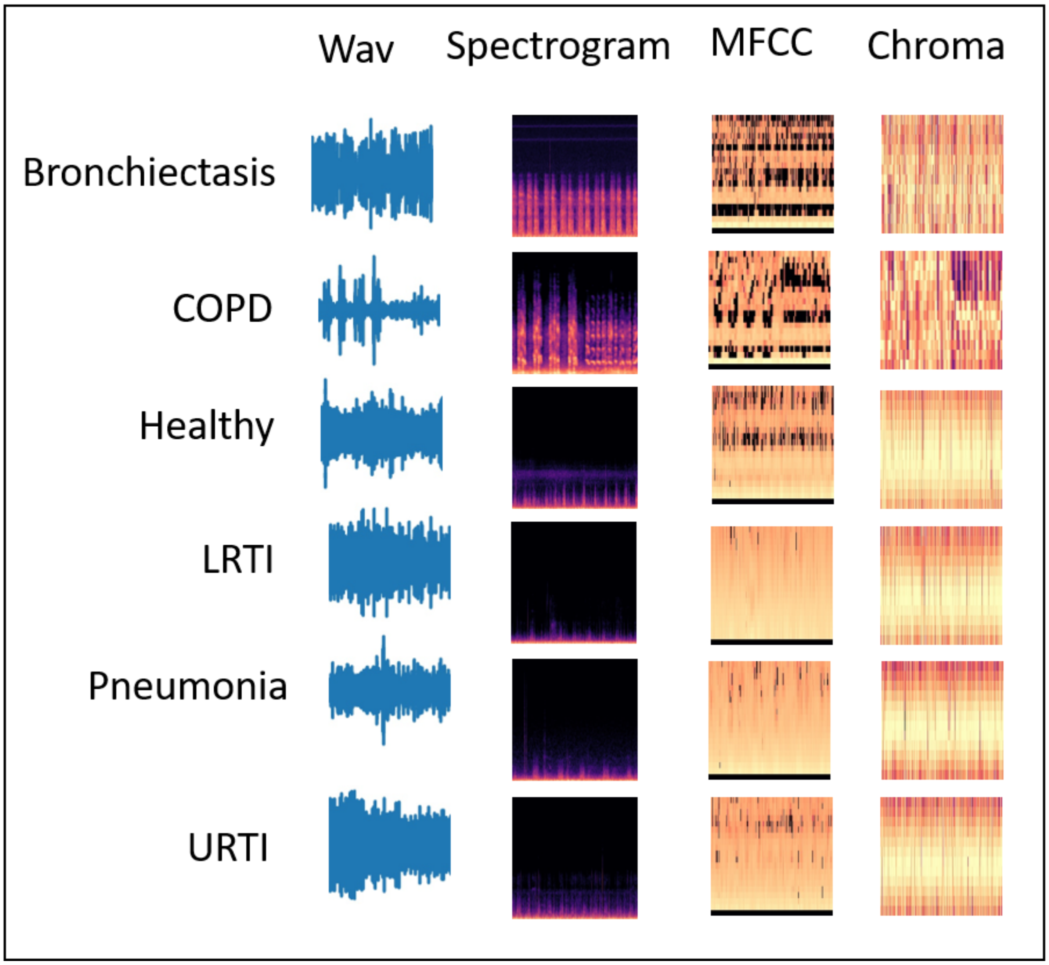


Figure 24: Lung Sound Features: (1) Wav (2) Spectrogram (3) MFCC (4) Chromagram

a more accurate prediction model, we used the sliding window technique to make each clip of 3 seconds. We extract audio features for the given input in terms of Spectrogram, MFCC, and Chromagram. The feature vectors of the extracted features for these three types are converted as JPG images in the dimension of [128x128] using the CV2 image and NumPy libraries.

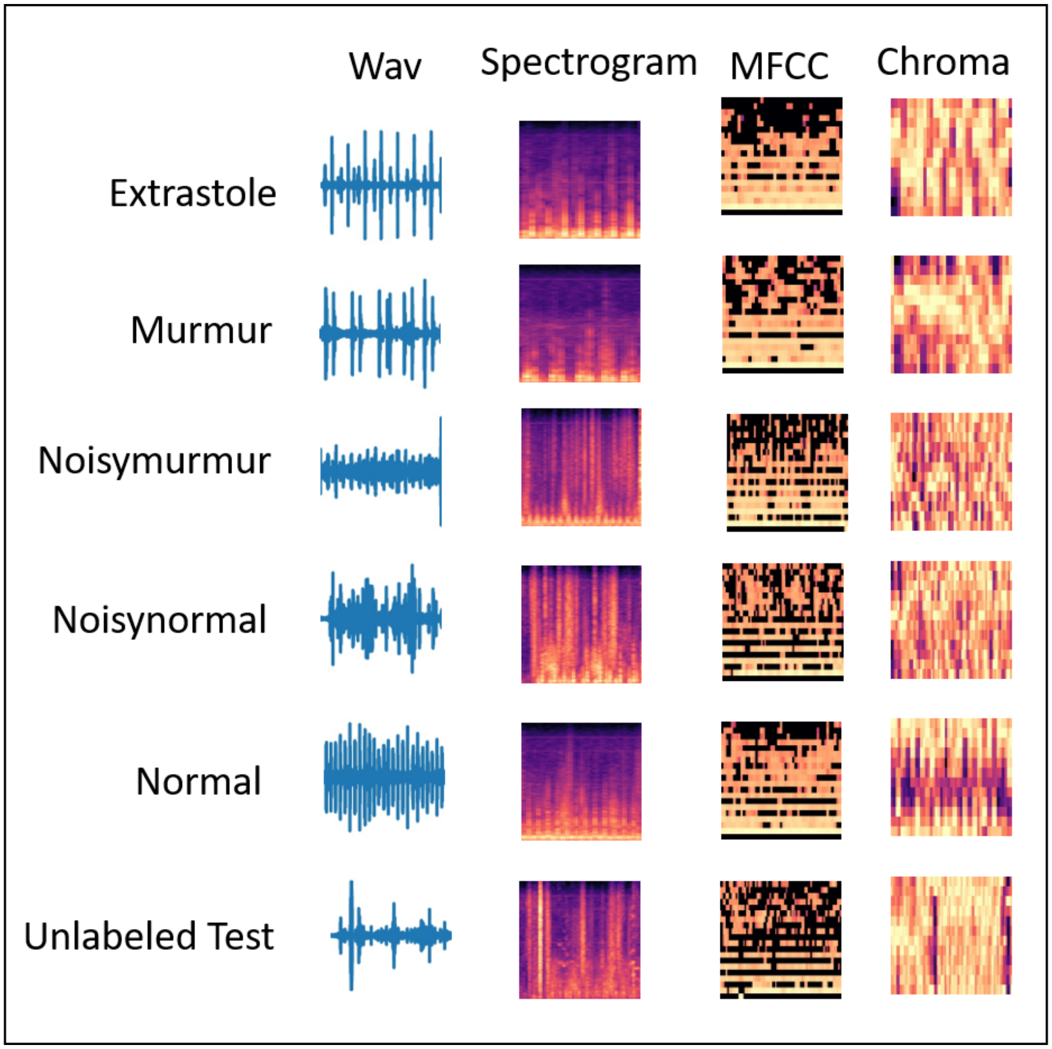


Figure 25: Heart Sound Features: (1) Wav (2) Spectrogram (3) MFCC (4) Chromagram

5.4.3.1 Spectrogram Generation

The spectrogram is a visual representation of a signal in the time-frequency domain. These are generated by the application of the short-time Fourier transform (STFT) [44]. According to the theorem, a single Fourier analysis may not see a nonstationary signal's spectrum variation. The Fourier transform can be used to determine the frequency and sequence of signals and their changes over time. Hence, the spectrogram considers the stationary signal by computing the Fourier transform of the segmented signal into slices. The spectrogram can be calculated as:

$$STFT_x^f(t,f) = \int_{\infty}^{\infty} [x(t)w(t-\tau)e^{-j2\pi ft}dt \tag{5.2}$$

where x(t) is the time-domain signal, τ is the time localization of STFT, and $w(t-\tau)$ is a window function to cut and filter the signal. The length of the window function must be selected and adjusted according to the signal's length because it affects the time and frequency resolution [45]. The spectrogram will be converted into a grayscale image, and the image will be used to generate a feature vector that will be feed for deep learning.

The scaling process is applied to the spectrogram to expand the values range between 0-255 because the range of the spectrogram is usually comprehensive. The method of scaling is done in a linear manner, which can be expressed as follow:

$$S(m,n) = \frac{|Spec(m,n)|}{max|Spec|} \times 255 \tag{5.3}$$

where Spec(m, n) is the value of the spectrogram and S(m, n) is the expanded value from a spectrogram.

5.4.3.2 Mel-frequency Cepstral Coefficient

The Mel-frequency Cepstral Coefficient (MFCC) coefficients are a set of discrete cosine transform (DCT) derived from a type of cepstral representation of the audio clip. The frequency warping allows a better representation of sound by containing the difference between the cepstrum and the Mel-frequency cepstrum. It computed through logarithmic spectrum scale after it was transformed to the Mel scale [105] calculated as:

$$\mathrm{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.4}$$

For the input signal $y(n)$, an N-point discrete Fourier transformation (DFT) is given as:

$$Y(k) = \sum_{n=1}^{M} y(n).e^{\left(\frac{-j2\pi nk}{M}\right)} \tag{5.5}$$

MFCCs are commonly derived as follows: first, we obtain the spectrum by taking the Fourier to transform to a signal. Second, we map the spectrum onto the Mel scale and then take a log-based transform of the Mel-frequency scaled spectrums. Finally, we take the discrete cosine transform of the Mel-log-frequency scaled spectrums and amplitude the spectrum.

5.4.3.3 Chromagram

Chromagram features or Chromagram are Pitch Class Profile whose pitches can be meaningfully categorized. The Chromagram technology was applied to generate a robust set of acoustic features by capturing harmonic and melodic characteristics of music in a signal whose pitches can be classified in the categories of lung or heart sounds. Since the heart and lung sounds have subtle differences in pitch, Chromagram has features that make it a good source of lung and heart sound classification.

5.4.4 Classification Model

5.4.4.1 Overview of the FDC Model

Convolutional neural network (CNN) has been recognized as a popular and powerful deep neural network model in audio and image classification applications.

We have developed a fusion model based on CNN-based architecture for heart and lung disease classification. Our model is a 2D CNN model composed of the input layer, convolutional 2D layer, max-pooling layer, and fully connected layers. The invention is a fusion model, FDC-FS, by combining multi-featured models, including FDC-1, FDC-2, and FDC-3. The fusion was conducted on their final layer to append all model parameters.

There are two essential components in the design of multi-feature models with CNNs. 1) the feature extractor collected features (Spectrogram, MFCC, and Chromagram) from the audio signals and transformed the features into images to generate visual feature vectors. 2) Each model (FDC-1, FDC-2, FDC-3) was uniquely designed to optimize learning for the specific features (Spectrogram, MFCC, and Chromagram), respectively which is composed of multiple convolutional and pooling layers, activation, and fully connected layers with several hidden units. Finally, the fusion model was built by composing the features from these three models. After the models were trained, the visual feature vectors of the input signals were classified by the models into their appropriate categories.

The mathematical form of the convolutional layers is given in Equation 5.6 and 5.7

$$[x^l_{i,j,k} = \sum_a \sum_b \sum_c w^{(l-1,f)}_{i,j,k} y^{(l-1)}_{i+a,j+b,k+c} + bias^f] \tag{5.6}$$

$$[y^l_{i,j,k} = \sigma(x^{(l)}_{i,j,k})] \tag{5.7}$$

The output layer is represented by $y^l_{i,j,k}$ where as the 3-dimensional input tensor is denoted by i, j, k. The weights for the filters are denoted by $w^{(l)}_{i,j,k}$ and $\sigma(x^{(l)}_{i,j,k})$ describes the sigmoid function for linear activation. The fully connected is the final layer represented by Equation 5.8 and 5.9.

$$[x^{(l)}_i \sum_j w^{l-1}_{i,j} y^{l-1}_j + bias^{l-1}_j] \tag{5.8}$$

$$[y^l_{i,j,k} = \sigma(x^l_{i,j,k})] \tag{5.9}$$

Table 30: Literature Survey for Lung Sound Classification

Publication	Network	Feature	Approach	Limitation
Dalal (2018)	SVM, KNN, GMM, CNN	MFCC, LBP	An ensemble-based approach for classification lung sounds	Machine utilization was very high by applying almost 1 million or more epochs
Hai (2019)	ResNet50	Sp	A visual-based approach to propose a detection method using optimized S-transformed (OST) and deep residual networks also known as ResNets.	Large number of trainable parameters and low data
Fatih (2020)	CNN	Sp Images, Deep Features	An ensemble-based approach for classification lung sounds	Low performance in terms of accuracy
Demir(2020)	CNN, SVM	Sp, Sp Images	Transfer learning method for classification of lung diseases	Large number of trainable parameters with high computational power. Low performance in terms of accuracy
Samiul (2020)	CNN	Hybrid Scalogram	A lightweight CNN model for detecting the respiratory diseases through lung sounds	Trainable parameters are high
Luay (2021)	SVM, KNN, DT, KNN	Entropy features	Examines the application of different homogeneous ensemble learning methods to perform multi-class classification of respiratory diseases	Lack of deep learning performance

Table 31: Literature Survey for Heart Sound Classification

Publication	Network	Feature	Approach	Limitation
Potes(2016)	CNN	MFCC	A feature-based ensemble technique for classification of normal vs abnormal heart sounds. The combination of AdaBoost classifier and CNN classifier to classify the heart sounds	Low performance in terms of accuracy
Zhang(2017)	CNN+SVM	SP	A method for heart sound classification without segmentation using CNN and achieved the last classification with SVM	Low performance in terms of accuracy
Bozkurt(2018)	CNN	MFCC, Mel-SP	Focus on segmentation and time-frequency representation components of the CNN-based designs with data augmentation techniques	In comparison to our work low performance in terms of accuracy
Wu(2019)	CNN	Sp, Mel-SP, MFCC	An ensemble approach for classification of normal and abnormal heart sounds	Large number of trainable parameters with low performance
Shu(2020)	WaveNet	Multiple features	A novel deep WaveNet model for classification of heart sounds	Obtained high training accuracy with low data however, the testing performance is low in comparison
Xiao(2020)	1D CNN	Raw signal w/t band filter	A transition approach with 1D CNN for classification of PhysioNet data	Low performance in terms of accuracy with low data
Muqing(2020)	CRNN, PRCNN	MFCC	An improved feature extraction method using MFCC. For classification of heart sounds, the convolutional neural network (CNN) and recurrent neural network (RNN) models were combined	Results obtained with balanced dataset
Mehmat(2021)	1D CNN	LBP+LTP	A one-dimensional CNN approach for classification of PhysioNet data	Low performance in terms of accuracy

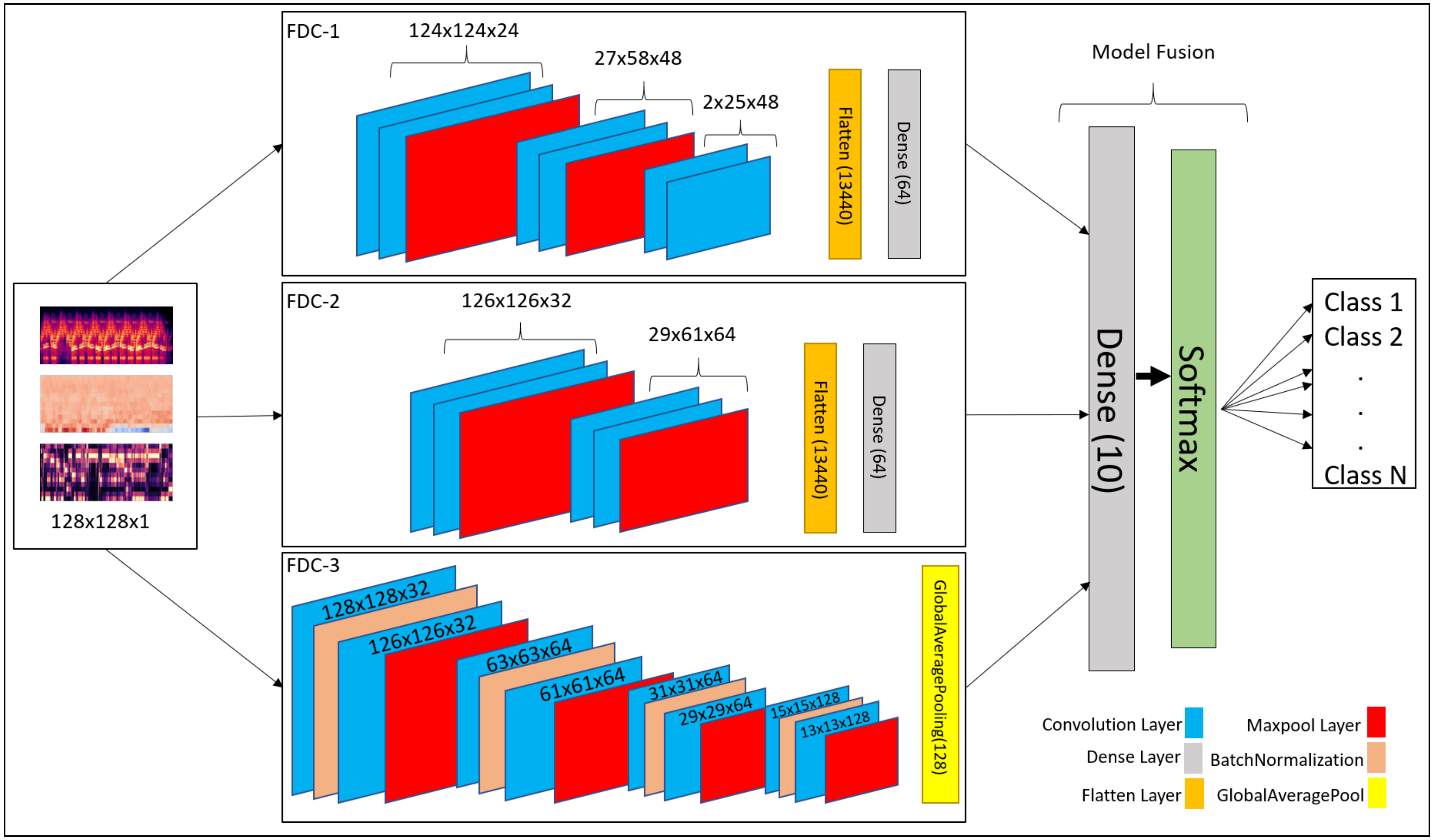


Figure 26: Overall FDC-FS Architectures (a) FDC-1 (b) FDC-2 (c) FDC-3

Our fusion model FDC-FS is composed of three different models, such as FDC-1, FDC-2, and FDC-3. They consist of the convolutional layers enclosed by the max pool layer, followed by fully connected layers, including dropout, batch Normalization, rectified linear units (ReLU), and LeakyReLU. During the extraction of features, we have used the window size and hop size of 23 ms. As the sound clips vary between 3 to 5 seconds, that is why we kept the extraction to 3 seconds to make every bit of the sound clip usable. In addition, we have reshaped the input taken from the sound clips to $X \in R^{128x128}$ shape. Further, we have sent these reshaped features to the classifier to predict heart or lung diseases.

5.4.4.2 FDC-1 Model

The FDC-1 model is designed for the classification based on the image feature vector of the three audio features (Spectrogram, MFCC, and Chromagram) for the given datasets, using convolutional neural network architecture with a total of five layers. Among the five layers, 3 are convolutional layers, and 2 are dense layers. We considered rectified linear units (ReLU) as the activation function between layers, a max-pooling is also applied, and we have also used dropout in different layers to avoid overfitting. The total number of trainable parameters based on the five layers of architecture is 241,174 (0.24 M). The hyper-parameters are shown in Table 32.

The first layers of the FDC-1 model consist of 24 filters with a 5*5 receptive field. The layer is also followed by a (4*2) strided max-pooling function. The activation function used in this layer is rectified linear units (ReLU). The second layer of FDC-1 is composed of 48 filters with 5*5 receptive files. It is followed by 4*2 strided Max Pooling and ReLU activation. The padding for these layers is kept as "valid." The third layer of FDC-1 consists of 48 filters with 5*5 receptive fields. The layers have "valid" padding, which is followed by the ReLU activation function. After the activation, the output is flattened, and the dropout of factor "0.5" is applied to avoid overfitting the output from layer to layer. The fourth layer is the first dense layer which is also called the hidden layer. It consists of 64 hidden units followed by ReLU activation and dropout rate of 0.5 to avoid overfitting the output result to the next layer. The fifth layer is a final dense layer that consists of output units. The output units are always equal to the number of classes used in the dataset. The last layer is followed by the "Softmax" activation function.

5.4.4.3 FDC-2 Model

The FDC-2 model is designed for classification based on the image feature vector of the three audio features (Spectrogram, MFCC, and Chromagram) for the given datasets. It is based on a convolutional neural network architecture consisting of 4 layers, including two convolutional layers, two hidden layers, and an L2 regularizer on the first layers to reduce likelihood and bias among the inputs. In addition, this model consists of Max Pooling to reduce unwanted features for training, dropout to avoid overfitting, ReLU, and Softmax activation. Feature vector flattening is also considered to convert 2-dimensional features to 1-dimensional features. The total number of trainable parameters based on the four layers of architecture is 879,430 (0.87 M). The overall hyper-parameters are shown in Table 33.

In the first layers of FDC-2, the layers take 32 filters with 3*3 receptive files. The first layers also consist of the L2 regularizer norm with the value of “0.0005.” Then, Strided Max Pooling of 4*2 follows it. Finally, ReLU is used as an activation function. FDC-2 takes 48 filters with 3*3 receptive filed and “valid” padding in the second layer. It is pursued by the ReLU activation function and Max Pooling of 4*2. After all the operations above, the 2-Dimensional input is flattened to 1-Dimensional and passed on to the hidden layers, i.e., dense layers. A dropout follows the flatten with a rate of 0.5 to avoid overfitting the input. The third layer is the first hidden (dense) layer of FDC-2 with hidden units of 64, followed by ReLU activation and dropout with a rate of 0.5. The fourth layer is a dense layer consisting of the output units, which is equal to the number of classes available in the dataset. The final activation function is Softmax.

5.4.4.4 FDC-3 Model

The FDC-3 model is designed for the classification based on the image feature vector of the three audio features (Spectrogram, MFCC, and Chromagram) for the given datasets, focused more on in-depth training and eventually reducing the number of trainable parameters. This model is composed of 8 convolutional layers and one dense layer. The layers consist of padding, ReLU, softmax activation, Max Pooling, Global Average Pooling, Batch Normalization, and dropout. The Batch Normalization is used to train the model intensely, and in return, it standardizes the input in a layer for each mini-batch. Hence, it has a perfect effect on the learning process, reducing the number of trainable parameters. The total number of trainable parameters based on the nine-layer architecture given below is 362,214 (0.36 M). FDC-3 hyper-parameters are shown in Table 34 in details.

The first and second layers of FDC-3 have 32 filters with 3*3 receptive fields and some padding and ReLU as activation function. Both layers also consist of 2*2 strided Max Pooling. Batch Normalization follows both layers to perform deep training and reduce the trainable parameters. However, the second layer is following by a dropout of 0.25 to overfitting the input to the next layer. The third-to-sixth layers of FDC-3 are the same as the first and second layers, but the third-to-sixth layers take 64 filters with 3*3 receptive fields. They use the same strided max-pooling, padding, activation, dropout, and Batch Normalization for deep training. The seventh and eighth layers of FDC-3 take 128 filters with 3*3 receptive files, followed by the same padding and ReLU activation function. Batch Normalization follows the activation function in both layers. However, the eighth layer is followed by the dropout of rate 0.25. Finally, the last convolutional layer is also followed by Global Average Pooling before the input is ready for output classification. The ninth layer is the final and only the dense layer in this architecture, consisting of output units equal to the number of classes in the dataset.

5.4.4.5 FDC-FS Model

The FDC-FS model is a fusion model resulting from transfer learning from all three models (FDC-1, FDC-2, FDC-3). Its architecture is composed of the softmax activation and dense layers consisting of the output units equal to the number of classes in the dataset at the last layer. Therefore, FDC-FS is composed of 13 convolutional layers and 3 dense layers model, more specifically three convolutional layers from FDC-1, two convolutional layers from FDC-2, eight convolutional layers from FDC-3, one dense layer from FDC-1, and one dense layer from FDC-2, and a final dense layer of an output unit. The FDC-FS's total trainable parameters for six classes are 1,482,806 (1.4M). Table 35 and Table 36 show the hyper-parameters of the final convolutional architecture, which is a fusion of our novel three architectures shown in Figure. 26.

5.5 Result and Evaluation

We have conducted comprehensive experiments for the FDC models using the lung and heart sound datasets [24, 106]. We now present the results obtained from the experiments with the three FDC models and the fusion model (FDC-FS) using the original and augmented datasets of the lung and heart datasets. The results includes the accuracy, loss, and class-wise accuracy for original and augmented datasets. The experimental results have been obtained as compared to the state-of-the-art methods.

5.5.1 Experimental Setup

We have conducted most of the experimentations using Google research collaboratory with 12 GB NVIDIA Tesla K80. For the data augmentation and feature extraction, we used the NVIDIA GeForce Â® GTX 1080 Ti, packed with 11 Gbps GDDR5X memory and an 11 GB frame buffer. The number of epochs was set at 50 while avoiding overfitting, and a batch size of 64 was considered. However, for the training for the augmented datasets, the epochs were set at 30 with a 128 batch.

The model training was set to 80% training and 20% testing. From the 80% training data, we have further split it into 80% training and 20% validation (64% training, 16% validation, 20% testing). We have reported classification accuracies for training and testing. The class-wise accuracy was also reported for four different models (FDA-1, 2, 3 & FS), two types of data (original and augmented), and for two different datasets (lung and heart sound). We observed that FDC-FS performed the best compared to others.

5.5.2 Dataset

Lung Sound Dataset: The research team from Greece and Portugal created the lung dataset [24]. There are 920 annotated recordings, ranging from 10s to 90s (the total of 5.5 hours), obtained from 126 patients using a digital stethoscope. Unfortunately, the complete dataset was not released. Therefore, the publicly available dataset used for lung sound classification modeling is mainly limited in data amount and sound quality. To overcome the data issues, we applied two approaches to balance the dataset: (1) The "Synthetic Minority Oversampling Technique (SMOTE)" replicates the same sample several times to balance the dataset with other classes. Most of the state-of-the-art research use SMOTE in their works. The second approach is to consider weighted average as our testing accuracy for the models.

We further applied data augmentation techniques to generate synthesized data. Table 37 shows the amount of the original and augmented data. We have removed the

Asthma category having only a single recording from the dataset. The number of the original recordings was 919 for six categories (*Bronchiectasis*, *COPD*, *Health*, *LRTI*, *Pneumonia*, *URTI*) while the number of augmented recordings are 10,109.

Heart Sound Dataset The heart dataset consists of 656 audio recordings for different heart classes such as *Extrastole*, *Murmur*, *Noisy Murmur*, *Noisy Normal*, *Normal*, *Unlabeled test*. The author of the dataset [106] made this dataset public for two challenges, using an iPhone app and a digital stethoscope. The initial dataset has both clean and noisy data without any data synthesis. In order to increase the accuracy of the data, data augmentation techniques were added to the initial heard sound samples. After applying the data augmentation to the initial data, the total number of files increased to 7216 as shown in Table 38.

Figure. 27 shows the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization for lung and heart sound dataset. It can be seen that the Lung dataset is very dispersed, and classes are mixed up with each other.

Table 32: FDC-1 Model Hyper-parameters

Layer(Type)	**Output Shape**	**Param #**
Conv2D	(None, 124, 124, 24)	624
MaxPooling2D	(None, 31, 62, 24)	0
Activation	(None, 31, 62, 24)	0
Conv2D	(None, 27, 58, 48)	28848
MaxPooling2D	(None, 6, 29, 48)	0
Activation	(None, 6, 29, 48)	0
Conv2D	(None, 2, 25, 48)	57648
Activation	(None, 2, 25, 48)	0
Flatten	(None, 2400)	0
DropOut	(None, 2400)	0
Dense	(None, 64)	153664
Activation	(None, 64)	0
DropOut	(None, 64)	0
Dense	(None, 6)	650
Activation	(None, 6)	0
Total Param: 241,174		
Trainable Param: 241, 174		
non-Trainable Param: 0		

Table 33: FDC-2 Model Hyper parameters

Layer(Type)	**Output Shape**	**Param #**
Conv2D	(None, 126, 126, 32)	320
Activation	(None, 126, 126, 32)	0
MaxPooling2D	(None, 31, 63, 32)	0
Conv2D	(None, 29, 61, 64)	18496
Activation	(None, 29, 61, 64)	0
MaxPooling2D	(None, 7, 30, 64)	0
Flatten	(None, 13440)	0
Dropout	(None, 13440)	0
Dense	(None, 64)	860224
Activation	(None, 64)	0
Dropout	(None, 64)	0
Dense	(None, 6)	650
Activation	(None, 6)	0
Total Parameters: 879,430		
Trainable Parameters: 879,430		
Non-Trainable Parameters: 0		

Table 34: FDC-3 Model Hyper-Parameters

Layer (Type)	**Output Shape**	**Param #**
Conv2D	(None, 128, 128, 32)	320
BatchNormalization	(None, 128, 128, 32)	128
Conv2D	(None, 126, 126, 32)	9248
BatchNormalization	(None, 126, 126, 32)	128
MaxPooling2D	(None, 63, 63, 32)	0
Dropout	(None, 63, 63, 32)	0
Conv2D	(None, 63, 63, 32)	18496
BatchNormalization	(None, 63, 63, 32)	256
Conv2D	(None, 61, 61, 64)	36928
BatchNormalization	(None, 61, 61, 64)	256
MaxPooling2D	(None, 31, 31, 64)	0
Dropout	(None, 31, 31, 64)	0
Conv2D	(None, 31, 31, 64)	36928
BatchNormalization	(None, 31, 31, 64)	256
Conv2D	(None, 29, 29, 64)	36928
BatchNormalization	(None, 29, 29, 64)	256
MaxPooling2D	(None, 15, 15, 64)	0
Dropout	(None, 15, 15, 64)	0
Conv2D	(None, 15, 15, 64)	73856
BatchNormalization	(None, 15, 15, 64)	512
Conv2D	(None, 13, 13, 128)	147584
BatchNormalization	(None, 13, 13, 128)	512
MaxPooling2D	(None, 7, 7, 128)	0
Dropout	(None, 7, 7, 128)	0
GlobalAveragePooling	(None, 128)	0
Dense	(None, 6)	1290
Activation	(None, 6)	0
Total Parameters: 363,366		
Trainable Parameters: 362,214		
Non-Trainable Parameters: 1,152		

Table 35: FDC-FS Model Hyper-Parameters (1)

Layer (Type)	**Output Shape**	**Param #**
Input	(None, 128, 128, 1)	0
Conv2D	(None, 128, 128, 32)	320
BatchNormalization	(None, 128, 128, 32)	128
Conv2D	(None, 126, 126, 32)	9248
BatchNormalization	(None, 126, 126, 32)	128
MaxPooling2D	(None, 63, 63, 32)	0
DropOut	(None, 63, 63, 32)	0
Conv2D	(None, 63, 63, 64)	18496
BatchNormalization	(None, 63, 63, 64)	256
Conv2D	(None, 61, 61, 64)	36928
BatchNormalization	(None, 61, 61, 64)	256
MaxPooling2D	(None, 31, 31, 64)	0
DropOut	(None, 31, 31, 64)	0
Conv2D	(None, 124, 124, 24)	624
Conv2D	(None, 31, 31, 64)	36928
MaxPooling2D	(None, 31, 62, 24)	0
BatchNormalization	(None, 31, 31, 64)	256
Activation	(None, 31, 62, 24)	0
Conv2D	(None, 126, 126, 32)	320
Conv2D	(None, 29, 29, 64)	36928
Conv2D	(None, 27, 58, 48)	28848
Activation	(None, 126, 126, 32)	0
BatchNormalization	(None, 29, 29, 64)	256
MaxPooling2D	(None, 6, 29, 48)	0
MaxPooling2D	(None, 31, 63, 32)	0
MaxPooling2D	(None, 15, 15, 64)	0
Activation	(None, 6, 29, 48)	0
Conv2D	(None, 29, 61, 64)	18496
DropOut	(None, 15, 15, 64)	0
Conv2D	(None, 2, 25, 48)	57648
Activation	(None, 29, 61, 64)	0
Conv2D	(None, 15, 15, 128)	73856
Activation	(None, 2, 25, 48)	0
MaxPooling2D	(None, 7, 30, 64)	0
BatchNormalization	(None, 15, 15, 128)	512
Flatten	(None, 2400)	0
Flatten	(None, 13440)	0
Conv2D	(None, 13, 13, 128)	147584
DropOut	(None, 2400)	0
DropOut	(None, 13440)	0
BatchNormalization	(None, 13, 13, 128)	512

Table 36: FDC-FS Model Hyper-Parameters (2) (Continued)

Layer (Type)	**Output Shape**	**Param #**
Dense	(None, 64)	153664
Dense	None, 64)	860224
MaxPooling2D	(None, 7,7, 128)	0
Activation	(None, 64)	0
Activation	(None, 64)	0
DropOut	(None, 7, 7, 128)	0
DropOut	None, 64)	0
DropOut	None, 64)	0
GlobalAveragePoolinh2D	(None, 128)	0
Concatenate	(None, 256)	0
Dense	(None, 6)	1542
Total Parameters: 1,483,958		
Trainable Parameters: 1,482,806		
Non-Trainable Parameters: 1,152		

Table 37: Lung Sound Dataset: Original and Augmented Data

ID	Name of Disease	Ori. Data	Aug. Data
1	Bronchiectasis	29	319
2	COPD	785	8635
3	Health	35	385
4	LRTI	2	22
5	Pneumonia	37	407
6	URTI	31	341
	Total	**919**	**10109**

Table 38: Heart Sound Dataset: Original and Augmented Data

S.No.	Category	Ori. Data	Aug. Data
1	Extra Systole	46	506
2	Normal	200	2200
3	Noisy Normal	120	1320
4	Murmur	66	726
5	Noisy Murmur	29	319
6	Unlabelled Test	195	2145
	Total	**656**	**7216**

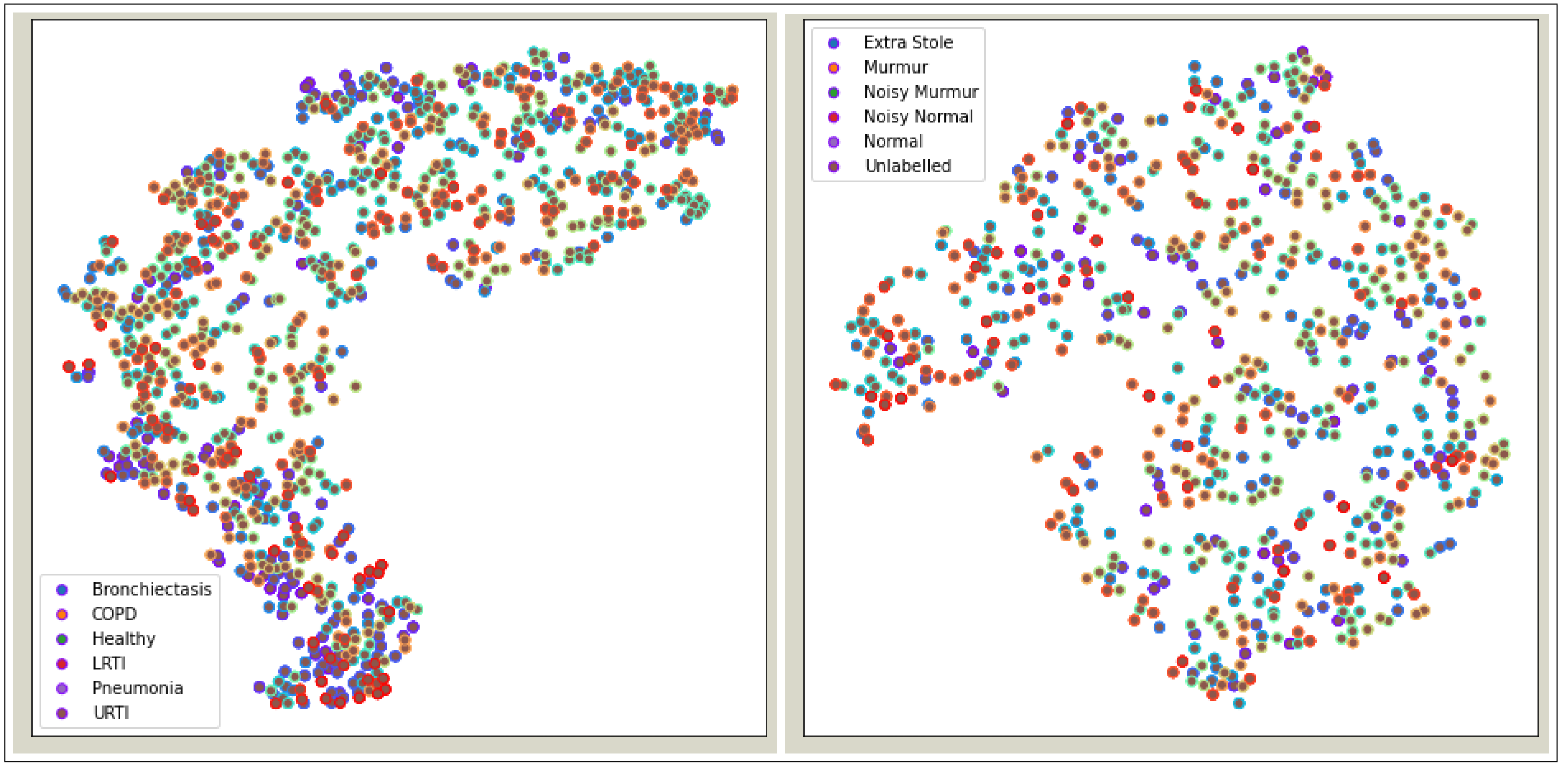


Figure 27: t-SNE Visualization: (a) Lung Sound Dataset (b) Heart Sound Dataset

5.5.3 Classification Results and Evaluations

Results on Original Lung Dataset: Based on the setup explained above, we have obtained the accuracy performance for the four models (FDC-1, FDC-2, FDC-3, and FDC-FS). FDC-FS model obtains the highest accuracy model in all three feature cases. Specifically, the highest accuracy is achieved by Spectrogram 97%, while MFCC reported accuracy of 91% and Chromagram reported accuracy of 95%. The learning and validation graphs for learning and loss are shown in Figure. 28. The class-wise accuracy for all models can be seen in Figure. 32 and Table 41.

Table 39: Lung & Heart Condition Detection with Original Data (Testing Accuracy: Weighted Average)

	Lung Sound Classification				Heart Sound Classification			
Features	FDC-1	FDC-2	FDC-3	FDC-FS	FDC-1	FDC-2	FDC-3	FDC-FS
Spectrogram	91%	87%	84%	**97%**	85%	81%	73.5%	**89%**
MFCC	90%	93%	83%	**91%**	89%	91%	89%	**93%**
Chromagram	86%	83%	89%	**95%**	89%	89%	84%	**92%**

Table 40: Lung & Heart Condition Detection with Augmented Data (Testing Accuracy: Weighted Average)

	Lung Sound Classification				Heart Sound Classification			
Features	FDC-1	FDC-2	FDC-3	FDC-FS	FDC-1	FDC-2	FDC-3	FDC-FS
Spectrogram	99%	98.6%	98.3%	**99.1%**	93%	92%	95.5%	**97%**
MFCC	99.3%	98%	98.2%	**99%**	95%	93%	96%	**96%**
Chromagram	97%	97%	95%	**98.4%**	94%	93%	95%	**97%**

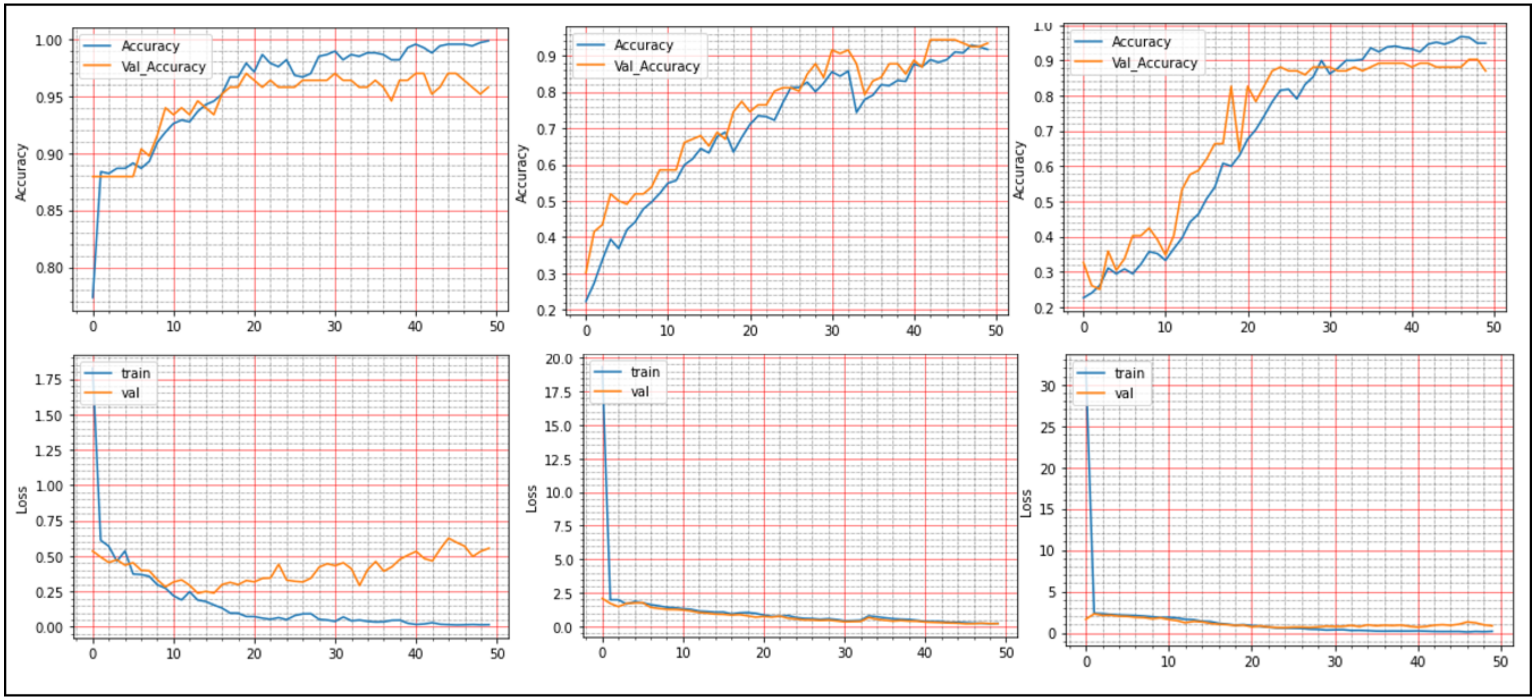


Figure 28: Lung Condition Classification (Accuracy vs. Loss): Fusion Network Model (FDC-FS) with Features: (a) Spectrogram (b) MFCC (c) Chromagram

Table 41: Class Wise Accuracy for Lung Dataset

Model	Feature	Class	Bronchiectasis	COPD	Healthy	LRTI	Pneumonia	URTI	W. AVG
FDC-1	Spec.	Ori.	92%	100%	67%	87%	87%	97%	91%
		Aug.	81 %	100 %	94 %	86 %	100 %	88 %	99%
	MFCC	Ori.	100%	89%	100%	76%	95%	88%	90%
		Aug.	91 %	100 %	95 %	100 %	94 %	91 %	99.3%
	Chroma	Ori.	79%	83%	100%	74%	86%	93%	86%
		Aug.	84 %	100 %	71 %	100 %	81 %	77 %	97%
FDC-2	Spec.	Ori.	73%	78%	62%	92%	84%	95%	87%
		Aug.	91 %	99 %	86 %	100 %	100 %	98 %	98.6%
	MFCC	Ori.	78%	88%	100%	71%	100%	96%	93%
		Aug.	81%	100%	88%	94%	92%	77%	98%
	Chroma	Ori.	100 %	80 %	40%	75%	94%	80%	83%
		Aug.	85 %	100 %	80 %	75 %	88 %	77 %	97%
FDC-3	Spec.	Ori.	86%	64%	100%	79%	100%	83%	84%
		Aug.	94 %	99 %	100 %	100 %	100 %	90 %	98.3%
	MFCC	Ori.	67%	85%	60%	77%	95%	79%	83%
		Aug.	69 %	100 %	59 %	50 %	77 %	74 %	98.2%
	Chroma	Ori.	86%	87%	80%	88%	98%	79%	89%
		Aug.	78 %	97 %	78 %	100 %	82 %	70 %	95%
FDC-FS	Spec.	Ori.	100%	100%	100%	99%	94%	93%	97%
		Aug.	95 %	100 %	90 %	100 %	100 %	96 %	99.1%
	MFCC	Ori.	75%	82%	67%	87%	100%	92%	91%
		Aug.	92 %	100 %	96 %	80 %	93 %	100 %	99%
	Chroma	Ori.	100%	100%	100%	97%	96%	89%	95%
		Aug.	90 %	100 %	88 %	100 %	89 %	91 %	98.4%

Results on Augmented Lung Dataset: We have achieved very impressive accuracy from the experiments with the augmented lung dataset that outperformed all state-of-the-art research. The highest accuracy was conducted by the FDC-FS model, which is the fusion of FDC-1, FDC-2, and FDC-FS. The weighted accuracy for the FDC-FS model for Spectrogram is 99.1%, the accuracy of 99% for MFCC, and 98.4% for Chromagram. Notably, it has been improved approximately 2%, 8%, 3% in Spectrogram, MFCC, and Chromagram, compared to the accuracy performance with the original data. The results for the augmented lung dataset are shown in Table 40. We have also reported the class-wise accuracy for the augmented dataset, which is visually demonstrated in Figure. 32. We obtained the highest class-wise accuracy for the COPD category based on the Spectrogram and MFCC features.

Results on Original Heart Dataset: Our experimental results are based on 50 epochs and 64 batch sizes and the categorical cross-entropy for the data validation. The average times taken for training the model were from 7 seconds to 1 minute for the FDC models. The learning graphs of the heart dataset for the FDC-FS model are shown in Figure. 30. Table 39 offers the accuracy of the heart original dataset. During the training and testing, we observed that FDC-FS performed the best compared to all others; specifically, we obtained the highest accuracy of 93% for MFCC. For the individual feature-model evaluation, the highest accuracy of Spectrogram was 85% with FDC-1; for MFCC, it was 91% with FDC-2. Chromagram was the accuracy of 89% with FDC-1 and FDC-2.

For the class-wise accuracy, the highest accuracy is reported by the FDC-FS model

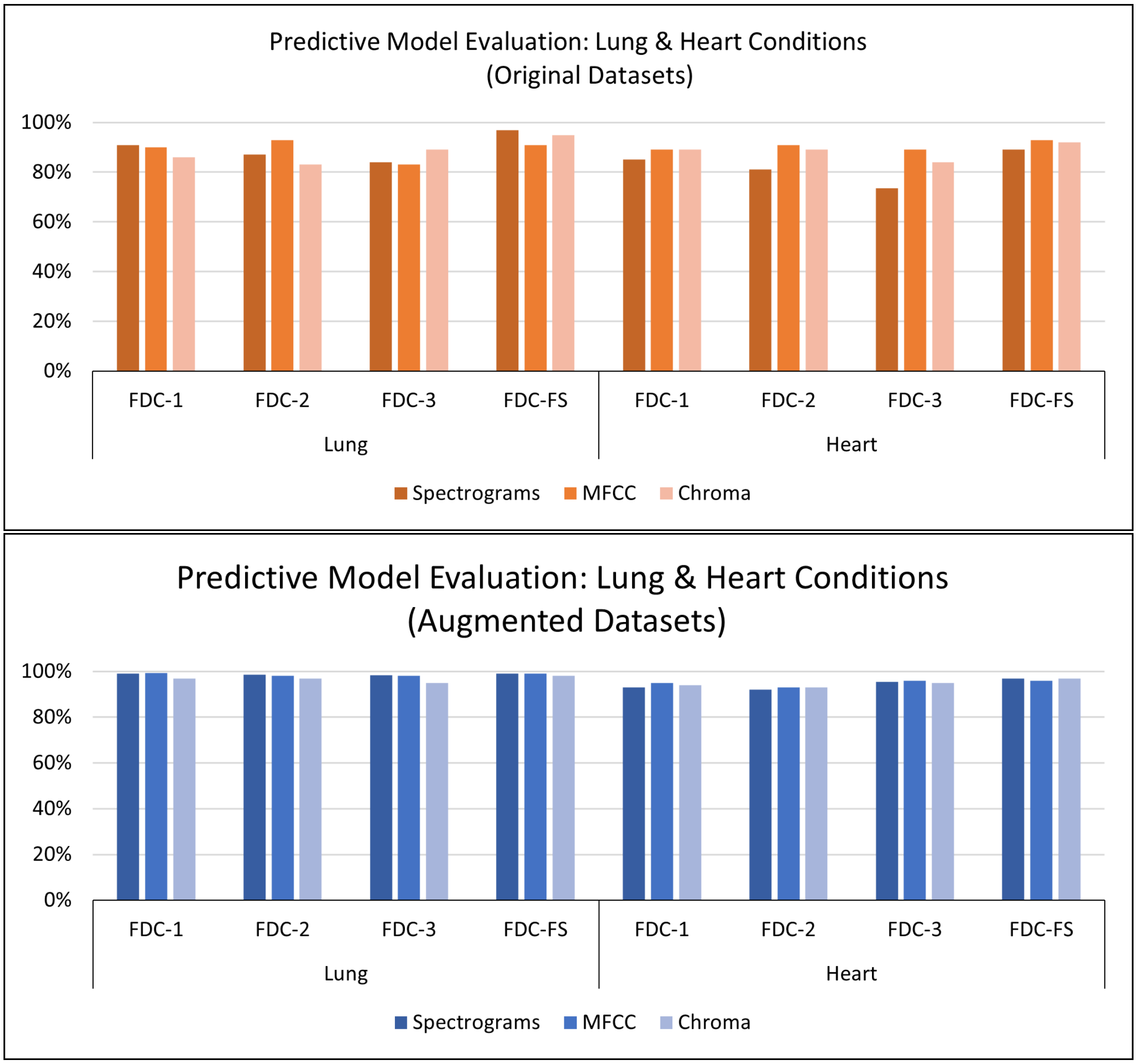


Figure 29: Accuracy for Lung/Heart Condition Detection (Original and Augmented Datasets)

as shown in Figure. 33 and Table 42. The heart data were unbalanced but showed consistent data patterns and characteristics among categories. Thus, as the average accuracy and weighted average accuracy were similar, we constantly reported the weighted accuracy strategy. As the heart data are similar to the musical dataset, MFCC and Chromagram performed better than Spectrogram. On the other hand, Spectrogram performed very well in FDC-FS for both the heart and lung datasets.

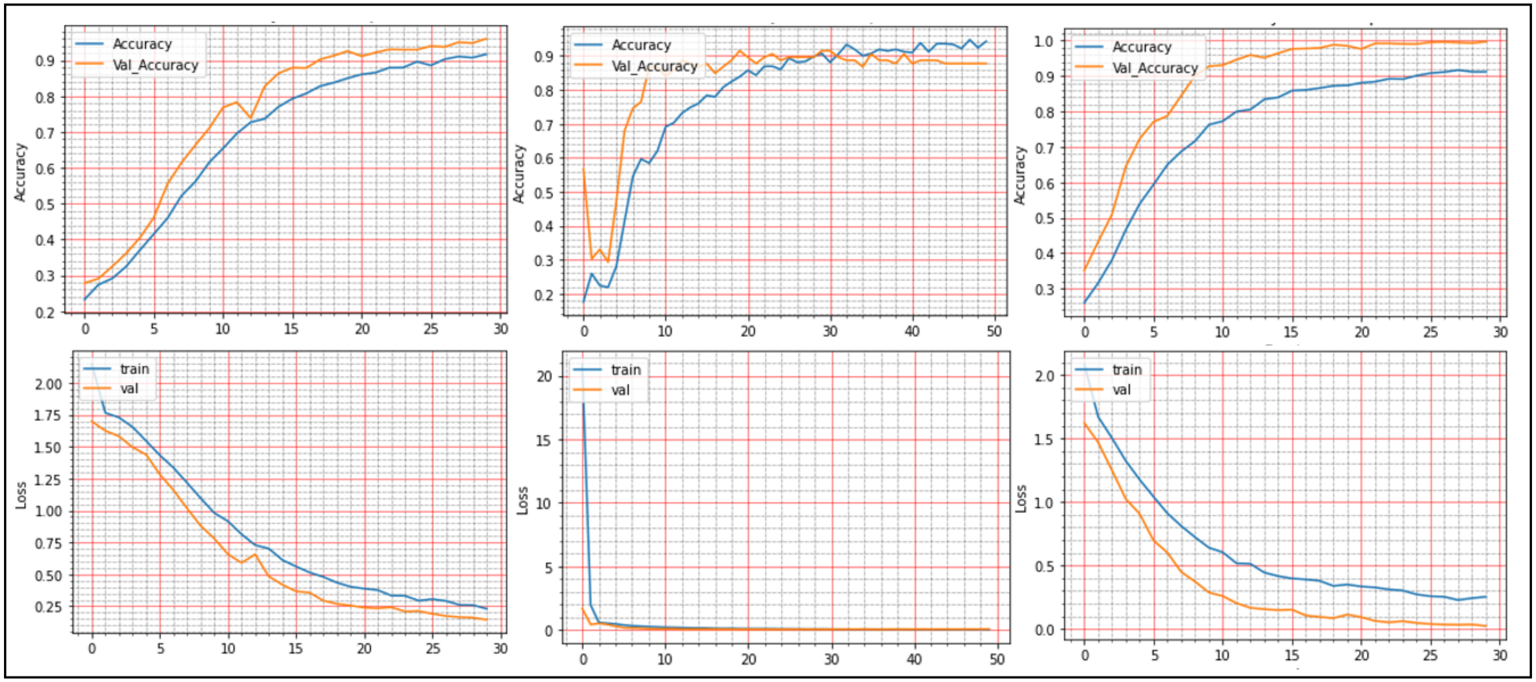


Figure 30: Heart Condition Classification (Accuracy vs. Loss): Fusion Network Model (FDC-FS) with Features: (a) Spectrogram (b) MFCC (c) Chromagram

Table 42: Class Wise Accuracy for Heart Dataset

Model	Feature	Class	Extrastole	Murmur	Noisy Murmur	Noisy Normal	Normal	Unlabeled	W. AVG
FDC-1	Spec.	Ori.	69%	89%	50%	97%	85%	79%	85%
		Aug.	94%	88%	87%	93%	95%	95%	93%
	MFCC	Ori.	83%	79%	83%	75%	98%	87%	89%
		Aug.	99%	93%	96%	96%	98%	90%	95%
	Chroma	Ori.	100%	67%	78%	100%	97%	92%	89%
		Aug.	95%	94%	95%	93%	99%	89%	94%
FDC-2	Spec.	Ori.	75%	55%	75%	90%	89%	76%	81%
		Aug.	93%	91%	89%	94%	93%	91%	92%
	MFCC	Ori.	75%	69%	67%	92%	100%	93%	91%
		Aug.	93%	91%	77%	96%	96%	90%	93%
	Chroma	Ori.	88 %	71 %	86%	78%	93%	92%	89%
		Aug.	91%	90%	83%	96%	95%	90%	93%
FDC-3	Spec.	Ori.	82%	70%	75%	81%	70%	68%	73%
		Aug.	97%	96%	100%	99%	96%	98%	92%
	MFCC	Ori.	67%	62%	88%	96%	98%	82%	89%
		Aug.	95%	96%	100%	96%	97%	95%	96%
	Chroma	Ori.	88%	100%	75%	66%	84%	89%	84%
		Aug.	97%	96%	96%	97%	100%	96%	95%
FDC-FS	Spec.	Ori.	94%	78%	100%	97%	83%	84%	89%
		Aug.	99%	87%	96%	95%	96%	94%	97%
	MFCC	Ori.	67%	82%	100%	92%	96%	95%	93%
		Aug.	97%	94%	96%	97%	98%	92%	96%
	Chroma	Ori.	67%	100%	100%	88%	92%	98%	92%
		Aug.	94%	98%	98%	96%	96%	91%	97%

Results on Augmented Heart Dataset: For the augmented heart data experiments, we observed that FDC-2 obtained the shortest training time of 42 seconds for MFCC, and FDC-FS took the longest time of 4 minutes and 11 seconds for Chromagram. The FDC-FS model obtained the highest accuracy of 97% with Spectrogram and Chromagram and 96% with MFCC. From the individual model evaluation, FDC-3 obtained the highest accuracy of 96% with MFCC. It is because the FDC-3 model is bigger/deeper compared to the other two models. The overall performance of the models is given in Table 40. The class-wise accuracy performance is also shown in Figure. 33 and Table 42. Our results are very competitive even with the state-of-the-art approaches, even with a small and reduced number of measurements.

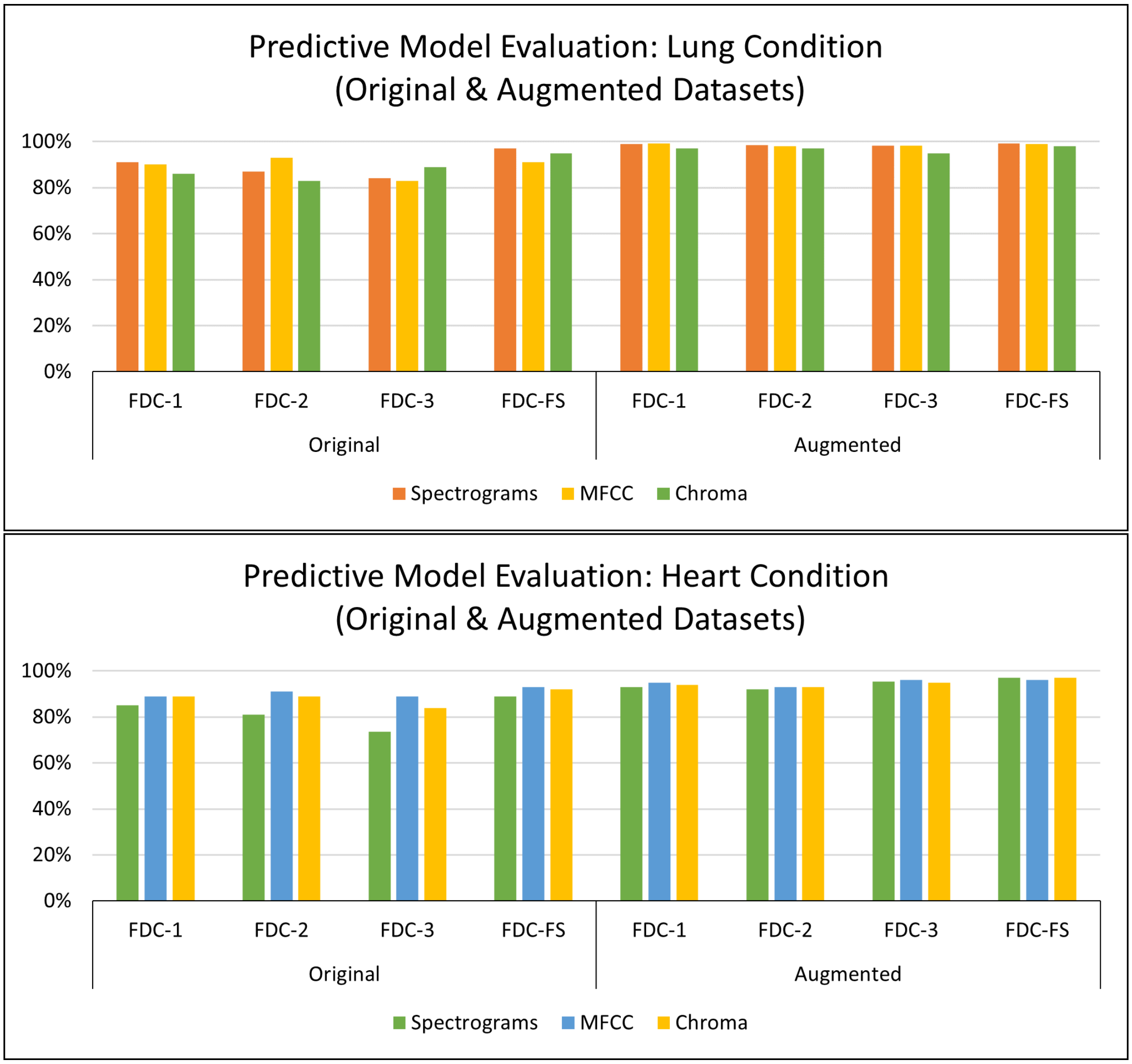


Figure 31: Accuracy for Lung and Heart Condition Detection

Table 43: State-of-the-art Lung Classification Models

Work	Method	Network	Class#	Data	Para#	Aug	Feature	Results
Dalal(2018)[14]	Ensemble	SVM, KNN, GMM, CNN	7	R.A.L.E data	NA	✓	MFCC, LBP	ACC: 95.56%
Hai(2019)[37]	NA	ResNet50	3	489	23M+		Sp	ACC: 98.79%
Fatih(2020)[94]	Ensemble	CNN	4	920	NA		Sp Images, Deep Features	ACC: 71.15%
Demir(2020)[95]	NA	CNN, SVM	4	6,898	138M		Sp	ACC: 65.9%
Demir(2020) [95]	Transfer learning	CNN, SVM	4	6,898	138M		Sp Images	ACC: 63.09%
Samiul(2020)[96]	NA	CNN	6	917	3.8M	✓	Hybrid Scalogram	98.70%
Luay(2021)[98]	Ensemble	SVM, KNN, DT, KNN	6	308/1,176	NA		Entropy features	ACC: 98.20%
FDC-FS (Ours)	Fusion	DCNN	6	Original/Augmented: 919/10,109	**1.48M**	✓	SP, MFCC, CH	**ACC: 99.1%**

Table 44: State-of-the-art Heart Classification Models

Work	Method	Network	Class#	Data	Para#	Aug	Features	Results
Potes(2016)[99]	Ensemble	CNN	2	Normal/Abnormal: 2,575/665	NA		MFCC	ACC: 85%
Zhang(2017)[100]	NA	CNN+SVM	3/4	Heart sounds 1 & 2	NA		SP	Precision: 77%/71%
Bozkurt(2018)[38]	NA	CNN	4	PhysioNet: Abnormal/Normal	NA	✓	MFCC, Mel-SP	ACC: 81.50%
Wu(2019)[107]	Ensemble	CNN	2	Normal/Abnormal: 2,575/665	61M		Sp, Mel-SP, MFCC	ACC: 86%
Shu(2020)[101]	NA	WaveNet	5	1,000	0.32M		Multiple features	Training ACC: 97%
Xiao(2020)[108]	Transition	1D CNN	4	PhysioNet: 3,153	0.19M		Raw signal w/t band filter	ACC: 93%
Muqing(2020)[102]	Concatenation	CRNN, PRCNN	4	PhysioNet: 3,240	NA		MFCC	ACC: 98%
Mehmat(2021)[109]	NA	1D CNN	4	PhysioNet	NA		LBP+LTP	ACC: 91%
FDC-FS (Ours)	Fusion	DCNN	6	Original/Augmented: 656/7,216	**1.48M**	✓	SP, MFCC, CH	**ACC: 97%**

5.6 Comparison with State-of-the-Art Research

For the comparative evaluation of our frameworks, we have considered the state-of-the-art research published in reputed journals and conferences between 2018 and 2021, commonly used benchmark datasets from the ICBHI [24] and the heart challenge [106]. First, we have conducted a comparative evaluation of the proposed framework (FDC) with different lung sound classification approaches [14], [37], [94], [95], [98], [96]. Second, we have conducted a comparative evaluation of the proposed framework (FDC) with different heart sound classification approaches [109], [38], [102], [101], [99], [107], [100], [108].

5.6.1 Lung Sound Classification

A comprehensive evaluation of the lung sound classification models has been conducted regarding feature selection and representation, network architecture design, accuracy, and the number of trainable parameters on the lung and heart sound datasets. The best state-of-the-art approach for lung classification that has obtained the highest accuracy of 98.20% is by [109]. However, they have mixed the ICHBI dataset with their own recorded sounds from a local hospital, and another factor is that they are using shallow learning models. [96] proposed methodology obtained 98.70% accuracy, their number of trainable parameters was 3.8 M. To train their model, they needed more computational power and time. Similarly, [37] used different features using the ResNet-50 model, which is a massive model with over 23 M trainable parameters. It can be seen from Table 43 that our model has shallow trainable parameters, i.e., 1.48 M, which is the lowest as compared to all state-of-the-art research. Thus, it requires very minimum resources (we mainly used CoLab for training and testing) and low epochs of only 50 for training our models. The accuracy performance is also slightly better than other approaches. FDC model also achieved the highest accuracy of approximately 99.1%. The overall comparison of our model performance for the lung sound classification with state-of-the-art research is shown in Table 43.

5.6.2 Heart Sound Classification

Similarly, we have conducted a comprehensive evaluation of the heart sound classification models. Based on the number of parameters for FDC-1, the total trainable parameters are 0.24 M, and we have obtained an accuracy of 93%. In contrast, for FDC-3, we received an accuracy of 96%, and the total parameters are 0.36 M. After applying the fusion technique, the parameters increased to 1.48 M with an accuracy of 97%. However, our accuracy for the fused model is near to the state-of-the-art accuracy (approximately 98%). Still, some of the works have not reported the trainable parameters of their proposed models [102]. Also, they used the dataset having additional samples equally balanced. Therefore, it can be assumed that their trainable parameters may slightly be higher due to their network architecture of paralleling recurrent convolutional neural network (CNN), i.e., input shape, the number of layers, max-pooling, strides, the output classification size, etc. However, our accuracy is comparable to Shuvo et al. [96], whose accuracy is 97% with a model with 0.32 M trainable parameters. The overall comparison of our model performance for the heart sound classification with state-of-the-art research is shown in Table 44.

5.6.3 Discussion

Our proposed framework demonstrated superior performance compared to the state-of-the-art research both in lung and heart sound classification. We summarize the primary reasons it is so well performed in lung or heart condition detection and why it consistently achieves high performance. (i)Selection of compelling audio features to maximize the characteristics of lung or heart sounds. (ii) Application of data augmentation techniques effectively to overcome the audio data issues such as the low quality and unbalanced datasets. (iii) Transformation of the selected audio features (Spectrogram, MFCC, and Chromagram) to visual feature vectors to maximize the learning performance from deep learning. (iv) Design three unique deep neural network models (FDC-1, FDC-2, FDC-3) to discover new image patterns of audio features involved in a specific disease in lung or heart domains. (v) The fusion model (FDC-FS) is based on the transfer learning from the three different models (FDC-1, FDC-2, FDC-3) from three unique features (Spectrogram, MFCC, and Chromagram) in the lung or heart sound domains.

The limitations of the proposed framework are (i) The proposed framework performs well in two domain lung and heart sound domains; however, there is a lack of generalization. Nevertheless, we will investigate well enough to offer scientific evidence to explain why some models or specific features are better than others. (ii) We will develop suitable pattern mining methods and practices for automatic network design according to the given datasets. (iii) We will incorporate more effective transfer learning or subsequent knowledge distillation through the fusion networks that might be further optimized for the excellent balance between conciseness (fusion) and detail (specific features).

5.7 Conclusion

we have developed the feature-based fusion network FDC-FS for the heart and lung disease classification. We used the two publicly available sound datasets with different numbers of samples and class imbalance ratios for this study. In addition, we performed our experimentation with the original dataset to compare our results with the current state-of-the-art research. The experimental results confirmed the superiority of FDC-FS that is a fusion network by combines the three unique models, i.e., FDC-1, FDC-2, and FDC-3, built with the images of specific audio features of Spectrogram, MFCC, and Chromagram. The accuracy reported for the lung dataset is 97% for Spectrogram, 91% for MFCC, and 95% for Chromagram. In contrast, for the heart data, the accuracy reported is 89% for Spectrogram, 93% for MFCC, and 92% for Chromagram.

We have further improved the results by applying the data augmentation techniques to the audio clips rather than images. We used three types of audio augmentation techniques, i.e., noise, pitch shifting, and time stretching, carefully selecting the ranges of values. As a result, the accuracy reported for the augmented lung dataset is 99.1% for Spectrogram, 99% for MFCC, and 98.4% for Chromagram. For the heart dataset, the reported accuracy is based on the accuracy of the dataset% augmentation is 97% for Spectrogram, 96% for MFCC, and 97% for Chromagram. (iv) We will improve preprocessing or data augmentation methods to help to overcome the data issues (noise and

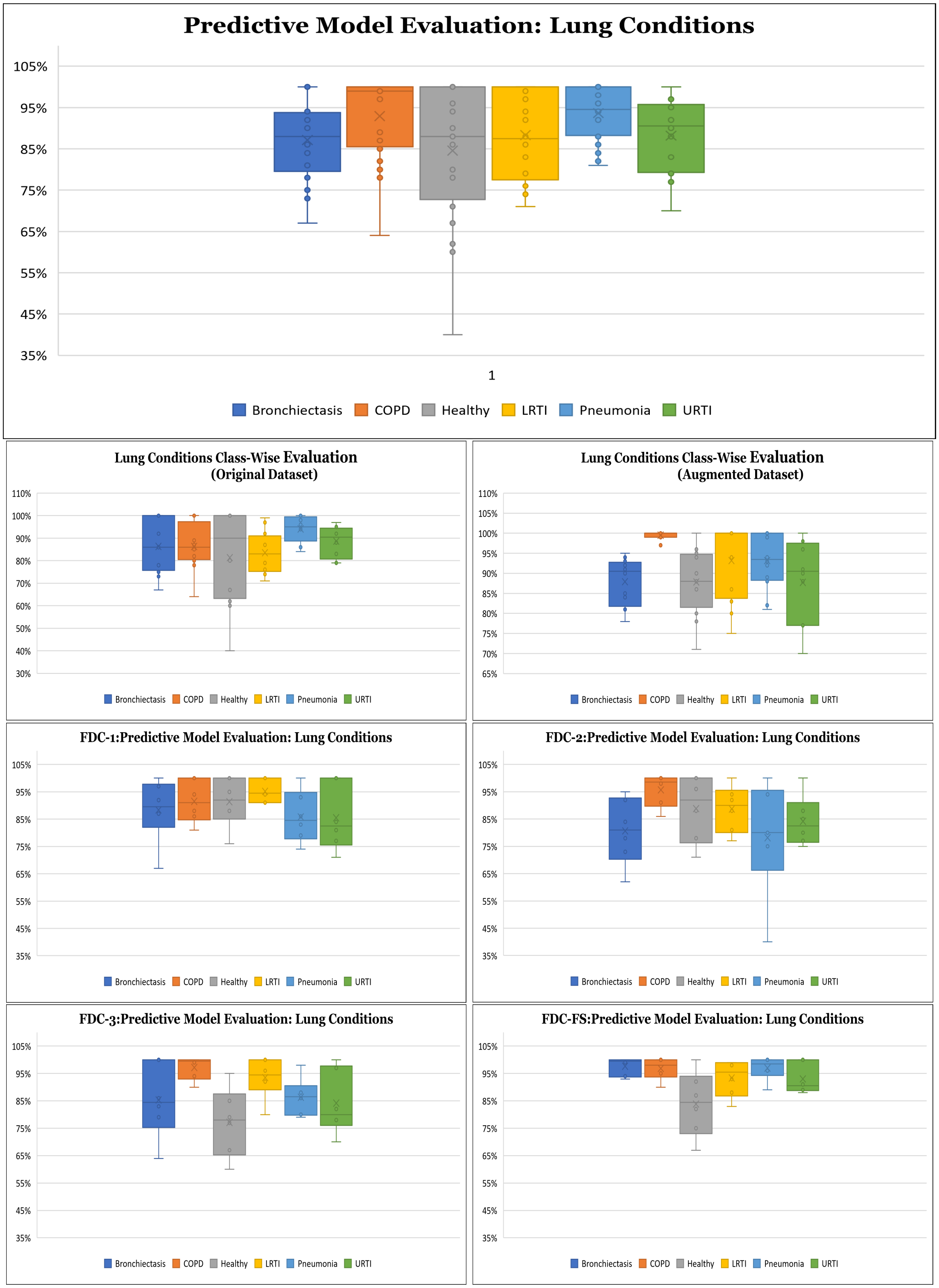
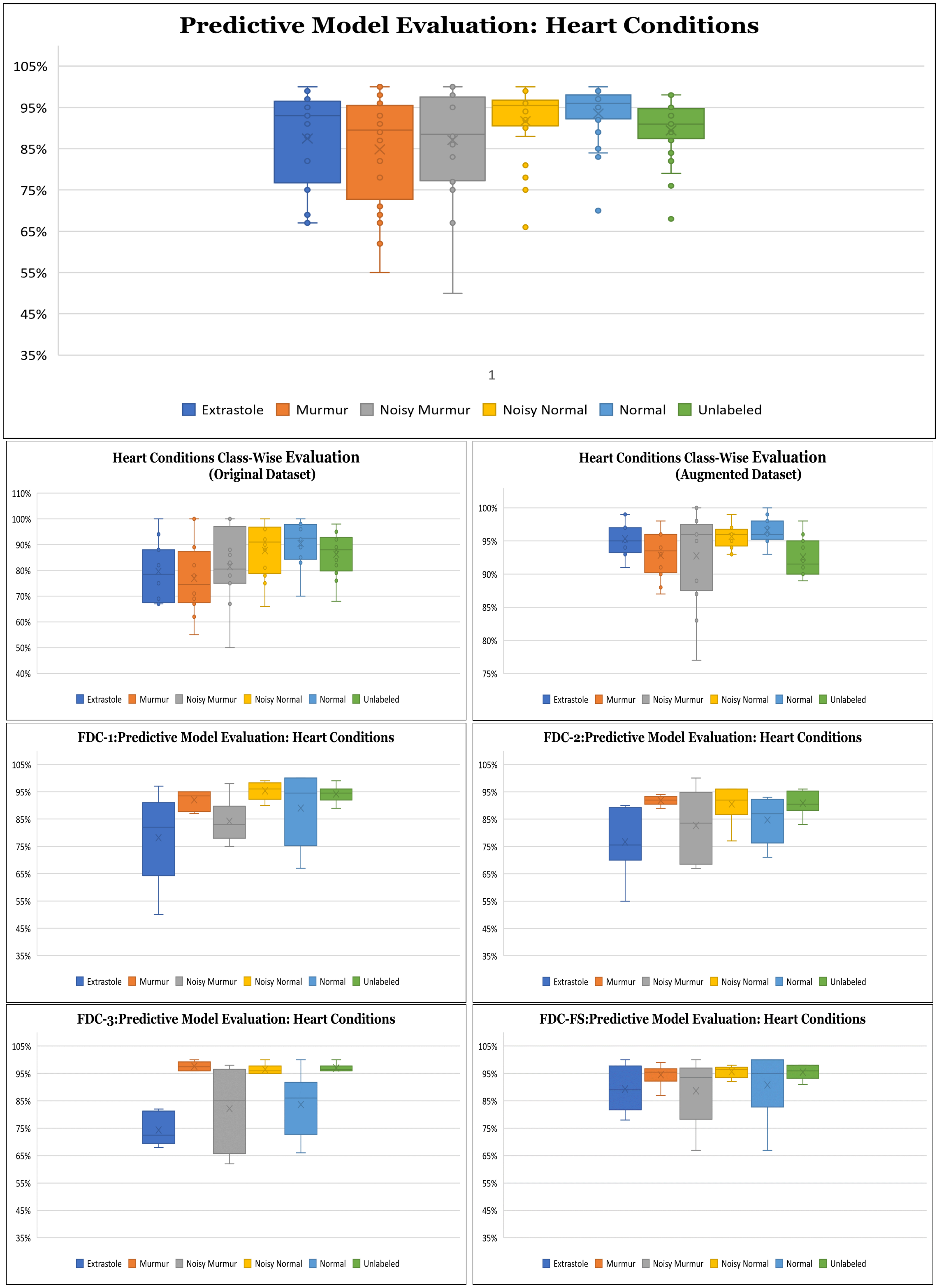


Figure 32: Class Wise Accuracy for Lung Condition Detection



Predictive Model Evaluation: Heart Conditions

Heart Conditions Class-Wise Evaluation (Original Dataset)

Heart Conditions Class-Wise Evaluation (Augmented Dataset)

FDC-1:Predictive Model Evaluation: Heart Conditions

FDC-2:Predictive Model Evaluation: Heart Conditions

FDC-3:Predictive Model Evaluation: Heart Conditions

FDC-FS:Predictive Model Evaluation: Heart Conditions

Figure 33: Class Wise Accuracy for Heart Condition Detection

data imbalance), which are common in medical research, resulting in poor performance and sometimes bias in network design and parameter estimates.

CHAPTER 6

NETWORK MODEL DISCUSSION

Our comparison is based on architecture, the network model, the number of parameters, data augmentation, and features. The comparison is made for the available benchmark environmental and health datasets, i.e., UrbanSound8K, ESC-10 & ESC-50, Heart sounds and Lung sounds. Our results are very competitive with the state-of-the-art approaches discussed earlier, even with a small and reduced number of measurements.

6.1 Multimodal Deep Acoustic (MDA)

The performance with the audio and image data was influenced by data issues such as noise or imbalance. However, we have considered applying well-known data augmentation techniques. The MDA model performs better due to Max pooling's larger size, which is known for downsampling the features and selecting the next layer's essential elements. Secondly, better performance is due to striding, which is a concept that considers data compression, i.e., the data block considered for (2 x 2) in the first layers and (4 x 2) in the next layers is taken as an input. It moves 1 unit ahead and provides the output volume. Hence, reducing the number of parameters from 4 and 8 units in the top two layers to 1 and 1 unit, respectively. The overall results for Multimodal Deep Acoustics are shown in Table 45.

6.2 Network Fusion Model (FDA-NET)

This model is a fusion of FDA-NET-1 and FDA-NET-2 that is concatenated before the final dense layer, depending on the number of classes in the training setâthe last layer of this model uses Softmax activation due to the multiple classification problem. The network was trained with more extensive training parameters with FDA-NET-1 and FDA-NET-2, and their loss and accuracy were computed as the training matrix. As per the working of this network model, features were selected from FDA-NET-1 and FDA-NET-2, and the network was trained with the features, hence, having better parameters, strides, and output data.The overall results for FDA-NET model are shown in Table 46. Table 47 describes the average runtime for environmental sound classification.

6.3 Feature-Fusion Model (FDC-FS)

The FDC-FS model is a fusion model resulting from transfer learning from all three models (FDC-1, FDC-2, FDC-3). Its architecture is composed of the softmax activation and dense layers consisting of the output units equal to the number of classes in the dataset at the last layer. Therefore, FDC-FS is composed of 13 convolutional layers and 3 dense layers model, more specifically three convolutional layers from FDC-1, two convolutional layers from FDC-2, eight convolutional layers from FDC-3, one dense layer from FDC-1, and one dense layer from FDC-2, and a final dense layer of an output unit. The FDC-FS's total trainable parameters for six classes are 1,482,806 (1.4M).The overall results for FDC-FS model are shown in Table 48. Table 49 describes the average runtime for lung and heart sound classification.

Table 45: Overall Results for Multimodal Deep Acoustics

	Lung		Heart		UrbanSound8k	
Mode	Ori.	Aug	Ori.	Aug	Ori.	Aug
Audio	83%	96%	80%	93%	89%	95.80%
Image	85%	97.50%	82%	94%	92%	97%

Table 46: Overall Results for Network Fusion Model (FDA-NET)

	Audio Classification		Image Classification	
Data	Ori.	Aug.	Ori.	Aug.
US-8K	96.19%	96.96%	97%	98.53%
ESC-10	91.25%	93.47%	96.67%	97.50%
ESC-50	87.25%	93.10%	92.75%	96.10%

Table 47: Average Runtime for Environmental Sound Classification

	ESC-10		ESC-50		UrbanSound8k	
	Ori.	Aug.	Ori.	Aug.	Ori.	Aug.
FDA-NET 1	2.1	2.9	2.5	3.01	2.99	3.21
FDA-NET 2	2.6	3.1	2.9	3.23	3.12	3.59
FDA-NET 3	2.9	3.35	3.25	3.59	3.55	3.98

Table 48: Overall Results for Feature-Fusion Model (FDC-FS)

	Lung Sound Classification		Heart Sound Classification	
Features	Ori.	Aug.	Ori.	Aug.
Spectrogram	97%	99.10%	89%	97%
MFCC	91%	99%	93%	96%
Chromagram	95%	98.40%	92%	97%

Table 49: Average Runtime for Lung and Heart Disease Classification

	Lung		Heart	
	Ori.	Aug.	Ori.	Aug.
FDC-1	1.41	1.78	1.32	1.56
FDC-2	1.79	2.45	1.85	2.26
FDC-3	1.65	2.15	1.66	2.01
FDC-FS	2.1	2.96	1.93	2.71

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this research, we propose novel neural network model for deep acoustics, mainly in environment and health domain. The deep acoustics is based on classification of audio and images. We divide our implementation of neural network objective into five main objectives. (1) Data Normalization, (2) Data Augmentation and (3) Feature generation (4) Construction of neural network models (5) Evaluation on bench mark datasets. As a part of our research, we published several papers for classification of audio in a different context during the PhD program [110, 111, 112, 113, 114, 115, 116, 117].

To achieve the objectives, our contributions can be summarized as follows:

7.1 Summary of Deep Acoustics Learning

We have designed an integrated network model with advanced normalization and augmentation techniques for deep acoustic. We named the model as Deep Acoustics (DA) model. The DA model is tested using the three benchmark datasets i.e., lung sounds, heart sounds, and urbansound8k. Our model has shown significant performance interms of accuracy and data insufficient and data imbalance issues for classification of different types of sounds.

7.2 Summary of Multimodality for Deep Acoustics Learning

In addition, for high-performance classification, we developed the Multimodal classification system for deep acoustics (MDA), which combines advanced data normalization and data augmentation approaches. The model has been put to the test in terms of audio and image classification. The audio classification was shown to be a costly method of classification, with the best accuracy for 100 epochs. Image classification, on the other hand, produced the same results in half the time that audio did. In only 50-60 epochs, the image classification achieved the greatest accuracy. To avoid model over-fitting, we employed Early-stopping. In health care, our proposed approach can be utilized to diagnose lung and heart disorders using lung and heart sounds.

7.3 Summary of Network-based Fusion for Deep Acoustics Learning

Our third contribution we have worked on three different datasets using a deep learning approach. We have proposed an enhanced Fusion technique on two individual models, i.e., we have combined i.e., we have combined two networks to produce a third network named FDA-NET-3. The model is based on the fusion of the other two convolutional neural network models. Furthermore, we have considered an audio feature for classification. We further generated spectrogram images for the above datasets in the same format. The performance with the audio and image data was influenced by data issues such as noise or imbalance. However, we have considered applying well-known data augmentation techniques for improving our results. For augmentation of images, we did not consider image augmentation. Instead, we generated spectrograms of augmented audio samples. The presented technique of the data augmentation and normalization provided excellent results. Our validation studies provide fundamental design strategies for improving the classification performance handling the data issues. Our model is also efficient interms of time performance and computational power.

7.4 Summary of Feature-based Fusion for Deep Acoustics Learning

Our main contribution for this objective is to design a feature-based fusion model transferred from the three unique feature-based convolutional neural network models. Our main goal is to show the effectiveness of our model to classify heart or lung diseases with images transformed from three different sound features, i.e., Spectrogram, MFCC, and Chromagram. Furthermore, our objective is to apply different types of data augmentation, such as Noise, Pitch-Shift, and Time-Stretch, effectively to the audio dataset for optimal deep learning training and testing performance.

7.5 Future Work

we can further enhance our results using image augmentation or by using actual data that can be used by deep learning. Transfer learning and handcrafted audio features can help merge different features, improving our results further. Moreover, fusing other datasets would be a good research prospect for our research, i.e., combining audio-based networks with a visual approach.

We will further apply the proposed models and techniques to more various datasets. Moreover, we will take our research towards the multi-tasks classification by combining lung and heart models. Finally, we will extend the work for interpretable deep learning and explainable AI by providing evidence of unique patterns discovered for specific conditions.

Bibliography

[1] Murat Aykanat et al. “Classification of lung sounds using convolutional neural networks”. In: *EURASIP Journal on Image and Video Processing* 2017.1 (2017), p. 65.

[2] Yandre MG Costa, Luiz S Oliveira, and Carlos N Silla Jr. “An evaluation of convolutional neural networks for music classification using spectrograms”. In: *Applied soft computing* 52 (2017), pp. 28–38.

[3] Jongpil Lee et al. “Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification”. In: *Applied Sciences* 8.1 (2018), p. 150.

[4] Justin Salamon and Juan Pablo Bello. “Deep convolutional neural networks and data augmentation for environmental sound classification”. In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 279–283.

[5] Li Deng, Dong Yu, et al. “Deep learning: methods and applications”. In: *Foundations and Trends® in Signal Processing* 7.3–4 (2014), pp. 197–387.

[6] Maryam M Najafabadi et al. “Deep learning applications and challenges in big data analytics”. In: *Journal of big data* 2.1 (2015), pp. 1–21.

[7] Musab Coşkun et al. “An overview of popular deep learning methods”. In: *European Journal of Technique* 7.2 (2017), pp. 165–176.

[8] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. “DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2015, pp. 283–294.

[9] Karol J Piczak. “Environmental sound classification with convolutional neural networks”. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2015, pp. 1–6.

[10] Justin Salamon and Juan Pablo Bello. “Unsupervised feature learning for urban sound classification”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 171–175.

[11] Ilyes Rebai et al. “Improving speech recognition using data augmentation and acoustic model fusion”. In: *Procedia computer science* 112 (2017), pp. 316–322.

[12] Luke Taylor and Geoff Nitschke. “Improving deep learning using generic data augmentation”. In: *arXiv preprint arXiv:1708.06020* (2017).

[13] Hamada Rizk, Ahmed Shokry, and Moustafa Youssef. “Effectiveness of data augmentation in cellular-based localization using deep learning”. In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2019, pp. 1–6.

[14] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. “Lung sounds classification using convolutional neural networks”. In: *Artificial intelligence in medicine* 88 (2018), pp. 58–69.

[15] Tomoya Koike et al. “Audio for audio is better? An investigation on transfer learning models for heart sound classification”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 74–77.

[16] Karol J Piczak. “ESC: Dataset for environmental sound classification”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 1015–1018.

[17] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A dataset and taxonomy for urban sound research”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 1041–1044.

[18] Nithya Davis and K Suresh. “Environmental sound classification using deep convolutional neural networks and data augmentation”. In: *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE. 2018, pp. 41–45.

[19] U Rajendra Acharya et al. “Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals”. In: *Information Sciences* 415 (2017), pp. 190–198.

[20] Shu Lih Oh et al. “Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats”. In: *Computers in biology and medicine* 102 (2018), pp. 278–287.

[21] Wootaek Lim, Daeyoung Jang, and Taejin Lee. “Speech emotion recognition using convolutional and recurrent neural networks”. In: *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE. 2016, pp. 1–4.

[22] Rupesh Dubey and Rajesh M Bodade. “A Review of Classification Techniques Based on Neural Networks for Pulmonary Obstructive Diseases”. In: *A Review of Classification Techniques Based on Neural Networks for Pulmonary Obstructive Diseases (April 1, 2019)* (2019).

[23] Zhichao Zhang et al. “Learning attentive representations for environmental sound classification”. In: *IEEE Access* 7 (2019), pp. 130327–130339.

[24] BM Rocha et al. “A respiratory sound database for the development of automated classification”. In: *International Conference on Biomedical and Health Informatics*. Springer. 2017, pp. 33–37.

[25] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

[26] R EBU–Recommendation. “Loudness normalisation and permitted maximum level of audio signals”. In: (2011).

[27] Rafael L Aguiar, Yandre MG Costa, and Carlos N Silla. “Exploring data augmentation to improve music genre classification with convnets”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–8.

[28] Shengyun Wei et al. “Sample mixed-based data augmentation for domestic audio tagging”. In: *arXiv preprint arXiv:1808.03883* (2018).

[29] Brian McFee et al. “librosa: Audio and music signal analysis in python”. In: *Proceedings of the 14th python in science conference*. 2015, pp. 18–25.

[30] Jeffrey J Ward. *Rale lung sounds 3.1 professional edition*. 2005.

[31] P Bentley et al. “The pascal classifying heart sounds challenge 2011 (chsc2011) results”. In: *See http://www. peterjbentley. com/heartchallenge/index. html* (2011).

[32] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A dataset and taxonomy for urban sound research”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 1041–1044.

[33] Lei Cai, Jingyang Gao, and Di Zhao. “A review of the application of deep learning in medical image classification and segmentation”. In: *Annals of translational medicine* 8.11 (2020).

[34] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. “Deep learning for medical image processing: Overview, challenges and the future”. In: *Classification in BioApps* (2018), pp. 323–350.

[35] Basil M Harris et al. *Non-invasive system and method for breath sound analysis*. US Patent App. 16/465,353. Dec. 2019.

[36] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[37] Hai Chen et al. “Triple-classification of respiratory sounds using optimized s-transform and deep residual networks”. In: *IEEE Access* 7 (2019), pp. 32845–32852.

[38] Baris Bozkurt, Ioannis Germanakis, and Yannis Stylianou. “A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection”. In: *Computers in biology and medicine* 100 (2018), pp. 132–143.

[39] Gari D Clifford et al. “Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016”. In: *2016 Computing in cardiology conference (CinC)*. IEEE. 2016, pp. 609–612.

[40] Rahatul Jannat et al. “Ubiquitous emotion recognition using audio and video data”. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM. 2018, pp. 956–959.

[41] Fatih Demir, Daban Abdulsalam Abdullah, and Abdulkadir Sengur. “A New Deep CNN Model for Environmental Sound Classification”. In: *IEEE Access* 8 (2020), pp. 66529–66537.

[42] Wenhao Bian et al. “Audio-Based Music Classification with DenseNet and Data Augmentation”. In: *Pacific Rim International Conference on Artificial Intelligence*. Springer. 2019, pp. 56–65.

[43] Ohini Kafui Toffa and Max Mignotte. “Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration”. In: *IEEE Transactions on Multimedia* (2020).

[44] Leon Cohen. *Time-frequency analysis*. Vol. 778. Prentice hall, 1995.

[45] John L Semmlow and Benjamin Griffel. *Biosignal and medical image processing*. CRC press, 2014.

[46] K Yamini et al. “Image Colorization With Deep Convolutional Open CV”. In: *Journal of Engineering Science* 11.4 (2020), pp. 533–543.

[47] Yong Xu et al. “Large-scale weakly supervised audio classification using gated convolutional neural network”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 121–125.

[48] Juncheng Li et al. “A comparison of deep learning methods for environmental sound detection”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 126–130.

[49] Jürgen T Geiger and Karim Helwani. “Improving event detection for audio surveillance using gabor filterbank features”. In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE. 2015, pp. 714–718.

[50] Xuehao Liu, Sarah Jane Delany, and Susan McKeever. “Sound Transformation: Applying Image Neural Style Transfer Networks to Audio Spectograms”. In: *International Conference on Computer Analysis of Images and Patterns*. Springer. 2019, pp. 330–341.

[51] Sajjad Abdoli, Patrick Cardinal, and Alessandro Lameiras Koerich. “End-to-end environmental sound classification using a 1D convolutional neural network”. In: *Expert Systems with Applications* 136 (2019), pp. 252–263.

[52] Yuni Zeng et al. “Spectrogram based multi-task audio classification”. In: *Multimedia Tools and Applications* 78.3 (2019), pp. 3705–3722.

[53] Koji Abe et al. “Sound classification for hearing aids using time-frequency images”. In: *Proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. IEEE. 2011, pp. 719–724.

[54] Aditya Khamparia et al. “Sound classification using convolutional neural network and tensor deep stacking network”. In: *IEEE Access* 7 (2019), pp. 7717–7727.

[55] Zhao Ren et al. “Learning image-based representations for heart sound classification”. In: *Proceedings of the 2018 International Conference on Digital Health*. 2018, pp. 143–147.

[56] Ossama Abdel-Hamid et al. “Convolutional neural networks for speech recognition”. In: *IEEE/ACM Transactions on audio, speech, and language processing* 22.10 (2014), pp. 1533–1545.

[57] Jonathan Dennis, Huy Dat Tran, and Eng Siong Chng. “Overlapping sound event recognition using local spectrogram features and the generalised hough transform”. In: *Pattern Recognition Letters* 34.9 (2013), pp. 1085–1093.

[58] Nils Günther Peters. *Normalization of ambient higher order ambisonic audio data*. US Patent 9,875,745. Jan. 2018.

[59] Loris Nanni et al. “An Ensemble of Convolutional Neural Networks for Audio Classification”. In: *arXiv preprint arXiv:2007.07966* (2020).

[60] Shaobo Li et al. “An ensemble stacked convolutional neural network model for environmental event sound recognition”. In: *Applied Sciences* 8.7 (2018), p. 1152.

[61] Jederson S Luz et al. “Ensemble of handcrafted and deep features for urban sound classification”. In: *Applied Acoustics* 175 (2021), p. 107819.

[62] Michael Glodek et al. “Multiple classifier systems for the classification of audio-visual emotional states”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer. 2011, pp. 359–368.

[63] Stavros Petridis and Maja Pantic. “Prediction-based audiovisual fusion for classification of non-linguistic vocalisations”. In: *IEEE Transactions on Affective Computing* 7.1 (2015), pp. 45–58.

[64] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. “Rethinking CNN Models for Audio Classification”. In: *arXiv preprint arXiv:2007.11154* (2020).

[65] Zohaib Mushtaq, Shun-Feng Su, and Quoc-Viet Tran. “Spectral images based environmental sound classification using CNN with meaningful data augmentation”. In: *Applied Acoustics* 172 (2020), p. 107581.

[66] Zohaib Mushtaq and Shun-Feng Su. “Environmental sound classification using a regularized deep convolutional neural network with data augmentation”. In: *Applied Acoustics* 167 (2020), p. 107389.

[67] Venkatesh Boddapati et al. “Classifying environmental sounds using image recognition networks”. In: *Procedia computer science* 112 (2017), pp. 2048–2056.

[68] Keunwoo Choi et al. “Convolutional recurrent neural networks for music classification”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 2392–2396.

[69] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. “Recurrent neural networks for polyphonic sound event detection in real life recordings”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 6440–6444.

[70] Yasser Alsouda, Sabri Pllana, and Arianit Kurti. “IoT-based Urban Noise Identification Using Machine Learning: Performance of SVM, KNN, Bagging, and Random Forest”. In: *Proceedings of the International Conference on Omni-Layer Intelligent Systems*. 2019, pp. 62–67.

[71] Zhichao Zhang et al. “Deep convolutional neural network with mixup for environmental sound classification”. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer. 2018, pp. 356–367.

[72] Lonce Wyse. “Audio spectrogram representations for processing with convolutional neural networks”. In: *arXiv preprint arXiv:1706.09559* (2017).

[73] L Rafael Aguiar, MG Yandre Costa, and N Carlos Silla. “Exploring Data Augmentation to Improve Music Genre Classification with ConvNets”. In: *International Joint Conference on Neural Networks*. IEEE. 2018, pp. 1–8.

[74] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. “Harmonics tracking and pitch extraction based on instantaneous frequency”. In: *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. Vol. 1. IEEE. 1995, pp. 756–759.

[75] Golnooshsadat Elhami and Romann M Weber. “Audio feature extraction with convolutional neural autoencoders with application to voice conversion”. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. CONF. 2019.

[76] Gustavo Z Felipe et al. “Identification of infantsâ cry motivation using spectrograms”. In: *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE. 2019, pp. 181–186.

[77] Mohan Mishra et al. “Classification of normal and abnormal heart sounds for automatic diagnosis”. In: *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE. 2017, pp. 753–757.

[78] Chloë Brown et al. “Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 3474–3484.

[79] Diogo Marcelo Nogueira et al. “Heart sounds classification using images from wavelet transformation”. In: *EPIA Conference on Artificial Intelligence*. Springer. 2019, pp. 311–322.

[80] M Cobos, JJ Perez-Solano, and LT Berger. “Acoustic-based technologies for ambient assisted living”. In: *Introduction to Smart eHealth and eCare Technologies* (2016), pp. 159–180.

[81] Charalampos Doukas and Ilias Maglogiannis. “Advanced patient or elder fall detection based on movement and sound data”. In: *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*. IEEE. 2008, pp. 103–107.

[82] Sarika Hegde et al. “A survey on machine learning approaches for automatic detection of voice disorders”. In: *Journal of Voice* 33.6 (2019), 947–e11.

[83] Amit Krishna Dwivedi, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. “Algorithms for automatic analysis and classification of heart sounds–a systematic review”. In: *IEEE Access* 7 (2018), pp. 8316–8345.

[84] Musaed Alhussein, Ghulam Muhammad, and M Shamim Hossain. “EEG pathology detection based on deep learning”. In: *IEEE Access* 7 (2019), pp. 27781–27788.

[85] Shaikh Anowarul Fattah et al. “Stetho-phone: Low-cost digital stethoscope for remote personalized healthcare”. In: *2017 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE. 2017, pp. 1–7.

[86] K Davis. “Annual Report: Presidentâs MessageâHealth Care Reform: A Journey”. In: *Commonwealth Fund, New York* (2012).

[87] Yasemin P Kahya, E Cagatay Guler, and Serdar Sahin. “Respiratory disease diagnosis using lung sounds”. In: *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.'Magnificent Milestones and Emerging Opportunities in Medical Engineering'(Cat. No. 97CH36136)*. Vol. 5. IEEE. 1997, pp. 2051–2053.

[88] Anup Mandke and Kshitij Mandke. *Under diagnosis of COPD in primary care setting in Surat, India*. 2015.

[89] Salvatore Mangione and Linda Z Nieman. “Pulmonary auscultatory skills during training in internal medicine and family practice”. In: *American journal of respiratory and critical care medicine* 159.4 (1999), pp. 1119–1124.

[90] Agnieszka Mikołajczyk and Michał Grochowski. “Data augmentation for improving deep learning in image classification problem”. In: *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE. 2018, pp. 117–122.

[91] Truc Nguyen and Franz Pernkopf. “Lung Sound Classification Using Snapshot Ensemble of Convolutional Neural Networks”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 760–763.

[92] Kranthi Kumar Lella and Alphonse Pja. “Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice”. In: *AIMS Public Health* 8.2 (2021), p. 240.

[93] Kirill Kochetov and Andrey Filchenkov. “Generative Adversarial Networks for Respiratory Sound Augmentation”. In: *2020 International Conference on Control, Robotics and Intelligent System*. 2020, pp. 106–111.

[94] Fatih Demir, Aras Masood Ismael, and Abdulkadir Sengur. “Classification of Lung Sounds With CNN Model Using Parallel Pooling Structure”. In: *IEEE Access* 8 (2020), pp. 105376–105383.

[95] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj. “Convolutional neural networks based efficient approach for classification of lung diseases”. In: *Health information science and systems* 8.1 (2020), pp. 1–8.

[96] Samiul Based Shuvo et al. “A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram”. In: *IEEE Journal of Biomedical and Health Informatics* (2020).

[97] Elmar Messner et al. “Multi-channel lung sound classification with convolutional recurrent neural networks”. In: *Computers in Biology and Medicine* 122 (2020), p. 103831.

[98] Luay Fraiwan et al. “Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers”. In: *Biocybernetics and Biomedical Engineering* 41.1 (2021), pp. 1–14.

[99] Cristhian Potes et al. “Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds”. In: *2016 computing in cardiology conference (CinC)*. IEEE. 2016, pp. 621–624.

[100] Wenjie Zhang and Jiqing Han. “Towards heart sound classification without segmentation using convolutional neural network”. In: *2017 Computing in Cardiology (CinC)*. IEEE. 2017, pp. 1–4.

[101] Shu Lih Oh et al. “Classification of heart sound signals using a novel deep wavenet model”. In: *Computer Methods and Programs in Biomedicine* 196 (2020), p. 105604.

[102] Muqing Deng et al. “Heart sound classification based on improved MFCC features and convolutional recurrent neural networks”. In: *Neural Networks* 130 (2020), pp. 22–32.

[103] Pranav Rajpurkar et al. “Cardiologist-level arrhythmia detection with convolutional neural networks”. In: *arXiv preprint arXiv:1707.01836* (2017).

[104] Brian McFee, Eric J Humphrey, and Juan Pablo Bello. “A software framework for musical data augmentation.” In: *ISMIR*. Vol. 2015. 2015, pp. 248–254.

[105] Sirko Molau et al. “Computing mel-frequency cepstral coefficients on the power spectrum”. In: *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221)*. Vol. 1. IEEE. 2001, pp. 73–76.

[106] P. Bentley et al. *The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results*. http://www.peterjbentley.com/heartchallenge/index.html.

[107] Jimmy Ming-Tai Wu et al. “Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model”. In: *Applied Soft Computing* 78 (2019), pp. 29–40.

[108] Bin Xiao et al. “Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption”. In: *Neurocomputing* 392 (2020), pp. 153–159.

[109] ER Mehmet Bilal. “Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features”. In: *Applied Acoustics* 180 (2021), p. 108152.

[110] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. “Smart 311 request system with automatic noise detection for safe neighborhood”. In: *2018 IEEE International Smart Cities Conference (ISC2)*. IEEE. 2018, pp. 1–8.

[111] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. “Lung Disease Classification using Deep Convolutional Neural Network”. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 732–735.

[112] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. “Speech Emotion Detection using IoT based Deep Learning for Health Care”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 4191–4196.

[113] Sayed Khushal Shah, Zeenat Tariq, and Yugyung Lee. “Audio iot analytics for home automation safety”. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 5181–5186.

[114] Sayed Khushal Shah, Zeenat Tariq, and Yugyung Lee. “IoT based Urban Noise Monitoring in Deep Learning using Historical Reports”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 4179–4184.

[115] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. “Multimodal Lung Disease Classification using Deep Convolutional Neural Network”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 2530–2537.

[116] Sayed Khushal Shah et al. “Real-Time Machine Learning for Air Quality and Environmental Noise Detection”. In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 3506–3515.

[117] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. “Automatic Multimodal Heart Disease Classification using Phonocardiogram Signal”. In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 3514–3521.

VITA

Zeenat Tariq is a doctoral candidate in computer science at the School of Computing and Engineering, University of Missouri-Kansas City. Her research is supervised by Prof. Yugyung Lee. Her dissertation research focuses on developing novel integrated neural network models for deep acoustics learning. She has published several conference papers in top computer science conferences such as IEEE Big Data, IEEE BIBM, and IEEE Smart city and has contributed to three journal papers in the environment and health acoustics and edge intelligence.

Along with her Ph.D. program, Zeenat is actively involved in teaching and curriculum design before joining UMKC and during her time at UMKC. Zeenat has participated in data science education through OCEL.AI, sponsored by the National Science Foundation-funded. She was also a part of the UMKC Institute for Women summer school in 2019. Zeenat is the Graduate Assistant Fund (GAF) recipient with the highest award in three consecutive years (2019, 2020, and 2021). She started her academic career as an Assistant Professor in the computer science department at the University of North Texas.