Spring 5-23-2021

Multi-Label/Multi-Class Deep Learning Classification of Spatiotemporal Data

Natalie Sommer
Syracuse University

Abstract

Human senses allow for the detection of simultaneous changes in our environments. An unobstructed field of view allows us to notice concurrent variations in different parts of what we are looking at. For example, when playing a video game, a player, oftentimes, needs to be aware of what is happening in the entire scene. Likewise, our hearing makes us aware of various simultaneous sounds occurring around us. Human perception can be affected by the cognitive ability of the brain and acuity of the senses. This is not a factor with machines. As long as a system is given a signal and instructed how to analyze this signal and extract useful information, it will be able to complete this task repeatedly with enough processing power.

Automated and simultaneous detection of activity in machine learning requires the use of multi-labels. In order to detect concurrent occurrences spatially, the labels should represent the regions of interest for a particular application. For example, in this thesis, the regions of interest will be either different quadrants of a parking lot as captured on surveillance videos, four auscultation sites on patients' lungs, or the two sides of the brain's motor cortex (left and right). Since the labels, within the multi-labels, will be used to represent not only certain spatial locations but also different levels or types of occurrences, a multi-class/multi-level schema is necessary. In the first study, each label is appointed one of three levels of activity within the specific quadrant. In the second study, each label is assigned one of four different types of respiratory sounds. In the third study, each label is designated one of three different finger tapping frequencies.

This novel multi-labeling/multi-class schema is one part of being able to detect useful information in the data. The other part of the process lies in the machine learning algorithm, the network model. In order to be able to capture the spatiotemporal characteristics of the

data, selecting Convolutional Neural Network and Long Short Term Memory Network-based algorithms as the basis of the network is fitting.

The following classifications are described in this thesis:

- In the first study, one of three different motion densities are identified simultaneously in four quadrants of two sets of surveillance videos. Publicly available video recordings are the spatiotemporal data.

- In the second study, one of four types of breathing sounds are classified simultaneously in four auscultation sites. The spatiotemporal data are publicly available respiratory sound recordings.

- In the third study, one of three finger tapping rates are detected simultaneously in two regions of interest, the right and left sides of the brain's motor cortex. The spatiotemporal data are fNIRS channel readings gathered during an index finger tapping experiment.

Classification results are based on testing data which is not part of model training and validation. The success of the results is based on measures of Hamming Loss and Subset Accuracy as well Accuracy, F-Score, Sensitivity, and Specificity metrics. In the last study, model explanation is performed using Shapley Additive Explanation (SHAP) values and plotting them on an image-like background, a representation of the fNIRS channel layout used as data input. Overall, promising findings support the use of this approach in classifying spatiotemporal data with the interest of detecting different levels or types of occurrences simultaneously in several regions of interest.

MULTI-LABEL/MULTI-CLASS DEEP LEARNING CLASSIFICATION OF SPATIOTEMPORAL DATA

By

Natalie Sommer

M.S.E.E., Union College, Schenectady, NY , USA, 1991

B.S.E.E., Union College, Schenectady, NY, USA, 1991

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering.

Syracuse University

May 2021

DEDICATION

I dedicate my work to God and my family.

Acknowledgements

- Syracuse University: First, I would like to thank my advisors, Dr. Senem Velipasalar-Gursoy and Dr. Leanne Hirshfield, for their exceptional help and guidance. I would also like to dedicate this dissertation in part to the memory of Dr. Carlos Hartmann who helped me get acclimated back to my doctoral studies and finish my degree. Also, thank you to Danushka Bandara, Yantao Lu, and Burak Kakillioglu for their assistance and support. Finally, I would like to thank my doctoral defense committee members, Dr. James H. Henderson, Dr. Garrett E. Katz, Dr. Fanxin Kong, Dr. Qinru Qiu, and Dr. Reza Zafarani.

- Family: First, and foremost, I thank my husband, Andrei, for his tremendous support along with my three sons, Sebastian, Gabriel, and Phillip, and my daughter-in-law, Ruby. Likewise, I am fortunate to have had my parents' and sister's unending support. Wanting to follow in my father's (Michael Rudko '74) footsteps has pushed me to finally complete my studies. I would also like to dedicate this dissertation in part to the loving memory of my grandparents.

- DeVry College of New York: I would like to thank my colleagues at my workplace for their support along with my students. My students, many of whom were non-traditional college students, taught me that all educational goals can be achieved through grit and perseverance. I often thought of them as I resumed my doctoral studies which I had originally begun in 1991.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Initially, an interest in researching machine learning as it relates to the classification of human emotions launched the exploration of different methods of achieving this. Two types of data, video and Functional Near Infrared Spectroscopy (fNIRS) were considered. For the former, a webcam was used to acquire human reaction data through the acquisition of facial images in response to auditory stimuli [1]. The features of interest in this research were changes in pupil size, mouth curvature, eyebrow curvature, and distance between eyebrows. These types of variances are often linked to a person's emotional reaction to a stimulus. For example, the Autonomous Nervous System (ANS), which controls pupil dilation and constriction, is also linked to our sense of emotion. This means that a change in emotion can trigger a change in pupil size. Likewise, changes in other facial features in response to an emotion-evoking stimulus is often instinctive and worth noting.

Using the International Affective Digital Sounds (IADS) database [9], sounds which elicit various levels of emotion on the valence and arousal scales were chosen. Sounds rated to evoke positive, neutral, and negative emotions were played while subjects were to look at a webcam and react naturally to the sounds. The video recordings were analyzed and various image processing techniques were applied, including the use of wavelets in the extraction of pupil size, and parameters to analyze facial expressions. Changes in the shape

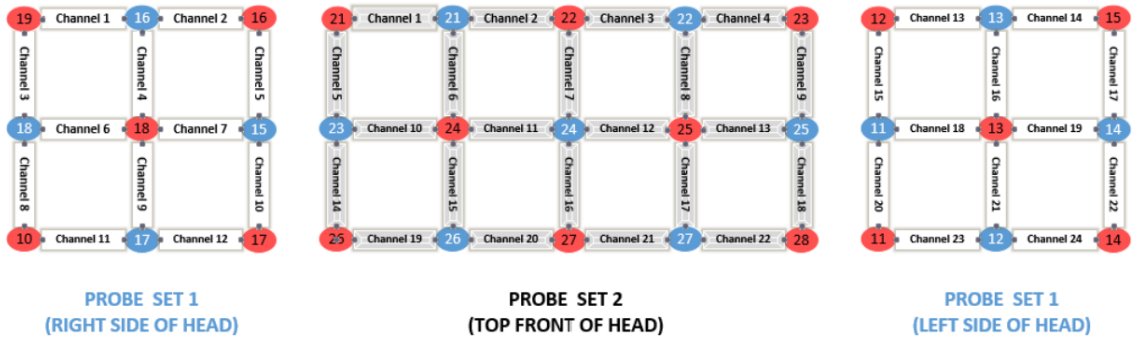
of eyebrows, pupil sizes, and mouth, as well as distances between eyebrows were the facial parameters of interest. A classifier was built based on a simple neural network that was trained on the aforementioned facial and pupil measurement features to detect one of three valence levels of emotion evoked by auditory stimuli. A testing result of 90% was achieved when discerning unpleasant, neutral, and pleasant sounds in this small research.

The exploration of Human Computer Interaction (HCI) as it pertains to the detection of emotions continued with research based on fNIRS data collected during a study of how advertisements affect us. Four groups of emotions were constructed to correspond to three main categories of advertisements. The groupings were made to represent: (i) highly engaging advertisement which could elicit two groups of emotions: 'Excited Happiness' along with 'Pleased & Content', (ii) neutral advertisement which would most likely elicit 'Neutral' emotions, and (iii) badly perceived advertisement which could bring about 'Displeased' emotions.

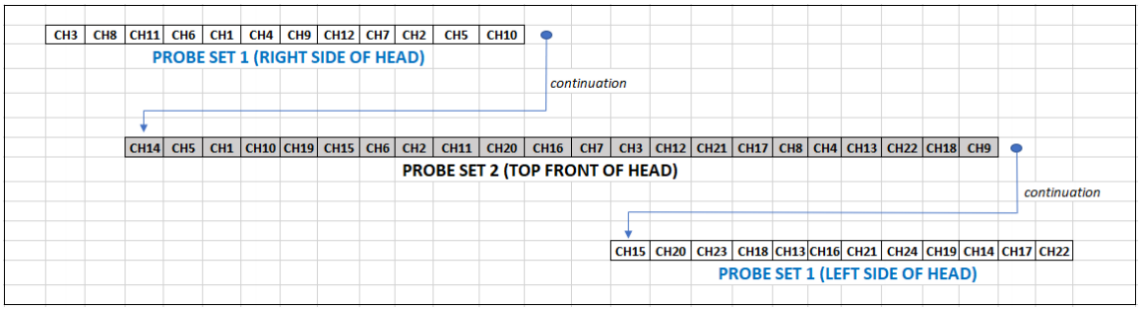
This research was able to support improvement in binary classification results of fNIRS data with a model based on Convolutional Neural Networks (CNNs) as opposed to shallow learning algorithms such as Naïve Bayes and Support Vector Machine. Having gained significant success and popularity in image processing, CNNs have also been used in building models which are trained on fNIRS brain signals. For example, CNNs have been used for gender classification, location of Regions of Interest (ROI), and for the classification of different BCI cognitive tasks [2], [3], [4], [5]. Finally, Bandara et al. [6] used CNNs to classify the three main emotional responses (negative, neutral, positive) on the valence scale due to music video stimuli captured by fNIRS experimental data. With these examples, it is apparent that great interest lies in finding a standard in classifying fNIRS signals with deep learning. Deep learning is attractive in that it does not require the generation of handcrafted features and can, therefore, handle raw data. In the case of the CNN algorithm, one of its advantages is its ability to capture the spatial attributes of input data.

The data for the advertisement study was collected by an ETG-4000 Hitachi fNIRS ma-

chine and was based on a dual probe system which provided 46 channels of oxyhemoglobin (oxyHb) data and 46 channels of deoxyhemoglobin (deoxyHb) data. Neural activity is correlated with an increase in cerebral blood flow which results in an increase in oxyHb (when hemoglobin is transporting oxygen) and decrease in deoxyHb (when hemoglobin is releasing oxygen). Since the absorption spectrum of NIR light is dependent on oxygenation levels, a reaction to a stimulus can be recorded. Labeling was based on each volunteer's Self-Assessment Manikin (SAM) survey results. The order of the channels was not changed in the first dataset and then, for comparison, was adapted to reflect the fNIRS probe layout for the second dataset as shown in Fig. 1.1



(a)



(b)

Figure 1.1: Spatial Channel Order Rearrangement for Advertisement Study
(a)Actual Probe Layout; (b)Final fNIRS Data Channel Set-Up for Dataset #2

A CNN model was trained with two convolutional layers applied to a randomly selected 70% of the data for each of fifteen participants. Block cross-validation and testing were performed, each on half, of the remaining 30% of the total data for individual participants.

Comparing classification results with ground truth participant survey ratings, we sought to distinguish between the four different groups of emotions defined in the advertisement study, namely, 'Excited Happiness', 'Pleased & Content', 'Neutral', and 'Displeased'.

As a result of the pattern recognition capabilities of the CNN algorithm, it became evident that rearranging the channels in the data files according to the spatial configuration of the fNIRS probes was beneficial. Table 1.1 showcases the group of statistical measures, Accuracy, AUC and F-Scores, generated with a CNN algorithm and spatially rearranged data files (Spatial) next to the original CNN classification results (No Spatial Ordering).

Comparison	**Average Participant Results (No Spatial Ordering)**	**Average Participant Results (Spatial)**
Excited Happiness vs. Pleased and Content	Accuracy: 69.4% AUC: 0.787 F-Score: 0.732	Accuracy: 76.6% AUC: 0.816 F-Score: 0.785
Excited Happiness vs. Neutral	Accuracy: 65.1% AUC: 0.899 F-Score: 0.726	Accuracy: 74.6% AUC: 0.902 F-Score: 0.777
Pleased and Content vs. Neutral	Accuracy: 71.6% AUC: 0.770 F-Score: 0.743	Accuracy: 78.6% AUC: 0.837 F-Score: 0.793
Excited Happiness vs. Displeased	Accuracy: 77.4% AUC: 0.842 F-Score: 0.855	Accuracy: 83.5% AUC: 0.942 F-Score: 0.889
Pleased and Content vs. Displeased	Accuracy: 80.2% AUC: 0.902 F-Score: 0.870	Accuracy: 86.2% AUC: 0.918 F-Score: 0.895
Neutral vs Displeased	Accuracy: 75.1% AUC: 0.859 F-Score: 0.821	Accuracy: 80.5% AUC: 0.898 F-Score: 0.837

Table 1.1: Comparison of Average Participant CNN Accuracies, AUC and F-Score Values Based on fNIRS Data Files with the Absence of Channel Ordering and Spatially Rearranged Channel Ordering.

The overall improvement in testing metrics for the spatially rearranged input data in the advertisement study supported the use of the novel approach of adding a spatial component to input data to enhance the performance of CNN-based models.

Research continued with further exploration into the CNN's ability to learn the spatial

characteristics of input data. More specifically, we began to study how CNNs perform with other types of spatiotemporal data. Also, as a way of putting emphasis on the spatial characteristics of input data, a multi-labeling schema which assigned spatially descriptive labels to different regions of interest was developed. Additionally, this type of multi-labeling gave us the ability to detect simultaneous activity in the regions of interest. Finally, seeking a way to catch important temporal information in the input data, an RNN was added to the model.

1.1 Research Impact

Being able to detect simultaneous activity in different regions of multidimensional data with spatiotemporal characteristics can be useful for many application domains. For example, integrating an algorithm, which can automate the process of surveillance by simultaneously detecting anomalies in different regions of interest captured through video can enhance security and provide peace of mind. In a survey of recent research related to intelligent surveillance monitoring techniques, Sreenu et al. [7] review different methods to discern unusual motion in crowd analysis. Jiang et al. [8] focus primarily on detecting anomalies in vehicular traffic at an intersection and consider both single object anomalies and co-occurring anomalies.

As research continued, a novel and promising approach to autonomously detect different levels of simultaneous and spatiotemporal activity in multidimensional data was established. This was aided by a new multi-labeling technique, which assigns different labels to different regions of interest in the data while enhancing the spatial aspect of the model. Each label is built to describe the level of activity/motion to be monitored in the spatial location that it represents, in contrast to existing approaches in current research which only provide a binary result as the presence or absence of activity. This novel spatially and motion-level descriptive labeling schema is combined with a CNN and Long Short

Term Memory (LSTM)-based network for classification to capture different levels of activity both spatially and temporally without the use of any foreground or object detection. The proposed approach can be applied to various types of spatiotemporal data captured for completely different application domains. Initially, it was evaluated on surveillance video data as well as respiratory sound data. Metrics commonly associated with multi-labeling, namely Hamming Loss and Subset Accuracy, as well as confusion matrix-based measurements were used to evaluate performance. Promising testing results were achieved with an overall Hamming Loss for video datasets close to 0.05, Subset Accuracy close to 80% and confusion matrix-based metrics above 0.9. In addition, the proposed approach's ability in detecting frequent motion patterns based on predicted spatiotemporal activity levels was explored. Encouraging results were also obtained on a small respiratory sound dataset while detecting abnormalities in different parts of the lungs. The experimental results demonstrated that the proposed approach can be applied to various types of spatiotemporal data captured for different application domains.

Due to favorable results obtained from applying this novel multi-labeling/multi-class spatially descriptive deep learning classification to a small dataset of respiratory sound data, the investigation into this model's compatibility with this type of spatiotemporal data continued. In this case, the approach was applied to multi-labeling spatiotemporal data to detect different classes in several regions of interest simultaneously enabling us to autonomously detect different types of breathing sounds in audio recordings as a supplement to traditional auscultation. In this scenario, the multi-labeling technique assigns labels to different auscultation sites (i.e. regions of interest), and a given label describes a type of breathing sound (normal, wheezing only, crackling only, wheezing and crackling) to be monitored in the spatial location that it represents. By considering several areas of the chest and corresponding labels simultaneously, a CNN and LSTM-based network was trained to classify the aforementioned pulmonary sounds spatially. Moreover, a Mel-Spectrogram representation of the audio data as well as raw waveforms were used and their

performances were compared when using this multi-labeling approach. In addition, the performance of the spatially informative multi-location/label model was compared with a single location/label model to support the former's ability to learn label dependency. The evaluation and comparison of outcomes was performed with Hamming Loss, along with confusion matrix-based measurements. The best testing results were generated by the Mel-Spectrogram multi-location/label model with an average Hamming Loss of 0.10, and average F-Score of 0.90. The experimental results supported the extended use of this novel approach to classify different forms of spatiotemporal data.

In the continued investigation into multi-labeling spatiotemporal data to detect different classes in several regions of interest simultaneously, we were interested in revisiting working with fNIRS data. In this study, the goal was to apply this novel approach to autonomously detect different finger tapping levels simultaneously in regions of interest. In order to do this, our multi-class multi-labeling technique assigned labels to the left and right index fingers, and a given label described one of the three different finger tapping frequencies (rest, 80 bpm, and 120 bpm) to be monitored in the corresponding contralateral spatial location in the brain's motor cortex. We trained a CNN/LSTM-based network to classify the aforementioned finger tapping levels spatially and simultaneously. The evaluation, based on simultaneous multi-label predictions for two brain regions, was performed with Hamming Loss, along with confusion matrix-based measurements. Promising testing results were obtained with an average Hamming Loss of 0.20, average F-Score of 0.80, and average Accuracy of 0.80. Moreover, we explained our model and novel multi-labeling approach by generating Shapley Additive Explanation values and plotting them on an image-like background, which represented the fNIRS channel layout used as data input. Using these Shapley values helped to add transparency and interpretability to our deep learning models, which aligns with the recent push to build out explainable and trustworthy AI.

The research presented in this thesis resulted in several publications including respected and peer-reviewed journals and international conference proceedings.

1.2 Publications

Peer-Reviewed Journals:

- N. Sommer, B. Kakillioglu, T. Grant, S. Velipasalar, L. Hirshfield, Classification of fNIRS Finger Tapping Data with Multi-Labeling and Deep Learning, Submitted to IEEE Sensors Journal, April 2021.
- N. Sommer, B. Kakillioglu, S. Velipasalar, L. Hirshfield, Detecting Wheezes and Crackles in Respiratory Sound Data Through Multi-Labeling and Deep Learning, Submitted to Computers in Biology and Medicine, April 2021.
- N. Sommer, S. Velipasalar, L. Hirshfield, Y. Lu, and B. Kakillioglu, Simultaneous and Spatiotemporal Detection of Different Levels of Activity in Multidimensional Data, IEEE Access Vol. 8, IEEE; 2020, p. 118205–118218.
- L. Hirshfield, P. Bobko, A. Barelka, N. Sommer, and S. Velipasalar, Toward interfaces that help users identify misinformation online: Using fNIRS to measure suspicion, Augmented Human Research, Vol. 4, No. 1, Springer; 2019, p. 1–13.

Peer-Reviewed Conferences:

- L. Hirshfield, T. Williams, N. Sommer, T. Grant, and S. Velipasalar-Gursoy, Workload-driven modulation of mixed-reality robot-human communication, Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, 2018, p. 1–8.
- N. Sommer, L. Hirshfield, and S. Velipasalar, Our Emotions as Seen through a Webcam, International Conference on Augmented Cognition, Springer; 2014. p. 78–89.

1.3 Organization of Thesis

Further details of the three main studies that were conducted are provided in the following chapters:

- Chapter 2 presents research based on applying our novel approach of detecting different activity levels simultaneously in four quadrants of video surveillance frames. In the context of videos, the multi-label vector represents distinct spatial regions (i.e. regions of interest) of an image frame, unlike common uses of the multi-label vector as a representation of the presence or absence of multiple descriptors of interest. Each region is assigned one of three possible levels of activity to train and validate a deep learning model. One of three different motion densities are identified simultaneously in four quadrants (without loss of generality) of a video. Label imbalance is also addressed, which is a common problem that is magnified in multi-labeling, as testing results are presented. Metrics commonly associated with multi-labeling, namely Hamming Loss and Subset Accuracy, as well as confusion matrix-based measurements are used to evaluate performance.

- Chapter 3 presents the use of our approach in detecting various types of breathing sounds simultaneously in four auscultation sites using respiratory sound data. The multi-labeling technique assigns labels to four different auscultation sites (i.e. regions of interest), and a given label describes a type of breathing sound (normal, wheezing only, crackling only, wheezing and crackling) to be monitored in the spatial location that it represents. By considering several areas of the chest and corresponding labels simultaneously, a deep learning model classifies the aforementioned pulmonary sounds spatially. Two audio data formats are used, namely Mel-Spectrograms and raw waveforms, to compare performance when using the multi-labeling approach. In addition, the performance of the spatially informative multi-location/label model is compared with a single location/label model to support the

former's ability to learn label dependency. The evaluation and comparison of outcomes is performed with Hamming Loss, along with confusion matrix-based measurements.

- Chapter 4 illustrates how the third main form of spatiotemporal data studied in this research, Functional Near Infrared Spectroscopy (fNIRS) data, is classified with spatially descriptive multi-labels for the simultaneous detection of neural activity due to different finger tapping rates in the two sides of the brain's motor cortex. The multi-labeling technique assigns labels to the two index fingers which correspond to the two sides of the brain's motor cortex (i.e. regions of interest). A given label describes a finger tapping frequency (rest, 80 bpm, and 120 bpm) to be monitored in the spatial location that it represents. A deep learning-based network is trained to classify the aforementioned finger tapping levels spatially and simultaneously. The evaluation of outcomes is performed with Hamming Loss, along with confusion matrix-based measurements. Also, an explanation of the network is presented with the generation of Shapley Additive Explanation Values.

- The Conclusion and Future Work Chapter provides a summary of what was learned during this research and planned future work.

Chapter 2

Simultaneous and Spatiotemporal Detection of Different Levels of Activity in Video Data

2.1 Introduction

Motivated by an interest in the autononous detection of concurrent activity in several regions of interest, a study was conducted with videos. The approach is unique since a novel multi-label/multi-class way of annotating video input data is introduced. Such a study can be beneficial for a wide range of application domains. For example, it can provide improvements in autonomous surveillance systems. Let's consider maritime surveillance for national security [10], [11]. Being able to detect a higher level of activity in regions of interest in SAR (Synthetic Aperture Radar) imagery may trigger other parts of the surveillance system or prompt an alert to human supervision. In general, integrating an algorithm, which can automate the process of surveillance by simultaneously detecting anomalies in different regions of interest captured through video can enhance security and provide peace of mind. In a survey of recent research related to intelligent surveillance monitoring techniques,

Sreenu et al. [7] review different methods to discern unusual motion in crowd analysis. Jiang et al. [8] focus primarily on detecting anomalies in vehicular traffic at an intersection and consider both single object anomalies and co-occurring anomalies. The last two examples represent models which are trained to specifically detect either pedestrians, in the first case, or vehicles, in the second case. This model was not built based on detecting the motion of specific objects but on general activity. That is why this method can be applied to other domains besides video surveillance such as Neuroscience and other computer vision applications. For the former, we can gain a better understanding of the concurrent use of different functional cognitive brain regions, which can help us to understand the types of cognitive load experienced by the user of an adaptive system. For example, an airline pilot's autopilot could provide the pilot with a visual overlay to support them in flight if it is determined that they have a high working memory load. However, the same system may choose to provide this information through the auditory channel via a speaker in the cockpit if it is determined that the pilot is experiencing both high working memory load and high visual perceptual load [12], [13]. For the latter, detecting simultaneous and various activity levels in different parts of video frames can lead to an improvement in event annotation. Ballan et al. [14] emphasize the importance of detecting multiple events happening at different times and locations throughout various types of videos as key to video annotation. Zhao et al. [15] use concurrent group activity classification to achieve their best results in annotating mobile videos. Similarly, Liu et al. [16], support the necessity of simultaneous event detection in videos through the use of multi-labeling.

In this chapter, a novel and promising approach to detecting different levels of simultaneous and spatiotemporal activity in multidimensional data will be presented through the use of a multi-labeling technique initially inspired by León's research [17], wherein EEG data is used to look at levels of activation in different regions of the brain simultaneously. The goal was to help people with disabilities (often paralyzed) to control their environment simply by thinking about moving fingers on one or both hands. Most publications,

on spatiotemporal tracking from videos, from the last five years do not use this type of approach [18], [19], [20], [21], [22]. The use of multiple labels to represent multiple actions has grown in popularity due to the interest in detecting and recognizing simultaneous activity in videos. For example, concurrent action recognition in hockey videos [23] could indicate that a 'Play', 'Face Off' and 'Fight' took place at the same time. Similarly, the ability to tag multiple facial expressions in videos can be accomplished using multi-labels to detect emotions, a crucial component of HCI [24]. However, in all of the aforementioned references, an action or a combination of actions is assigned to a label, which in turn is given a binary assignment representing its absence (0) or its presence (1).

In addition to the fact that interest lies in monitoring non-binary levels of movement in multidimensional data, the essence of this multi-labeling technique is also unique from a spatial perspective. Previous work using multi-labels to classify multidimensional data, while also interested in recognizing simultaneous actions, focused on describing specific types of actions, scenes and objects. In addition, the multi-labels used in previous work do not offer information about the locations of the actions, i.e. they do not address the spatial aspect. For example, Monfort et al. [25] add multi-labels to their Moments in Time Dataset to be able to assign a set of distinct actions concurrently to three second videos. Similarly, Yeung et al. [26], use multi-labels to add detail to the description of human actions. For example, an extra label will make it possible to distinguish a video of two people sitting and talking (multi-label: Sit,Talk) versus two people standing and talking (multi-label: Stand,Talk). Ray et al. [27] expand their multi-labels to include descriptions of background scenes and objects. In all three of these multi-labeled datasets, labels are chosen from a pool of choices to describe the events in the entire frame of a video. As mentioned above, these labels do not provide location information.

This chapter includes the following: i) a novel spatial aspect to a multi-label by assigning each label to a different region of interest. To the best of our knowledge, this is a unique way of handling the description of concurrent actions within a spatial context; ii) a proposed

approach that is able to detect different levels of activity, in contrast to existing approaches providing only a binary result as the presence or absence indicator [23], [24], [17]. Instead of simply assigning a '0' (absence of motion) or a '1' (presence of motion) to each label, we can assign, without loss of generality, a '0' (absence of motion), '1' (low level of motion) or '2' (high level of motion) to describe the level of motion in different regions of interest at a certain moment in time.; iii) an approach that is not specific to or fine-tuned for one type of application, but instead is suitable for different types of spatiotemporal data, ranging from videos to brain signals to lung sounds captured over time. To demonstrate this, we evaluate this approach on video data as well as on respiratory sounds, more specifically detecting the presence/absence of wheezing and crackling sounds in the lungs.

Also, in aerial maritime surveillance, it is advantageous to be able to detect concurrent anomalies in different regions of interest such as a coastal area. Moreover, the ability to not only detect presence, but also describe the level of activity in a region offers an additional facet to the alert that such a schema could provide in detecting motion. Therefore, designing a labeling technique that provides the option of assigning a non-binary level of activity to each label is necessary. For example, in Fig. 2.1, we show two possible maritime surveillance scenarios. On the left, we consider two regions of interest to monitor activity and on the right, the number increases to six to accommodate the complexity of the coastal area. In both cases, the use of a multi-labeling schema which allows for the concurrent detection of different levels of activity (i.e.: number of ships) in separate regions of interest would add to the ability of a surveillance system to identify specific patterns of activity. For example, concurrent activity or events at specific ports (designated as regions of interest) could indicate that some form of nefarious trafficking is occurring, and the Coast Guard could be notified.

To achieve spatially and motion-level descriptive classification outputs, we employ a deep learning approach, more specifically a Convolutional and Long Short Term Memory (ConvLSTM) Network, as the core of our model. The need for object detection, or

Figure 2.1: Examples of Maritime Surveillance Regions of Interest and Presence of Concurrent Activity

generation of region proposals is avoided with the proposed method.

In summary, we present an unconventional, though promising, method of using multi-labels to detect different levels of spatiotemporal motion in multidimensional data. In this chapter, without loss of generality, we will apply our proposed approach to video data as an example. As mentioned above, this approach can be applied to different types of spatiotemporal data captured for completely different application domains. In the context of videos, the multi-label vector represents distinct spatial regions (i.e. regions of interest) of an image frame, unlike common uses of the multi-label vector as a representation of the presence or absence of multiple descriptors of interest [28], [29], [30]. Each region is assigned one of three possible levels of activity to train and validate our deep learning model. We also address label imbalance, which is a common problem that is magnified in multi-labeling, as testing results are presented.

The remainder of this chapter is organized as follows: The related work is discussed in Section 2.2. The proposed approach and the network model are presented in Section 2.3. The datasets, labeling structure and the evaluation criteria are described in Section 2.4. The experimental results are presented in Section 2.5. An example of applying this method to another type of spatiotemporal data along with a comparison to published research is discussed in Section 2.6 . The chapter is then concluded in Section 2.7.

2.2 Related Work

The advent of multi-labeling is often linked to realizing that in describing multimedia resources, a single category is insufficient [31]. In particular, the need to assign simultaneous categories to videos is evidenced in the complexity of our environment, and plays an important role in artificial intelligence (AI) [25], [26], [27]. A conventional multi-class classification algorithm assumes that there is only one label to be used in the description of an event. Since the choice for the label is from a set of multiple descriptors (classes), the term multi-class is used [32]. Oftentimes, multi-labels are treated as subsets of multi-class labeling [28], [29], [30].

When it comes to dealing with label imbalance in the multi-label domain, there are two generally accepted approaches. In one case, the model is adapted to minimize the impact of the imbalance and, in the other case, the dataset is adjusted to minimize the labeling skew. One method that is commonly used when adapting the model is that the multi-label vector is split up into smaller sets of labels, each with its own classifier [33], [17], [34]. Another method involves assigning a new label to each multi-label and using multi-class classification [35], [36], commonly referred to as the Label Powerset transformation. The disadvantage of both methods resides with the loss of information about the correlation between the individual labels. The importance of gaining a better understanding of label correlation is emphasized by Zhang et al. [37]. Finally, another method in this category considers the occurrences of the classes for each label. The classification algorithm then assigns class weights to offset the skew between the minority and majority classes [23], [31].

The second approach to seeking a balanced set of labels is to make changes to the dataset itself. This includes using resampling techniques. Two common approaches use random undersampling and random oversampling. Applying these approaches to multi-labels becomes quite complicated. Charte et al. [38] test out these two resampling techniques. The best result was obtained by randomly oversampling each label separately within the multi-labels. The multi-labels were assigned to various datasets with domains

ranging from text to images as a preprocessing step before applying the data to different types of algorithms. Charte et al. [39] also take it one-step further by proposing a SMOTE (Synthetic Minority Oversampling Technique) that is specific to multi-labeled datasets (MLSMOTE). MLSMOTE is shown to outperform the other resampling techniques. Liu and Tsoumakas [61] improve the SMOTE resampling technique by considering the local distributions of labels when determining the minority class as opposed to the global distribution of labels. This algorithm, called MLSOL, shows promising results although it has yet to be applied to videos. Even though the outcome of using these resampling techniques leads to better balanced labels and data, there are certain disadvantages worth mentioning. In the case of undersampling, important data may be overlooked in the training process of the model. As for synthetic oversampling, creating artificial multi-labels and data is quite intricate. This challenge is augmented with raw video data.

Since videos are spatiotemporal in nature, we look to an algorithm which can handle both characteristics in our model design. There has been research conducted for spatiotemporal action detection. Zhu et al. [18] use a hidden two stream network, which combines a spatial stream CNN with MotionNet/temporal stream CNN to recognize human activity in geo-tagged videos. Similarly, Weinzaepfel et al. [19] make use of static and motion CNN features to detect actions in space and time. Also, temporal localization is accomplished with a sliding window. On the other hand, Saha et al. [21] and Yang et al. [22], use action tubes to track the actions detected by their networks. These works focus mostly on object tracking and rely on feature extraction, object detection or region proposal generation. In our research, we are interested in detecting various levels of activity simultaneously in different spatial locations. As seen with the aforementioned papers, along with others [40], [41], the world of computer vision has greatly benefited from the CNNs. Their application to multi-labeled datasets is also common [23], [42], [43]. We also employ CNNs as part of our design. Furthermore, to address the issue of label dependencies inherent to multi-labels, and temporal information contained in videos, a combination of CNN

and RNN is appropriate [43], [44], [45], [46], [20]. According to Wang et al. [43], the RNN framework can focus on the corresponding image regions when predicting different labels, which is very similar to a human's multi-label classification process. This quality of an RNN makes it highly suitable as part of the deep learning process. Additionally, according to Medel et al. [44], Convolutional LSTM (ConvLSTM) is deemed to be effective in modeling and predicting video sequences. Although CNNs are not developed with temporal features in mind, integrating them as part of an LSTM cell addresses this limitation. Deep learning allows the network to learn which features are important. ConvLSTM is able to temporally propagate spatial characteristics as supported by Shi et al. [47] and Zapata et al. [48]. One of the papers reviewed by Sreenu et al. [7] uses the ConvLSTM as the deep learning algorithm to discern unusual motion in crowd analysis [107]. However, unlike the approach to be described in the next section, the pedestrian trajectories of video inputs are represented with displacement vectors and used in training the model in lieu of the video frames. This necessitates the use of an encoder and a decoder.

2.3 Proposed Method

In our case, we will use a multi-label vector to represent distinct spatial regions. The spatial regions correspond to the regions of interest in the data. We determine the number of regions and the number of labels depending on the application and desired spatial resolution. For instance, in the case of video surveillance, regions of interest may depend on points of entry into the area captured by the camera. Without loss of generality, based on the motion to be detected in the chosen video datasets, as will be illustrated in Sec. 2.5, we visually determined that assigning four regions of interest to the video frames would be appropriate. Additionally, if the goal was to monitor the activity in a finer or coarser resolution, then the number of regions and, consequently, labels would have been adjusted.

In completing our labeling schema, each region is assigned to one of three possible mo-

tion level descriptors as will be presented in Sec. 2.4. Subsequently, it does seem as if each label does conform to a multi-class description system. However, in order to avoid confusion, we will refer to our data labeling schema as Spatially and Motion-Level Descriptive (SMLD) multi-labeling. The spatially descriptive multi-labels provide the ability to recognize the presence of concurrent actions in different regions of a frame/image, and the motion level identification offers a $non - binary$ and multi-level descriptor for each label. It is important to note that the classification of level of motion is performed independently for each region of interest. Therefore, viewing the results concurrently (i.e. in parallel) or sequentially would result in the same outcome.

Although the SMLD labels offer a novel way of describing different levels of activity in spatiotemporal multidimensional data, they too, like most multi-labeling schemas, are more prone to label imbalance. As stated by Herrera et al. [31], the learning from imbalanced data is another of the casuistics intrinsically linked to multi-label classification. Several techniques to offset this problem have been proposed in the past several years [34], [36], [39]. We have developed a network model that employs ConvLSTM in order to address the spatial and temporal properties of the videos. The labeling structure, which will be further discussed below, was designed to represent different activity levels at different spatial locations simultaneously. ConvLSTM-based networks have been used successfully in different applications, such as weather forecasting, robotic grasp slippage detection and understanding crowd behavior [47], [48], [107]. The ConvLSTM structure [50] is based on an LSTM cell with convolutional operators replacing the typical matrix multipliers. It is able to make use of a CNN's ability to extract spatial features and the LSTM's ability to extract temporal information while avoiding the problem of the 'vanishing gradient' of the RNN. Unlike a standard cascaded CNN-LSTM structure, it accepts multidimensional inputs making it suitable for video data applications. The ConvLSTM uses 3D tensors to learn spatial and temporal features simultaneously. The inner workings of the ConvLSTM structure is shown in Fig. 2.2.

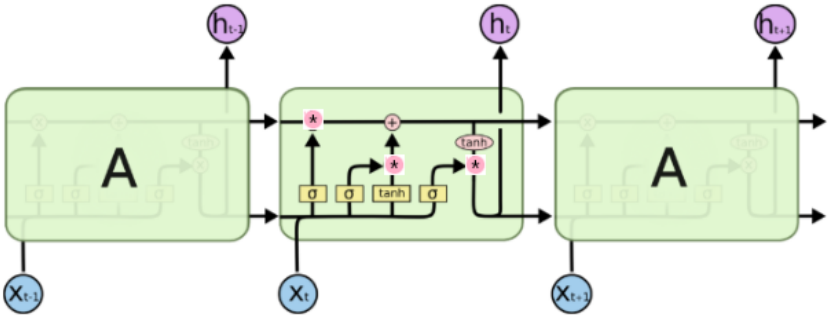


Figure 2.2: ConvLSTM Structure [50]

The details of our network model, which employs a Convolutional LSTM structure and the SMLD labeling schema, are shown in Fig. 2.3. The parameter values have been obtained after extensive evaluation. The model was compiled using binary cross-entropy loss and rmsprop optimization. Binary cross-entropy was a suitable choice, since the values used in the multi-labeling schema are one-hot encoded. Our empirical studies have shown better conversion with rmsprop optimization than with Adam optimization. We use early stopping, which monitors the validation loss. The latter automatically stops the training if the validation loss has not been decreasing after two epochs. After training, the proposed model is tested on unseen videos.

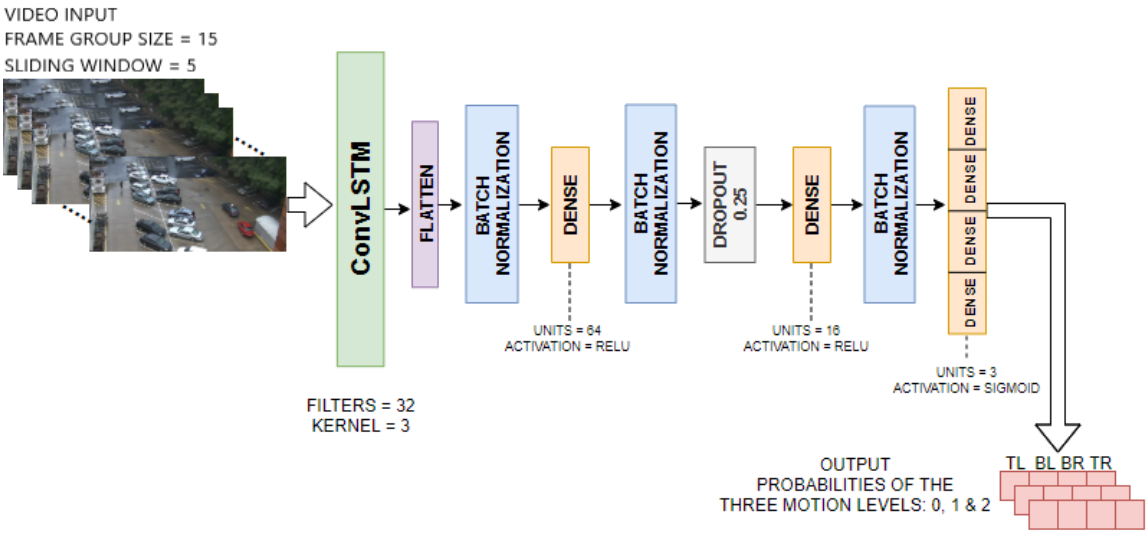


Figure 2.3: Proposed Model Structure

2.3.1 Detecting Motion Patterns

With this proposed approach, in addition to detecting different levels of activity in separate spatial locations, it is also possible to use these predicted levels of activity to detect motion patterns from one quadrant to another. We developed a new method that uses the detected levels of activity to autonomously detect inter-quadrant motion patterns. Therefore, another advantage of our proposed approach is that it provides the ability to detect motion patterns or frequent paths without relying on background subtraction, object detection or region proposal generation. In order to detect most frequent paths of motion from video datasets, we implemented a pattern detection script that can capture inter-quadrant motion by only using the predicted activity levels in different quadrants of the testing videos.

2.4 Datasets, Proposed Multi-Labeling Structure and Evaluation Criteria

We used two different sets of hand-annotated video data, obtained from open-access surveillance video datasets, as an example scenario to evaluate our proposed approach. Again, without loss of generality, we considered each video frame to have four quadrants. Each quadrant was represented by a label in the multi-labels. The decision to look at four regions of interest in the video frames suited our goal of being able to correctly classify the activity level in specific areas of a parking lot video, and then support these results by applying our model to another set of surveillance videos. The choice in the number of regions of interest is dependent on the information to be gathered from the spatiotemporal data. Depending on the spatial resolution and type of spatiotemporal data, the number of regions can be modified accordingly. If, for example, we want to focus on smaller regions, and monitor the activity in a finer resolution, then the number of regions can be increased. However, if a smaller number of regions is appropriate, as later shown with the respiratory data in

Sec. 2.6, our SMLD multi-labeling schema can be adjusted accordingly.

In the first two datasets, we employ the SMLD multi-labeling schema, described below, for each second of video, which is then fed to a 'doubly deep' [20] machine learning algorithm. After learning and capturing the spatial and temporal information contained in the training videos, our classifier is then tested on unseen video sequences from these two different sets of video data, and results are compared to ground-truth labels.

The datasets were prepared for training and validation by reading in the videos (frame rate = 30 fps), and resizing the frames, to 192x108x3 pixels, to reduce the memory requirements. Pixel values were normalized during this process. For temporal analysis via ConvLSTM, optimal results have been obtained with groups of 15 frames and a sliding window size of 5 frames. Each group of 15 frames was then assigned the multi-label of the middle frame in the group, i.e. the multi-label of the eighth frame in this case. The multi-labeled data was used for training using a batch generator to accommodate its size. Training was based on 80% of the data, and the remaining 20% was used for validation.

We used different subsets of videos from the VIRAT Video Dataset [51] to perform two sets of experiments, and evaluate the performance of the proposed work. The description of the video datasets, the details of labeling, and the discussion of the results will be provided in detail in Sec. 2.5.

Considering the fact that one possible application area for this proposed approach is autonomous surveillance, with videos being the multi-dimensional data, we tested our spatiotemporal classification method as a first-level alert indicator. We used a set of parking lot videos and another set of videos providing the view of a public outdoor space. These videos, from the VIRAT Video Dataset [51], were captured by a static camera. The types of motion in these videos included cars entering and leaving the parking lot, people walking in groups and separately, trees swaying with the wind, moving shadows and glare.

For our analysis, without loss of generality, we considered all the video frames to have four regions of interest, namely Top Left (TL), Top Right (TR), Bottom Left (BL) and

Bottom Right (BR). Four quadrants were chosen for illustration, and if finer/coarser resolution is desired for spatial regions, this number can be increased/decreased. We then used labels of '0', '1' and '2' to describe three different levels of motion in each quadrant. The definition for each level of motion that was used in our labeling schema is shown in Table 2.1.

LEVEL	DEFINITION
0	No Activity
1	Less Than 25% of Quadrant Shows Activity
2	More Than 25% of Quadrant Shows Activity

Table 2.1: Three Different Levels of Activity Are to Be Detected in Each Quadrant.

Each second of video was hand-labeled and given a unique SMLD multi-label assignment with four labels which represented the quadrants of the video frames. This SMLD multi-label was to be assigned to all 30 frames within this second of video. Before training and validating our model, the video frames were grouped in chunks of 15 frames with a sliding window size of 5 frames and assigned a final multi-label, corresponding to the middle (eighth) frame of the 15-frame group. The optimization of our empirical results led us to choose groups of 15 frames for temporal processing. This is also supported by the research of Sager et al. [52] who demonstrate that using chunk sizes of approximately half of a video's frame rate leads to the best activity detection accuracy. Similarly, Sozykin et al. [23] also use a chunk size of 15 frames with a sliding window of 5 frames to classify activity in hockey videos. This process aids in increasing the number of training samples and improves continuity in tracking motion.

For each set of surveillance videos, testing was performed on unseen videos from the same dataset. After testing, predictions were made for each group of frames. These predictions were a set of probabilities for each activity level assigned to the respective quadrants. The activity level with the highest prediction probability was chosen to be part of the resulting multi-label. Then, to manage the additional labels generated by the sliding window, the multi-label that occurred the most often defined the final assignment for the corresponding

second of video. These final multi-labels were then compared to ground truth labels.

2.4.1 Evaluation Criteria

The analysis of the results of multi-label classification is more complex than single label classification, and a standard set of metrics is yet to be established. Multi-label classification results can be instance-based or label-based. In other words, when a metric is based on an instance, the entire quadruple frame-wise multi-label assigned to a second of testing video is analyzed. In essence, the labels are looked at concurrently. On the other hand, a metric that is label-based will consider each single quadrant-wise label in the testing results separately. In this case, the labels are viewed sequentially. Therefore, the evaluation criteria will include frame-wise quadruple label classification metrics, such as Hamming Loss and Subset Accuracy, along with single quadrant-based classification metrics, which include Micro-averaged Precision, Micro-averaged Recall and Micro-averaged F-score [34], [37], [54]. Pereira et al. [54] recommend using Hamming Loss as a measure due to its popularity in multi-label research, and ability to report a classifier's overall prediction error.

i) Hamming Loss

We adapted the conventional binary-based Hamming Loss computation to take into account the three possible levels of motion by comparing the predicted value for a label during an instance (a second) with the ground truth value at that time. Average Hamming Loss was calculated as shown in Eq. (2.1), where S is the number of seconds in a testing video, and $p_{i,j}$ and $g_{i,j}$ represent the predicted label and the ground truth label, respectively. As seen in this equation, discrepancies were assigned a '1' when the predicted activity level for each label within a multi-label did not match the ground truth activity level. These values were then added and averaged over the product of the number of labels (4) in a multi-label and the time span of the video in seconds (S).

$$\frac{1}{4S}\sum_{i=1}^{S}\sum_{j=1}^{4}[if(p_{i,j}=g_{i,j},0,1)] \tag{2.1}$$

ii) Subset Accuracy

Another instance-wise metric (i.e. frame-based quadruple label metric in our case), presented by Pereira et al. [54], is Subset Accuracy. This measure considers how often the predicted activity levels of the SMLD multi-label matched the ground truth activity levels for each second. Despite the fact that there is no way of discerning a mismatch for 1, 2, 3 or all 4 quadrants, this measure will still provide useful information in assessing the quality of the classifier. Subset Accuracy was calculated as shown in Eq. (2.2), where S is the number of seconds in a testing video, and P_i and G_i represent the 4-quadrant prediction label and the 4-quadrant ground truth label, respectively.

$$\frac{1}{S}\sum_{i=1}^{S}[if(P_i=G_i,0,1)] \tag{2.2}$$

iii) Micro-Averaged Precision, Recall and F-Score

Shifting focus to label-based (i.e. single quadrant-based) metrics, we also looked at Micro-averaged Precision, Recall and F-Score, as suggested by Pereira et al. [54]. The Spearman correlation coefficient values between this set of metrics and Subset Accuracy are low, and the Spearman correlation coefficient values between this set of metrics and Hamming Loss are even lower. Thus, that is why this group of metrics provides a complete set of performance criteria.

The micro-averaging metrics adapt well to the multi-class nature of the labels since we deal with three different motion level assignments: '0', '1' and '2'. For example, instead of averaging out the calculated precision/recall for each activity level ("class"), we consider the overall precision/recall for all three activity levels ("classes") at once. The calculations

of Micro-Averaged Precision (μAP) and Micro-Averaged Recall (μAR) are shown in Eq. (2.3) and Eq. (2.4), respectively. In these equations, TP_i, FP_i and FN_i represent the number of True Positives, the number of False Positives and the number of False Negatives for activity level i, respectively, where $i \in \{0, 1, 2\}$.

$$\mu AP = \frac{TP_0 + TP_1 + TP_2}{TP_0 + TP_1 + TP_2 + FP_0 + FP_1 + FP_2} \tag{2.3}$$

$$\mu AR = \frac{TP_0 + TP_1 + TP_2}{TP_0 + TP_1 + TP_2 + FN_0 + FN_1 + FN_2} \tag{2.4}$$

The harmonic mean of Micro-Averaged Precision and Micro-Averaged Recall is the Micro-Averaged F-score.

2.5 Experimental Results

First, without loss of generality, we evaluated the proposed approach on different video data as an example case. The results of this evaluation are detailed in this section. Then, to demonstrate that the proposed approach can be applied to various types of spatiotemporal data captured for completely different application domains, we also evaluated it on respiratory sounds, more specifically detecting the presence/absence of wheezing and crackling sounds in the lungs. We provide results of this analysis together with a discussion in Sec. 2.6.

2.5.1 First Set of Experiments

We used a parking lot video for this set of experiments. Our labeling schema can best be described with an illustration. Suppose that Fig. 2.4 is the representation of what is seen for one second of video. In the Top Left quadrant, there are 2 people walking (activity level = '1'); no motion in the Bottom Left quadrant (activity level = '0'); a car moving in the

LABEL #1 TL ACTIVITY	LABEL #2 BL ACTIVITY	LABEL #3 BR ACTIVITY	LABEL #4 TR ACTIVITY
1	**0**	**1**	**2**

Table 2.2: Example of the Spatially and Motion Level Descriptive (SMLD) Multi-label Schema

Bottom Right quadrant (activity level = '1') and trees swaying in the Top Right quadrant (activity level = '2'). The SMLD multi-label that would be assigned to all 30 frames within this second of video (determined based on majority) is shown in Table 2.2.

Figure 2.4: Four Quadrants Used to Assign Labels for the First Set of Videos.

Initially, the dataset included 30 videos. The number of instances for different classes/labels, for these 30 videos, are shown in Table 2.3. As can be seen, there is some skew in the distribution, causing activity level '0' to be considerably more prominent in the TL, BL and BR quadrants. In order to eliminate the problems inherently associated with imbalanced data, we sought to obtain a more uniform class distribution, and have more balanced data. At first, we tried to balance the labels using class weight assignments in our model. However, this did not provide satisfactory results. Then, we devised a custom dataset augmentation technique. We wanted to avoid undersampling or creating synthetic samples because of the

disadvantages mentioned in Section 2.2. Looking at our labels, we noticed that some videos had significantly more activity level '1' assignments, which is one of the underrepresented levels as can be seen in Table 2.3. Thus, we augmented the dataset with modified/replicated versions of these videos in order to make use of these labels for balancing purposes. More specifically, the new videos were generated by flipping the individual quadrants either horizontally or vertically. Therefore, each video in the augmented dataset was unique. An example of a rotated version of a video whereby each quadrant has been flipped vertically is shown in Fig. 2.5. Similarly, an example showing a video generated by flipping the quadrants in the horizontal direction is shown in Fig. 2.6. This way, six additional videos were generated bringing the total size of the dataset to 36 videos, totaling close to 49 minutes in length. All of these videos were manually labeled, and have been used for training and validation.

	QUADRANT			
LEVEL	TL	BL	BR	TR
0	41%	41%	39%	34%
1	28%	30%	30%	30%
2	31%	29%	31%	36%

Table 2.3: Initial Activity Level Distribution for the First Dataset

After this customized data augmentation, we successfully obtained a more balanced dataset with better distribution of the number of label instances as shown in Table 2.4.

	QUADRANT			
LEVEL	TL	BL	BR	TR
0	37%	38%	36%	34%
1	31%	32%	33%	33%
2	32%	30%	31%	33%

Table 2.4: Activity Level Label Distribution after Video Data Augmentation

After testing the proposed model and obtaining predicted labels, we calculated average frame-wise quadruple-label Hamming Loss using Eq. (2.1). These results are shown in

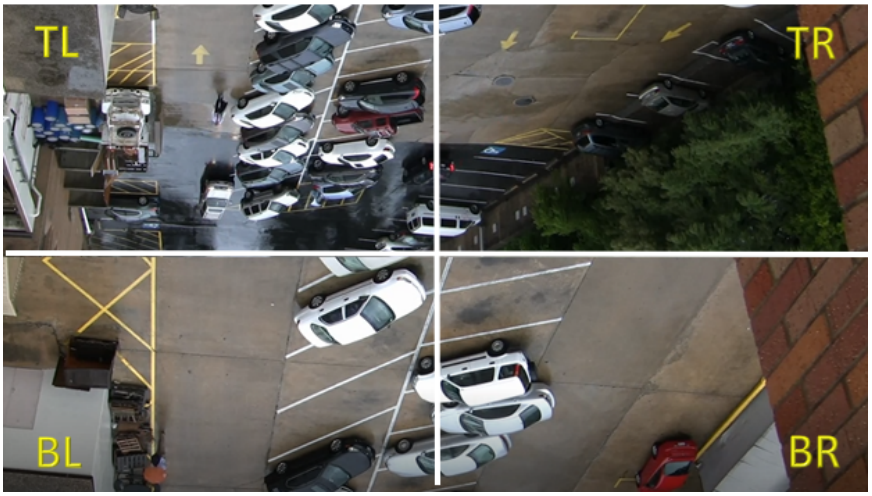


Figure 2.5: An Example Showing Four Vertically Rotated Quadrants for Generating a New Video for Balancing Label Distribution.

Figure 2.6: An Example Showing Four Horizontally Rotated Quadrants for Generating a New Video for Balancing Label Distribution.

Table 2.5. These Hamming Loss values indicate that, overall, our classifier made prediction errors about 5% of the time. The lower this value, the better the classifier is at predicting the correct motion level at each quadrant. It is difficult, and it would not be commensurate to compare these values to previous work [55], [56], since our experimental set-up and labeling schema are quite unique. For example, Dai et al. [55] use a concept-oriented multi-label for classification, which stores the physical attributes and design features of an object, e.g. shape, color and style of a smart watch, to offer comprehensive product

descriptive data. Trochidis et al. [56] use a multi-label to represent different emotions that a music piece might evoke. The type of information that these multi-labels carry is unlike the combined spatial and temporal characteristics of our SMLD multi-labels. Nonetheless, it can concluded that these low Hamming Loss results are promising.

Video 1	**Video 2**	**Video 3**	**Video 4**	**Video 5**
0.0541	0.0496	0.0438	0.0707	0.0476

Table 2.5: Average Hamming Loss Per Testing Video for the First Dataset

We then calculated the average frame-wise quadruple-label Subset Accuracy by using Eq. (2.2). The results are shown in Table 2.6. It should be noted that Subset Accuracy is quite unforgiving in that a small classification error for a 4-quadrant label will cause the entire predicted multi-label to be incorrect. This is because this metric does not distinguish between a single quadrant label mismatch and multiple quadrant label mismatches. If all 4-quadrant labels in an instance match the corresponding ground truth multi-label, the subset accuracy is set to '1'; otherwise, it is set to '0'. That is why, an average Subset Accuracy for all testing videos that is close to 80% can be considered to be a good result. It means that the algorithm is capable of predicting the correct 4-quadrant set of activity levels 80% of the time.

Video 1	**Video 2**	**Video 3**	**Video 4**	**Video 5**
81.1%	81.0%	81.3%	71.7%	76.2%

Table 2.6: Average Subset Accuracy Per Testing Video for the First Dataset

The set of metrics calculated for single quadrant-based labeling, namely Micro-Averaged Precision, Micro-Averaged Recall, and Micro-Averaged F-score values, are shown in Table 2.7. Considering that effects of data imbalance have been minimized in our training and validation data via data augmentation, our results are very promising and a good representation of our proposed SMLD model's ability to successfully predict different levels of simultaneous activity/motion in different spatial locations. This is exhibited with the confusion matrix parameters presented in Table 2.7. All four quadrants showcase robust values

for the testing videos. This is also supported by the frame-wise quadruple label-based Hamming Loss and Subset Accuracy results shown in Tables 2.5 and 2.6, respectively.

	QUADRANT			
METRIC	TL	BL	BR	TR
Micro-Averaged Precision	0.947	0.926	0.964	0.947
Micro-Averaged Recall	0.947	0.949	0.964	0.917
Micro-Averaged F-Score	0.947	0.937	0.964	0.932

Table 2.7: Single Quadrant-Based Label Metrics for the First Dataset

It should be noted that this algorithm was not designed for the purpose of 'object tracking', since the predicted activity levels produced by our approach do not distinguish between different types of objects or sources of motion. However, it can be observed that, with the SMLD labels, it is possible to detect certain singular and distinct motions as an added benefit. These motion patterns can be detected by looking at the sequence of predicted activity levels in adjacent quadrants. Most frequent motion patterns or common trajectories in the first set of videos included cars entering the TL quadrant from the top and driving to the BL quadrant, cars driving from the bottom of the BR quadrant to the TR quadrant, and people walking from the BL quadrant to the TL quadrant. We tested the ability of our proposed approach in detecting these motion patterns. By using the predicted motion levels at each quadrant over time, for all five test videos, we were able to successfully detect the aforementioned inter-quadrant motions simultaneously. Specifically, the TL to BL motion and BR to TR motion were detected 80% and 100% of the time, respectively. The accuracy of detecting BL to TL motion was 66.7%. Thus, the overall weighted accuracy for motion pattern detection is 83.3% for these three dominant trajectories/paths. This indicates that trajectory detection is possible with the predicted levels of activity without relying on object detection or tracking, when the inter-quadrant motion is distinct and not simultaneous. This accuracy was affected by classification errors during testing, and interference from additional intra-quadrant motions.

We also performed an experiment to investigate the ability of the proposed SMLD label-

ing approach in detecting a trajectory of activity across several quadrants (more than two). This would be possible by extending trajectory detection to less common inter-quadrant motion patterns. For this experiment, we also considered motion from the TR quadrant to the TL quadrant, an infrequent motion in the first set of testing videos. Adding this motion pattern to the more common BR to TR motion would enable us to detect motion from the BR quadrant up to the TL quadrant. This is still a challenging problem, since the quadrants of interest are not void of intra-quadrant motion interference, and our algorithm does not differentiate between different types and sources of motion within a quadrant.

As an example, based on the predicted levels of motion at different quadrants, we were able to detect a motion pattern from the BR quadrant to the TL quadrant in one of the test videos. By monitoring the predicted level of motion in the initial quadrant followed by consecutive motion in the subsequent quadrant, we cascaded the two inter-quadrant motions (BR to TR followed by TR to TL) in order to detect a trajectory from BR to TL. A comparison of the timestamps in the test video with those produced by our pattern detection script showed that the proposed approach was able to follow this type of trajectory of a car (activity level = '1'). Table 2.8 shows the timestamps detected for the BR to TR and the TR to TL motions. We then observed these moments in time in this test video to visualize the significance of our results. We present snapshots from the test video at the three main timestamps in Figures 2.7, 2.8 and 2.9. In Fig. 2.7, a car enters the BR quadrant at the 66th second. This correlates with the first detected start time for the BR quadrant seen in Table 2.8. At the 71st second of the video (Fig. 2.8), the car leaves the TR quadrant. Finally, the car leaving through the TL quadrant is seen at the 76th second (Fig. 2.9). The predicted activity levels in the BR, TR and TL quadrants, which were used to detect this pattern, are shown in Table 2.9. Looking at the highlighted '1' activity levels, which have a sequential pattern, the trajectory of the car can be followed as it travels from the BR quadrant to the TL quadrant while crossing through the TR quadrant. As mentioned above, our algorithm is designed to detect different levels of simultaneous motion in the

quadrants, without distinguishing the source of motion. Other non-highlighted activity levels in Table 2.9 correspond to different sources of motion. For instance, at the 66th second, a person is walking in the TL quadrant, and at the 74th second, the red parked car's hazard lights are turned on in the BR quadrant.

Start_Time	End_Time		Start_Time	End_Time
BR	TR		TR	TL
66 s	71s		69 s	76 s

Table 2.8: Start and End Times of the Detected Inter-Quadrant Motions from the Test Video of the First Set of Videos.

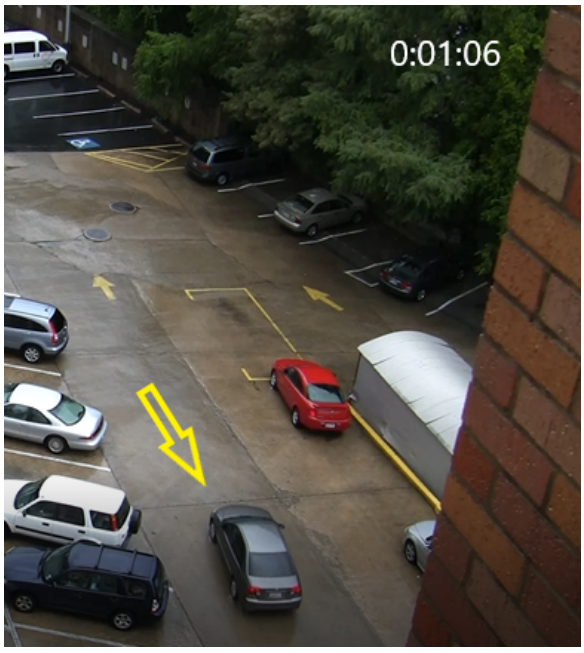


Figure 2.7: Snapshot from the Test Video for BR to TR Motion. A Car Enters at the 66th Second.

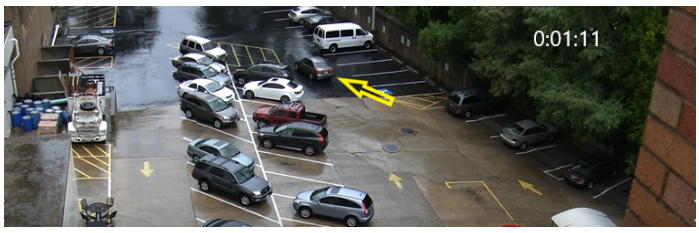


Figure 2.8: Snapshot from the Test Video for BR to TR Motion. A Car Leaves the TR Quadrant.

Figure 2.9: Snapshot from the Test Video for TR to TL Motion. A Car Leaves the TL Quadrant.

Time	BR	TR	TL
66 s	1	0	1
67 s	1	0	1
68 s	1	0	1
69 s	1	1	1
70 s	0	1	0
71 s	0	1	1
72 s	0	0	1
73 s	0	0	1
74 s	1	1	1
75 s	1	1	1
76 s	1	1	1
77 s	1	1	0

Table 2.9: Activity Levels in Regions of Interest to Illustrate Detected Multi-Quadrant Trajectory for the First Dataset

2.5.2 Second Set of Experiments

In order to further demonstrate the effectiveness of the proposed SMLD model in detecting different levels of activity, which happen simultaneously at different spatial locations in a multidimensional signal, we evaluated it on another group of videos from the VIRAT Video Dataset [51]. In these videos, a static camera provides the view of a public outdoor space. The types of activity in these videos include motion of trees, plants and patio umbrellas with the wind, people walking in groups and separately, and moving shadows as shown in Fig. 2.10.

Figure 2.10: Four Quadrants Used to Set Up Labels for the Second Set of Videos.

In this case, the threshold for level '2' activity was set slightly lower (to 20%) than for the first dataset (which had the value 25%). Since the motion in this dataset does not include larger objects such as vehicles, this adjustment to the threshold was necessary to be able to better represent an activity level of '2'. For example, looking at Fig. 2.10, the bottom left quadrant shows no activity so it would be assigned a '0'. The motions of individuals in the bottom right and top right quadrants would earn them labels of '1' (since the motion covers less than 20% of the respective quadrants). In the top left quadrant, a group of people

are walking and the trees are swaying at the same time resulting in a label of '2' for that quadrant. Putting these together, the multi-label for Fig. 2.10 is shown in Table 2.10.

LABEL #1 TL ACTIVITY	LABEL #2 BL ACTIVITY	LABEL #3 BR ACTIVITY	LABEL #4 TR ACTIVITY
2	**0**	**1**	**1**

Table 2.10: Example of Our Spatially and Motion Level Descriptive (SMLD) Multi-label Schema for the Second Dataset

Initially, this second set of data included 23 videos. Similarly to the previous dataset, we augmented the data with quadrant-wise flipped versions of the videos for label balancing purposes. In other words, the new videos were generated by flipping the individual quadrants either horizontally or vertically as described before. Therefore, each video in the augmented dataset was unique. This way, 5 additional videos were generated bringing the total size of the dataset to 28 videos totaling close to 28 minutes in length. All of these videos were manually labeled, and have been used for training and validation. Table 2.11 shows the distribution of the number of label instances for the second set of videos after label balancing via video data augmentation.

	QUADRANT			
LEVEL	TL	BL	BR	TR
0	33%	32%	33%	30%
1	33%	32%	31%	40%
2	34%	36%	36%	30%

Table 2.11: Activity Level Label Distribution for the Second Dataset

The set of videos was prepared for training and validation the same way as the first set of videos except that we had to account for a frame rate of 24 fps (instead of 30 fps as was the case for the first set). That is why the frames were grouped in chunks of 12 frames with a sliding window of 4 frames. The process of training and validation was identical to the one described above. Testing was also performed on five unseen test videos taken by the same camera.

The results for this experiment are presented in Tables 2.12, 2.13 and 2.14. Tables 2.12 and 2.13 show the frame-wise quadruple label-based Hamming Loss and the average Subset Accuracy, respectively. The single quadrant-wise label metrics are presented in Table 2.14. Applied to this second dataset, our classifier made prediction errors, on average, of about 4% of the time. The overall Subset Accuracy was close to 83% and, again, the Micro-Averaged Precision and Recall values were above 0.9. We can therefore conclude that the proposed method offers a promising model by providing a different approach to multi-labeling videos.

Video 1	**Video 2**	**Video 3**	**Video 4**	**Video 5**
0.0297	0.0341	0.0473	0.0411	0.0592

Table 2.12: Average Hamming Loss Per Testing Video for the Second Dataset

Video 1	**Video 2**	**Video 3**	**Video 4**	**Video 5**
88.1%	86.4%	81.1%	83.7%	76.3%

Table 2.13: Average Subset Accuracy Per Testing Video for the Second Dataset

	QUADRANT			
METRIC	TL	BL	BR	TR
Micro-Averaged Precision	0.987	0.948	0.905	0.934
Micro-Averaged Recall	0.987	0.948	0.968	0.934
Micro-Averaged F-Score	0.987	0.948	0.935	0.934

Table 2.14: Quadrant-Based Metrics for the Second Dataset

Frequent paths in the second dataset included people walking from the right of the TR quadrant to the TL quadrant as well as up and down the staircase on the left of the frames (i.e. motion between the TL and BL quadrants). Although our algorithm was not designed to support object tracking, by using the predicted motion levels at each quadrant over time for all 5 test videos, we obtained promising results in monitoring the aforementioned inter-quadrant motions simultaneously. Specifically, the TL to BL and BL to TL motions were

detected 75% and 100% of the time, respectively. The accuracy of detecting TR to TL motion was 87.5%. Thus, the overall weighted accuracy for motion pattern detection is 86.7% for these three dominant trajectories/paths. This indicates that trajectory detection is possible with our proposed approach and predicted levels of activity for the second dataset as well. This was accomplished simply through the detection of activity levels in groups of quadrants again without relying on any kind of object detection or tracking.

As an example, based on the predicted levels of motion at different quadrants, we were able to detect a motion pattern from the TR quadrant to the TL quadrant in one of the test videos. By monitoring the detected level of motion in the initial quadrant followed by consecutive motion in the subsequent quadrant, a pattern was detected. A comparison of the timestamps in the test video with those produced by our pattern detection script showed that the proposed approach was able to follow this type of trajectory of a person walking (activity level = '1'). Table 2.15 shows the timestamps detected for the TR to TL motion in this example. we then observed these moments in time in this test video to visualize the significance of our results. We present snapshots from the test video at the two timestamps in Figures 2.11 and 2.12. In Fig. 2.11, a person is seen in the TR quadrant at the 53rd second. This correlates with the detected start time for the TR quadrant seen in Table 2.15. At the 69th second of the video (Fig. 2.12), the person is on the way out of the TL quadrant headed to the BL quadrant. The video ends at that point so further motion cannot be detected.

Start_Time	End_Time
TR	**TL**
53 s	69 s

Table 2.15: Start and End Times of the Detected Inter-Quadrant Motion from the Test Video of the Second Set of Videos.

Figure 2.11: Snapshot from the Test Video for the Beginning of the TR to TL Motion. A Person Enters at the 53rd Second.

Figure 2.12: Snapshot from the Test Video for the End of the TR to TL Motion. A Person Leaves at the 69th Second.

2.6 Discussion

In order to show that the proposed approach can be applied to different types of spatiotemporal data, and illustrate the strength of this approach in providing multi-labels together with location information, we performed an experiment with respiratory sounds. Due to the current world health crisis that is taking place with COVID-19, we thought that it would be timely to assess how this approach could be used to detect sounds associated with lung disease. Using Kaggle's Respiratory Sound Database [57], we set up our multi-labels to detect the presence or absence of simultaneous wheezing and crackling sounds in the posterior left and right parts of the lungs. The simultaneous presence of these sounds in different parts of the lungs could be used to diagnose serious lung disease. Since we were interested in classifying an overall lung function, we chose two main regions of interest within

the dataset. This choice supported our goal of evaluating the performance of our SMLD multi-labeling and proposed approach on a different type of spatiotemporal data (other than videos) without addressing the medical significance of these sounds.

Since this study only focused on the presence or absence of wheezing and crackling sounds, the multi-labels became binary in nature. The Posterior Left (PL) and Posterior Right (PR) lung sound clips were concatenated and labeled with sets of four multi-labels representing PL crackling, PL wheezing, PR crackling and PR wheezing. The labels for each side of the lungs were provided in the database. As supported by Lee et al. [58], we used the raw waveform data instead of converting it to a spectrogram, a common practice when classifying signals in the audio domain [59], [60]. Lee et al.'s research shows that performance is not degraded by using raw waveform data. Training and validation were performed on a balanced label set. We compiled our classification results for comparison with the resulting F-scores from Kaggle's Notebook: CNN Detection of Wheezes and Crackles [59]. The predictions for the labels were generated independently which means that the results would be the same were we to consider the labels concurrently or sequentially. We were able to improve the F-Scores in all categories: absence of crackling/wheezing, presence of a crackling sound, presence of a wheezing sound and presence of both crackling and wheezing. It is important to note that our experiment is based on a subset of the overall respiratory sounds dataset with a focus on PL and PR lung sound clips. Also, the method presented in this Kaggle study differs from mine since a single multi-class label is used instead of multi-labeling and the algorithm is built out of a cascade of CNNs without any RNNs. The results showing the improvement by our method are presented in Table 2.16.

These results further strengthen the validity of our approach. Although our labeling technique and algorithm are unique, we were able to provide a commensurate comparison with the results presented in Kaggle. As we have shown, detecting two different types of patterns using lung sound clips was accomplished successfully with our approach. When

Sound	F-Score: Kaggle [59]	F-Score: Our Approach
No Crackling/Wheezing	0.81	**0.95**
Crackling	0.70	**0.91**
Wheezing	0.62	**0.86**
Crackling & Wheezing	0.57	**0.92**

Table 2.16: A Comparison of F-Scores in Detecting Respiratory Sounds

using the respiratory signals as input into our model, we treated the data with labels representing two spatial groups (PL and PR). The choice in the number of regions of interest is dependent on the information to be gathered from the spatiotemporal data. The improvement in classifying a certain type of sound or a combination of sounds inspires us to further explore this as a non-intrusive way of diagnosing lung disease in the following chapter.

Finally, the context of the spatiotemporal data is important in determining the significance of the correlations between regions of interest. For example, monitoring the path of motion is important in surveillance videos for security reasons. Another very important application domain, which would be well suited to this type of analysis, is in Neuroscience. If a stimulus elicits activity in different parts of the brain, it can be assumed that these activities are correlated, and different parts of the brain can be stimulated simultaneously and/or sequentially. The latter will be addressed in Chapter 4.

2.7 Conclusion

We have presented a novel and promising approach to detecting different levels of simultaneous and spatiotemporal activity in multidimensional data through the use of a new multi-labeling technique. In this chapter, without loss of generality, we applied our proposed approach to video data as an example, and will explore other domains in the next two chapters. The success of the proposed approach, as supported by a large spectrum of robust metric values, is encouraging in applying it to a variety of multidimensional datasets with spatiotemporal characteristics. Moreover, the ability of mapping out trajectories without

relying on object detection and tracking is an advantageous and preferable outcome. We also showed that our labeling technique and classification algorithm can produce successful classification results when applied to respiratory sounds. In the next chapter, we continue our study of classifying respiratory sounds by expanding our use of Kaggle's Respiratory Sound dataset [57]. Additional sound clips captured from other parts of the lungs are to be added with an increase in labels for our multi-labels. Also, another form of spatiotemporal data that we intend to apply to our method are brain data signals captured through a non-invasive brain measurement device. The nature of brain data offers the spatiotemporal data needed for the multi-label classifier outlined in this chapter. Separating the brain activity readings into separate regions of interest, as was done via four quadrants for the video datasets and two regions for the respiratory sound dataset, will enable us to use our novel multi-labeling system on specific functional brain regions of interest.

Chapter 3

Detecting Wheezes and Crackles in Respiratory Sound Data Through Multi-Labeling and Deep Learning

3.1 Introduction

Motivated by our continued interest in the autononous detection of concurrent activity in several regions of interest, a study was conducted with sound recordings. The uniqueness of this research is that, along with a novel multi-label/multi-class way of annotating the input data, the sound recordings from the various regions of interest were given a spatial configuration. Such a study may bring benefit to the correct interpretation of respiratory sounds and an assessment of the severity of pulmonary disease if various regions exhibit concurrent wheezing and crackling. Classical chest auscultation with an analog stethoscope has been used for the past two centuries since its invention by Rene Theophile Hyac in the Laënnec in 1816. Correct interpretation of breathing sounds has been shown to be dependent on the skill level and interpretation of the observer [62], [63]. A possible solution would be to use an electronic stethoscope, capable of capturing digital sound data,

in conjunction with a machine learning-based classification algorithm. Gurung et al. [64] support the fact that computerized sound analysis may improve diagnostic accuracy when used in conjunction with conventional chest auscultation. An improvement in diagnostic accuracy is crucial as auditory human perception can be quite subjective. For example, a study by Bohadana et al. [62] showed that a group of 143 healthcare professionals were able to identify normal breath sounds correctly about 17% of the time, wheezes about 85% of the time and crackles about 67% of the time. Similarly, Hafke-Dys et al. [65] found that even pulmonologists, who are respiratory disease specialists, achieved an average of 20% accuracy when identifying normal breathing sounds, an average of 62% accuracy when identifying wheezing and an average of 41% accuracy when identifying crackling.

Since the correct recognition of breathing sounds is challenging through chest auscultation with a traditional stethoscope, methods that offer improvements are sought after. As electronic stethoscopes gain popularity with their filtering and amplification abilities, human interpretation of the sounds has not been shown to improve significantly with these added features [66]. This could be explained by the undesirable difference in acoustic characteristics between electronic and analog stethoscopes [150]. Nonetheless, electronic stethoscopes offer the ability to capture sounds for automated analysis as a complement to professional interpretation. For example, electronic stethoscopes were also used in a small study [67] of positively-tested COVID-19 patients' abnormal breathing sounds. Disparity among the interpretation of these sounds, which included wheezing and crackling among others, by a group of physicians supported the need for additional diagnostic tools (i.e. signal processing) to minimize incorrect diagnoses. The use of automated analysis is further reinforced in the work by Brown et al. [68], when classifying coughing and breathing sounds from healthy, asthmatic and positively-tested COVID-19 patients. Preliminary machine learning results show an Area Under the Curve (AUC) average of 80% in distinguishing the coughing and breathing patterns of a subject infected with COVID-19. This supports the advantage of having an automated respiratory sound analyzer as part of a

pulmonary sound diagnosis.

Wheezing and crackling, among other adventitious sounds, are common abnormal breathing sounds, which point to lung disease. They can be present in lung diseases such as asthma, Chronic Obstructive Pulmonary Disease (COPD), and pneumonia. Accurate detection of these sounds is also important for early detection of COVID-19 symptoms associated with the lungs. These sounds, whether localized to one lung (unilateral) or diffused to both lungs (bilateral) can indicate the type of disease and its severity [69]. In addition to listening to both lungs, a thorough examination also comprises of auscultating anterior and posterior sides of the chest [63], [70], [71], [72].

In this chapter, we present a promising approach to detecting different types of breathing sounds concurrently at various spatial locations through our novel spatiotemporal multi-labeling technique. This type of multi-labeling does not have the spatial information that our multi-labeling approach has. The multi-location nature of chest auscultation to capture temporal lung sounds lends itself well to our proposed multi-labeling technique. This approach differs from traditional multi-labeling methods. In current research, multi-labeling means that one of more descriptors (classes) can be assigned to the data in question. For example, when classifying music genre, a musical piece might have elements of different styles such as pop, Deep House, and Raggae groove and these classes will become the elements of its multi-label [151]. In Chapter 2, we presented a spatiotemporal multi-labeling technique, and demonstrated promising results mainly in the classification of different levels of activity at different locations of surveillance video data, and on a very small subset of Kaggle's Respiratory Sounds database [57] (a.k.a. the ICBHI 2017 Challenge Respiratory Sound Database [73]). The subset of respiratory sounds included sound data from the posterior lung locations only. The differences of this work from our earlier study [84] include the following: (i) In this chapter, the Convolutional Neural Network (CNN) / Long Short Term Memory (LSTM)-based classification algorithm is adapted to not only handle the sound data in a raw waveform format, as in our earlier work, but also in a Mel-Spectrogram

format; (ii) for each of Mel-Spectrogram and raw waveform formats, we trained our models with the input data formatted in a spatial, quadrant-like configuration for simultaneous detection of abnormal lung sounds for a multi-output result, and compared the results; (iii) We also trained models by considering the data for each auscultation site separately and generating single output results. In the quadrant form, the four labels representing the four auscultation sites were assigned to the input data simultaneously. In the latter, the labels were assigned separately to each specific auscultation site's dataset; (iv) We use a larger set of respiratory sounds covering both posterior and anterior locations, and include four auscultation sites in the labels instead of two; (v) We increase the number of classes so that each label could be described with one of four sounds: normal breathing, wheezing only, crackling only and a combination of wheezing and crackling.

As mentioned in the previous chapter, León's research [17] influenced us in developing this labeling approach. In this research, various regions of the brain were monitored concurrently to detect levels of activation as a result of hand motor imagery captured with EEG data. Detecting levels of brain activation can ultimately help disabled (paralyzed) people interact with their environment as they think about moving fingers on their hands.

Automated spatiotemporal classification is usually associated with videos [8], [19], but is also being adapted to sounds. Spatiotemporal sound classification examples include geospatial sound modeling [74], audio scene detection in the realm of environmental sound analysis [75], and heart murmur detection [76], among others [77], [78]. Multi-labeling, also popular in videos, is commonly applied to the detection and recognition of concurrent events. For instance, simultaneous action detection in hockey videos [23] could categorize that various types of play in the game have taken place at the same moment. The use of many labels for spatiotemporal data classification has also been applied to sounds. For example, in their study, Trohidis et al. [56] assign multi-labels to music pieces which might give rise to various emotions. Nonetheless, in these studies, the multi-labels are a set of descriptors devoid of spatial context. For the study of sound abatement, Cartwright et

al. [79] created a database of multi-labeled spatiotemporal data. For the latter example, the sound tags, descriptive of urban sounds, are a set of descriptive classes based on the type of noise. Spatial information, provided with latitude and longitude values, is provided as a separate layer of input data. Spatial information is not ingrained in the actual labels and how they are defined. This is what sets our multi-labeling approach apart from those found in current research.

In our quest to classify breathing sounds, we apply this novel multi-labeling technique by designating a label to a specific auscultation site on the chest. Then, instead of the common multi-labeling technique of assigning a '0' or a '1' to each possible descriptor [37] to indicate its absence or presence, we establish a set of classes to choose from for each spatially descriptive label. Since we are interested in detecting wheezes and crackles, the four classes will be based on the absence of wheezing and crackling (i.e. normal breathing) ('0'), the presence of wheezing only ('1'), the presence of crackling only ('2'), and the presence of both wheezing and crackling ('3'). Acharya et al. [80] also use these four classes when detecting these two abnormal breathing sounds. However, their work differs from mine since only one type of sound is classified at a time. This research is based on a spatially informative set of multiple labels, which can detect abnormal breathing sounds simultaneously in various regions of the lungs.

The goals of our proposed approach encompass incorporating a novel spatial aspect to a multi-label in that each label gets assigned to a specific region of interest. The four regions of interest represent four lung auscultation sites, namely Anterior Left, Posterior Left, Anterior Right and Posterior Right. Each region is assigned a label representing one of the four classes. The four classes correspond to four breathing sounds, namely normal, wheezing only, crackling only, and a combination of wheezing and crackling. These multi-labels, assigned to sound data in either a raw waveform format or Mel-Spectrogram format, help train deep learning models to detect abnormal breathing sounds. To the best of our knowledge, this is a unique way of handling the detection of wheezing and/or crack-

ling simultaneously in different parts of both lungs. Our multi-label and spatiotemporal approach in labeling breathing data differs from the current state-of-the-art methods of classifying pulmonary auditory anomalies where only one type of sound is classified at a time [80], [81], [78], [83], [82].

Therefore, we propose an atypical but promising method of detecting wheezes and crackles in breathing sounds with the use of multi-labels. We use the Kaggle Respiratory Sounds database [57] (a.k.a. the ICBHI 2017 Challenge Respiratory Sound Database [73]) for our experiments. We compare our classification results from two different formats of the same input data to published research, which uses the same database. We also discuss label imbalance, an important consideration which becomes more prominent in multi-labeling.

The remainder of this chapter is organized as follows: The related work is discussed in Section 3.2. The proposed approach and the network models are presented in Section 3.3. The dataset, labeling structure and the evaluation criteria are described in Section 3.4. The experimental results, a comparison to published research, which uses the same database, and further analysis of our proposed method are presented in Section 3.5. The chapter is concluded in Section 3.6.

3.2 Related Work

Adventitious sounds, which are sounds that overlay normal breathing sounds, include wheezes and crackles. Wheezes are continuous in nature while crackles are discontinuous [63]. Wheezes are associated with COPD, asthma, COVID-19, and presence of a foreign object such as a tumor. On the other hand, crackles often occur with COPD, pneumonia, lung fibrosis as well as COVID-19. Wheezes are characterized by a dominant frequency of 1200 Hz and usually last about 100 ms, whereas crackles have a dominant frequency of 500 Hz lasting about 10 ms [63], [64], [71], [125], [86], [87], [88].

Being able to detect wheezes and crackles through an automated process can assist

in the diagnosis of the aforementioned diseases [63], [64]. The subjective interpretation of breathing sounds and disparity in auscultation training amongst medical professionals need to be addressed [62], [63], [65]. Attempts at improving auscultation skills through virtual patient simulators have proven to be successful in training medical students [89], [90]. Nonetheless, having an automated process, which can support a medical diagnosis through auscultation would be highly beneficial. Current research agrees with Gurung et al. [64], who state that "computerized analysis of recorded lung sounds may be a promising adjunct to chest auscultation as a diagnostic aid in both clinical and research settings".

Various methods have been studied for the implementation of an automated diagnostic tool in classifying abnormal breathing sounds. Using the same database as in our research, Acharya et al. [80] use Mel-Spectograms as inputs into a deep CNN/Recurrent Neural Network (RNN) model. An average score based on Sensitivity and Specificity for the four classes provided by the database's labeling system was reported as 66.38% with an improvement to 71.81% for patient specific models. In their study, the classification prediction is one label which indicates one of four possible breathing sounds. The work by Kochetov et al. [91], which also uses the same database and its four labeled classes, applies noise masking and Mel-frequency cepstral coefficient features to train an RNN-based model. Average Sensitivity/Specificity scores of 65.7% (based on all stethoscopes) and 67.9% (for a certain type of microphone, the AKG C417L) were reported. Perna [92], who also uses the same database, extracted features from a Mel-Spectrogram to train a CNN-based model. Although the results appear to be more robust, the classification categories used in this study [92] are quite general in nature. In one case, healthy and unhealthy sounds are sorted, and in the other case, sounds based on healthy, chronic and non-chronic diseases are differentiated. In the first case, the reported accuracy was 83% and the F-Score was 88%. The second case, the less general one, resulted in an accuracy of 82% and F-Score of 84%. In these last two studies, classification is also a single-label one. Chen et al. [125] classified three classes from the same database (normal breathing, wheezing,

crackling) with the use of an optimized S-transform and a deep Residual Network. The reported classification accuracy was 98.79% and the average Sensitivity/Specificity score was 98.14%. This latter study [125] was based on a smaller set of recordings, which did not include the combination of wheezes and crackles. Finally, Shuvo et al. [93] used hybrid scalograms derived from feature extractions along with a CNN model to classify three classes (chronic, non-chronic, and healthy) of lung disease and six classes of pathological diseases annotated in the database. An accuracy of 98.92% was reported for the former more general classification and 98.70% for the latter more specific one. These last two studies, unlike our approach, also classify respiratory sounds with one prediction, i.e. one label.

The aforementioned studies are based on the conventional multi-class classification algorithms using a single label to describe the type of breathing sound detected without taking into account the location of the recorded sound. In our case, we chose four different locations for our input data. The Kaggle Respiratory Sounds database [57] (a.k.a. the ICBHI 2017 Challenge Respiratory Sound Database [73]) contains data that is recorded from different chest locations, namely Trachea, Anterior Left, Anterior Right, Posterior Left, Posterior Right, Lateral Left, and Lateral Right. The four locations chosen for our study are: Anterior Left (AL), Anterior Right (AR), Posterior Left (PL), and Posterior Right (PR). Knowing the locations of wheezing and crackling sounds is very important. Finding these sounds in two or more auscultation sites can be a sign of greater risk of advanced disease and decreased lung function [149]. It is also important to compare the sounds produced by one side of the lungs to the other side to be able to detect a localized abnormal sound. For example, if a wheeze is only detected in a certain area of the lungs, it could indicate that a foreign body such as a tumor is blocking the airway [63], [71], [86]. Also, in a study by Zayet et al. [87], bilateral crackling sounds were found in the lungs of 24% of COVID-19 patients. This supports the significance of using multi-labels that can represent the simultaneous detection of normal breathing, wheezing only, crackling only, and wheezing and

crackling in different parts of the lungs. By comparing the sounds produced in one region and side of the lungs to other regions on the other side helps medical professionals ascertain whether the abnormal sound is localized or widespread. They can also tailor treatments for patients who have pre-existing lung issues with knowledge about the movement of sound in/toward specific regions of particular interest in the lungs.

The traditional approach to multi-labeling is prevalent in classifying environmental sounds. Environmental sounds are usually composed of simultaneous sounds from different sources. This means that in labeling a polyphonic sound event, multiple descriptors would be used bringing about the use of many labels. This type of labeling is useful in studying urban noise pollution, designing smart homes and security systems [74], [79], [94], [95]. Similar multi-labeling applications but, in this case, related to the medical field include classifying cardiac and pulmonary sounds, among others. For example, Baghel et al. [96] detect one of five cardiac disorders with a single prediction to aid in diagnosing cardio-vascular disease from heart sounds recorded by an electronic stethoscope. However, our interest lies in using a spatially descriptive set of labels to detect abnormal breathing sounds in different regions of interest simultaneously. To the best of our knowledge, designating labels to certain spatial locations and classifying the type of sound detected at each of these locations simultaneously sets our labeling schema apart from most research.

One approach which is closer to our method can be found in Nabi et al.'s research [97]. They present a technique of detecting asthmatic wheezing sounds in nine groups of auditory data separately. A group represents data from a specific auscultation location and/or the breathing phase (inspiratory/expiratory). Each group can be assigned one of three possible classes, which describes the severity of wheezing (mild, moderate, severe). Although this approach does have a spatial component in the way that the input data is presented, it does not offer a simultaneous classification of breathing sounds at different spatial locations. The classification of sounds for the groups was not handled concurrently, and the study was not geared towards detecting the presence of wheezes in different parts of the lungs at

the same time.

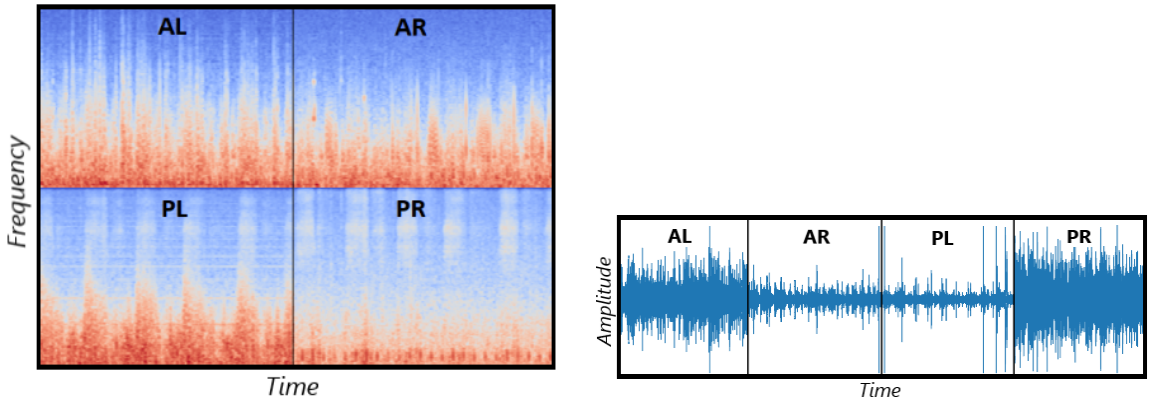
It is important to note that multi-labeling is prone to label imbalance [31]. This applies to all forms of spatiotemporal data that have been considered. In the previous chapter which was based on activity detection from videos, we augmented our video data to increase samples of minority labels by creating additional videos based on the original video dataset through rotation of frame quadrants. An interesting augmentation technique used by Baghel et al. [96], which was applied to cardiac audio signals, is called background deformation. This method adds background noise to the original signal. This type of data augmentation was also applied to pulmonary sounds to increase the number of samples of minority classes by Basu et al. [98]. Other techniques, which have also been applied to sounds, include assigning a greater cost to minority labels/classes during training [80]. Resampling techniques, such as random undersampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE) are also commonly considered to balance labels and classes [92]. The use of Sensitivity (Recall) and Specificity metrics to report classification results is also popular, since these metrics are less prone to label imbalance [80], [91], [97]. The F-Score metric is also less susceptible to class imbalance than accuracy, thus it is commonly used to report the success of classification in multi-labeling studies [94], [95].

3.3 Proposed Method

In this chapter, we use a multi-label vector to represent distinct spatial regions that correspond to chest auscultation sites. As previously mentioned, knowing the locations of wheezing and crackling sounds is very important. By comparing the sounds produced on one side of the lungs to the other side helps medical professionals ascertain whether the abnormal sound is localized or widespread. In our study, we used four auscultation sites of interest, namely AL, AR, PL and PR shown in Fig. 3.4, to maintain label balance when combining the pulmonary sounds from the four locations spatially and generating

one sound file. Although more auscultation sites were available in the database, adding more regions of interest would increase the possible number of different label vectors, and label balance would be harder to maintain.

In our attempt at improving the classification results, we created two different models based on different input data formats and representation. In the first case, we converted the sound input files into spectrograms, since this is a common approach in sound data classification [80], [81], [82], [86], [59], [60]. This offers an image-like representation of the spectrum of frequencies contained in the sound signal over time, fitting for the CNN part of our model. In the second case, we employed raw waveform files that we concatenated together to train and validate our second model. The input data, whether in a raw waveform representation or in a Mel-Spectrogram representation was formatted in adherence to custom quadrant-like configurations. A visualization of the two layouts is shown in Fig. 3.1.



(a) Mel-Spectrogram Layout (b) Raw Waveform Layout

Figure 3.1: Visualization of Input Data Formats

In our labeling schema, each spatial region is assigned to one of four possible breathing sounds: normal breathing, wheezing only, crackling only, and the combination of wheezing and crackling. Our spatially descriptive multi-labels enable the recognition of concurrent breathing sounds in different auscultation sites of the lungs simultaneously, and the identification of different breathing sounds through a $multi-class$ descriptor for each label. Our

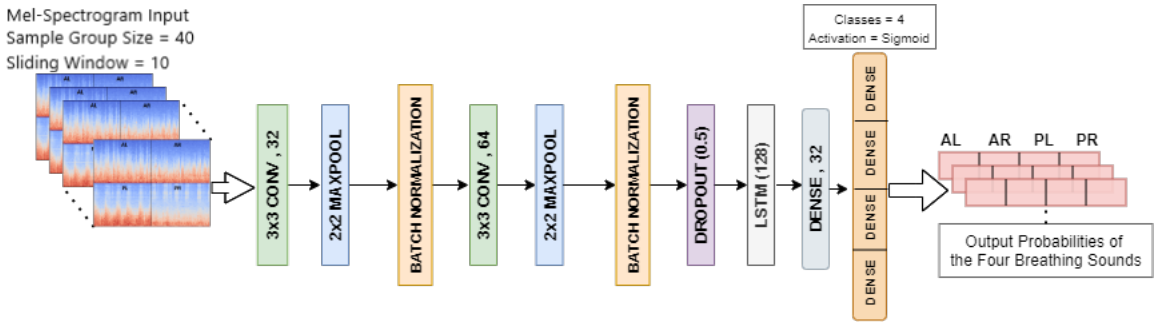


Figure 3.2: Proposed Spectrogram Model Structure

labeling schema, like traditional multi-labeling schemas, is more prone to label imbalance. That is why we were careful to maintain labeling balance as we chose sound files for the input data.

Since pulmonary sounds acquired through auscultation are spatiotemporal, we needed to choose an algorithm, which handles both types of characteristics for our model design. Our network model is based on CNN/LSTM to be able to detect the spatial and temporal properties of breathing sounds. Our labeling structure, which will be presented in detail in Sec. 3.4, was designed to represent different breathing sounds at different spatial locations simultaneously. By combining the CNN and RNN, we were also able to address the issue of label dependencies inherent to multi-labels and catch the temporal information [59], [60] contained in sound recordings. These network models are presented in detail in the next two subsections.

3.3.1 Network Model Using Spectrogram Input Data

The structure of the first network model, which uses both CNN and LSTM layers, quadrant formatted Mel-Spectrogram input data and our multi-labeling schema, is shown in Fig. 3.2. The network contains two 2D convolutional layers, each followed by max pooling and batch normalization layers. Then, following a dropout layer, there is an LSTM layer and a dense layer which leads to a final output layer. The parameter values shown in Fig. 3.2 were used in conjunction with an adaptive learning rate which decreased by a factor of 2

to a minimum of 10^{-6}. Binary cross-entropy loss and $rmsprop$ optimization are used in our model. Binary cross-entropy was a suitable choice, since the values used in our multi-labeling schema are one-hot encoded. We used an adaptive learning rate and early stopping, which monitors the validation loss. The learning rate decreases if the validation loss does not decrease after two epochs and then, training automatically stops if the validation loss has not been decreasing after three epochs due to early stopping. After training, we tested our model on unseen respiratory sound spectrogram files formatted the same way as the input data.

As seen in Fig. 3.2, the input to our proposed network model is formed like four quadrants, wherein top-left, top-right, bottom-left and bottom-right quadrants contain parts of the Mel-Spectrograms of the recordings of a participant's AL, AR, PL and PR regions of the lungs. In other words, for each set of recordings, the Mel-Spectrogram of each region of interest was divided up into chunks of 40 samples with a sliding window of 10 samples. Corresponding samples from the four auscultation sites formed the quadrants for each input sequence. Batches of these sequences were then used to train our network. As will be presented in Sec. 3.5, our experimental results show that the simultaneous training of the network and the analysis of the data from all quadrants provide better results compared to treating each of the four parts separately and independently.

3.3.2 Model Based on Raw Waveform Input Data

The details of the second network model, which also employs a CNN/LSTM structure, data formatted to contain raw waveform data from four regions of the lungs (AL, AR, PL, PR), and our multi-labeling schema, are shown in Fig. 3.3. The network contains two 1D convolutional layers, each followed by a batch normalization layer. Then, following a dropout layer, there is an LSTM layer and a dense layer which leads to a final output layer. The parameter values shown in Fig. 3.3 were used in conjunction with an adaptive learning rate which decreased by a factor of 2 to a minimum of 10^{-6}. Just as in the first model,

binary cross-entropy loss and $rmsprop$ optimization are used. We also used the same type of adaptive learning rate and early stopping. After training, the proposed model is tested on unseen respiratory sound raw waveform files formatted the same way as the input data.

As seen in Fig. 3.3, the input to our proposed network model is formed to provide a spatial format, wherein the arrangement (from left to right) contains parts of the raw waveforms of AL, AR, PL and PR regions of the lungs for each participant. For each recording, the raw waveform of each region of interest was divided up into chunks of 1000 samples with a sliding window of 200 samples. Corresponding chunks from the four auscultation sites were concatenated for each input sequence. Batches of these sequences were then used to train our network.

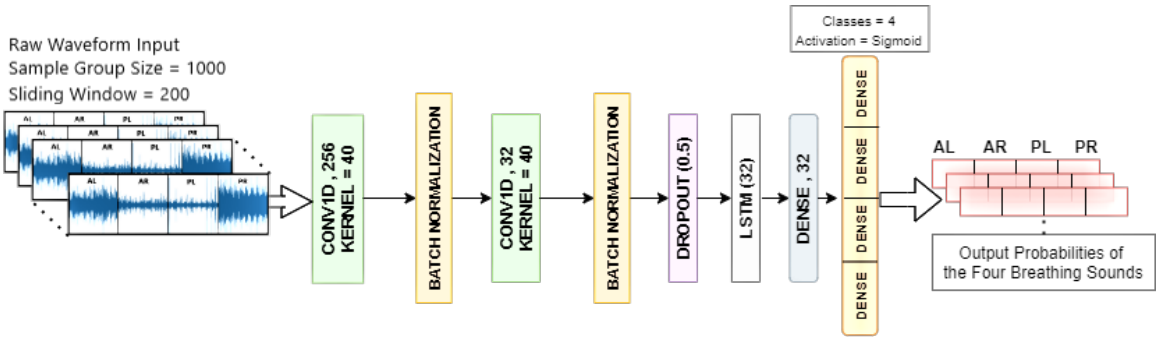


Figure 3.3: Proposed Raw Waveform Model

3.4 Dataset, Multi-Labeling Structure and Evaluation Criteria

3.4.1 Dataset

In this study, we have used a set of sound files from the Kaggle Respiratory Sounds database [57] (a.k.a. the ICBHI 2017 Challenge Respiratory Sound Database [73]), which contains 920 recordings from 126 patients in Portuguese and Greek hospitals compiled in 2017. Respiratory experts annotated each breathing cycle with one of four classes: normal

breathing, wheezing only, crackling only, and both wheezing and crackling. The recordings were captured with different stethoscopes and microphones. Our dataset is comprised of the recordings captured by the AKG C417L Microphone. As suggested by Kochetov et al. [91], we only used one type of recording source to maintain consistency. Although the database contained recordings from seven different auscultation regions (Trachea, Anterior Left, Anterior Right, Posterior Left, Posterior Right, Lateral Left, and Lateral Right), our dataset was built out of four regions of interest, Anterior Left (AL), Anterior Right (AR), Posterior Left (PL), and Posterior Right (PR). These were the locations, which were consistently recorded for the majority of the patients. Many recordings were missing Trachea and/or Lateral Lung sounds. By choosing data from these four spatial regions, we were able to maximize the sample size of our dataset. We included very important auscultation points and maintained a good balance between the labels' number of classes. Each of these four regions of interest was assigned a label and one of four possible classes.

Using samples provided for the AKG C417L Microphone (downsampled to 4 kHz), we generated 62 input files. Out of these files, 54 were chosen for training and validation. The rest were used for testing. We formed quadrants from the Mel-Spectrograms of the sound recordings of the four regions of interest (AL,AR,PL,PR), and shaped the input files as shown in Fig. 3.1(a). We concatenated the raw waveform data for the four regions of interest as shown in Fig. 3.1(b). These regions of interest define our multi-label and correspond to the auscultation sites shown in Fig. 3.4.

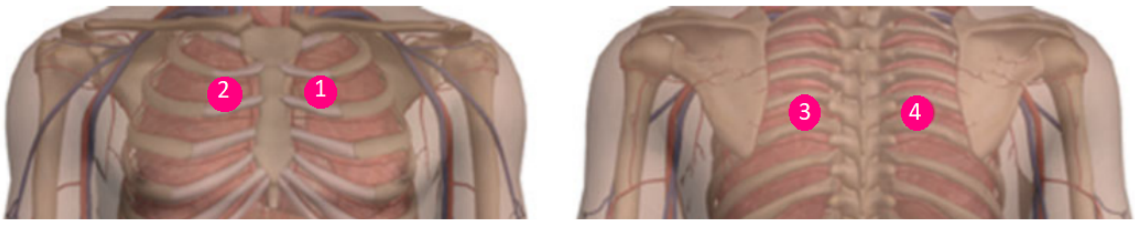


Figure 3.4: Spatial Multi-labels Correspond to Auscultation Sites: 1. Anterior Left (AL), 2. Anterior Right (AR), 3. Posterior Left (PL), 4. Posterior Right (PR) [73].

Since wheezes are characterized by a dominant frequency of 1200 Hz and crackles

have a dominant frequency of 500 Hz, a sampling rate of 4 kHz is considered appropriate [63], [64], [71], [125], [86]. The classes provided by the database for these four spatial locations were combined to create our multi-labels. Each of the 54 input files lasted 20 seconds resulting in 18 minutes of multi-labeled respiratory data for training and validation. After training, our classifiers were then tested on the remaining 8 similarly configured respiratory sound files that were not in the training and validation set. Results are then compared to ground-truth labels obtained from the expertly annotated files in the database.

We used two different formats of our dataset for comparison: Mel-Spectogram and raw waveform. For the Mel-Spectogram format, each sound waveform file from the dataset was converted to a 2D-like image with frequencies of the signal represented as pitches on a mel scale over time. Values were normalized during this process. With a sampling frequency of 4 kHz, each recording included a total of 80,000 sample values. For temporal analysis of the model, optimal results have been obtained with a chunk size of 10 ms and a 25% overlap. Therefore, groups of 40 samples, extracted from each of the four regions of interest, were formatted as shown in Fig. 3.1(a). The label of the middle sample in each group at each location, i.e. the label of the twentieth sample, in this case, was used as part of the multi-label. Our second data input format contained the raw waveform recordings of our dataset. Values were normalized and optimal results were obtained with a window size of 250 ms and a 20% overlap. Groups of 1000 samples, extracted from each of the four regions of interest, were concatenated and formatted as shown in Fig. 3.1(b). The label of the middle sample in each group, i.e. the label of the five hundredth sample in this case, was used as part of the multi-label. In both cases, our multi-labeled data was used for training with a batch generator. Training and validation were based on 54 input files from the dataset. The data was split so that 80% of it was used for training while keeping the remainder for validation.

3.4.2 Proposed Multi-labeling Structure

For our analysis, we considered all data, whether Mel-Spectrogram quadrants or raw waveform data from four regions of interest, to be assigned to a multi-label which describes four auscultation points of AL, AR, PL and PR. We then assigned each spatial region of interest a value of '0', '1', '2', or '3' to describe four different types of breathing sounds. The definition for each type of breathing sound that was used in our labeling schema is shown in Table 3.1.

TYPE OF SOUND	DEFINITION
0	Normal Breathing Sound
1	Wheezing Sound Only
2	Crackling Sound Only
3	Wheezing and Crackling Sounds

Table 3.1: Four Different Types of Sounds Are to Be Detected in Each of the Four Auscultation Sites of Interest.

As we combined the recordings from four regions of interest (AL, AR, PL, PR) for each input file, we combined the corresponding annotations from the database for our multi-label. For example, suppose that for a particular second of sound recording, the AL, AR, PL and PR regions are annotated to have wheezing (breathing type = '1'); normal breathing (breathing type = '0'); both wheezing and crackling (breathing type = '3') and crackling (breathing type = '2'), respectively. The multi-label of $[1, 0, 3, 2]$ would be assigned to the pertinent samples of the sound recording as shown in Table 3.2.

LABEL #1 AL Breathing Type	LABEL #2 AR Breathing Type	LABEL #3 PL Breathing Type	LABEL #4 PR Breathing Type
1	**0**	**3**	**2**

Table 3.2: Example of Our Spatially and Breathing Type Descriptive Multi-Label Schema

For each input data format, testing was performed on 8 quadrant-configured sound files that were not in the training and validation set. After testing, predictions were made for

each group of Mel-Spectrogram "frames" for the first case and group of raw waveform sound samples for the second case. Each type of breathing sound was assigned a probability for the respective region of interest. The class of the sound ('0','1','2', or '3') with the highest prediction probability was then chosen as part of the multi-label. Since additional labels were generated as the result of a sliding window used for the temporal analysis of our data, the multi-label with the greatest occurrence defined the final labels for each second of recording.

3.4.3 Evaluation Criteria

To be able to compare our results to research, which was conducted on the same database, we report the results based on the micro metrics of Sensitivity (Recall) and Specificity [80], [125], [91] along with F-Score and Accuracy [125], [92]. In addition, we use the measure of Hamming Loss, a common metric used in multi-label research. It offers a comprehensive look at a classifier's prediction error [54]. Metrics for multi-label classification results can be instance-based or label-based. An instance-based metric is based on the entire multi-label predicted for each time period of testing data. The Hamming Loss metric falls in this category. On the other hand, a label-based metric considers the label assigned to each region of interest in the testing results separately. Micro-averaged Sensitivity (Recall), Specificity, F-score and Accuracy [34], [37], [54] fall within this latter category.

i) Hamming Loss.

For each instance (in this case, one second of sound), we compared the group of four predicted classes of sound in our multi-label with the database's expert annotations. The average Hamming Loss for each testing audio recording was calculated as shown in Eq. (3.1), where S represents the number of seconds in a testing respiratory sound recording, and $p_{i,j}$ and $g_{i,j}$ indicate the predicted type of sound and the ground truth type of sound, respectively. Therefore, for each label within a multi-label, a mismatch is assigned a 1. These values are then added and averaged over the product of the number of labels (4) in a

multi-label and the time span of the recording in seconds (S).

$$\frac{1}{4S}\sum_{i=1}^{S}\sum_{j=1}^{4}[if(p_{i,j}=g_{i,j},0,1)] \tag{3.1}$$

ii) Micro-Averaged Sensitivity, Specificity, F-Score and Accuracy.

In order to provide a complete set of performance criteria, we also consider label-based (i.e. single auscultation site-based) metrics, such as Micro-averaged Sensitivity, Specificity, F-Score and Accuracy.

Since our labels can be assigned one of four classes: '0','1','2', or '3' to represent four types of breathing sounds, it is best to perform a micro-average of these parameters. For instance, instead of taking an average of the calculated Sensitivity (Recall) for each type of sound, we determine the overall Sensitivity (Recall) for all types of sound at once. The calculations of Micro-Averaged Sensitivity (μA_Sensitivity), Micro-Averaged Specificity (μA_Specificity), Micro-Averaged F-score (μA_Fscore), and Micro-Averaged Accuracy (μA_Accuracy) are shown in Equations (3.2), (3.3), (3.4), and (3.5), respectively. In the case of Micro-Averaged Specificity, since we are dealing with more than two classes (breathing sound types), we define True Negatives by using the one against all approach. In these equations, TP_i, FP_i, TN_i, FN_i represent the number of True Positives, False Positives, True Negatives, False Negatives for breathing sounds i, respectively, where $i \in \{0,1,2,3\}$.

$$\mu A\_\text{Sensitivity} = \frac{\sum_{i=0}^{3}(TP)_i}{\sum_{i=0}^{3}((TP)_i+(FN)_i)} \tag{3.2}$$

$$\mu A\_\text{Specificity} = \frac{\sum_{i=0}^{3}(TN)_i}{\sum_{i=0}^{3}((TN)_i+(FP)_i)} \tag{3.3}$$

The harmonic mean of Micro-Averaged Precision and Micro-Averaged Recall is the Micro-Averaged F-score, which can also be calculated the following way:

$$\mu A\_\text{FScore} = \frac{\sum_{i=0}^{3}(2 * TP)_i}{\sum_{i=0}^{3}((2 * TP)_i + (FP)_i + (FN)_i)} \tag{3.4}$$

$$\mu A\_\text{Accuracy} = \frac{\sum_{i=0}^{3}((TP)_i + (TN)_i)}{\sum_{i=0}^{3}((TP)_i + (TN)_i + (FP)_i + (FN)_i)} \tag{3.5}$$

3.5 Experimental Results

We evaluated our approach on two different formats of respiratory sound data. The goal was to detect different types of breathing sounds in four different spatial locations of lung auscultation simultaneously. Through our novel multi-labeling schema and deep learning-based algorithm, we built models to be able to perform automated classification of spatiotemporal breathing data and determine where (within our four regions of interest) normal sound, wheezing, crackling, and/or both wheezing and crackling were exhibited at the same time. The results of this evaluation are detailed in this section.

3.5.1 Simultaneous and Spatially Descriptive Region of Interest Analysis on Mel-Spectrogram Data

The training/validation dataset included 54 input files. The number of instances for different classes/breathing types, for these 54 files, was void of skew and imbalance. Each recording lasted 20 seconds resulting in 18 minutes of multi-labeled respiratory recordings for training and validation. Optimal results have been obtained with a window size of 10 ms and a 25% overlap. During training and validation, macro accuracy, macro F-Score and binary cross-entropy loss were calculated at the end of each epoch for each output label.

Training was monitored to make sure that validation loss was always smaller than training loss to prevent any possible overfitting. Once the validation loss did not improve for three epochs with an adaptable learning rate, training/validation would end. The final training and validation results are shown in Tables 3.3 and 3.4, respectively.

Label	**Accuracy**	**F-Score**	**Loss**
Label #1 (AL)	0.9740	0.9453	0.0670
Label #2 (AR)	0.9742	0.9670	0.0811
Label #3 (PL)	0.9728	0.9453	0.0677
Label #4 (PR)	0.9711	0.9416	0.0730

Table 3.3: Training Results on Mel-Spectrogram Data

Label	**Accuracy**	**F-Score**	**Loss**
Label #1 (AL)	0.9817	0.9632	0.0473
Label #2 (AR)	0.9742	0.9482	0.0613
Label #3 (PL)	0.9802	0.9605	0.0499
Label #4 (PR)	0.9758	0.9511	0.0578

Table 3.4: Validation Results on Mel-Spectrogram Data

After training on 54 input files, the classifier is then tested on an additional 8 similarly configured respiratory sound files captured with the AKG C417L Microphone. Results were compared to the labels provided by the expertly annotated files in the database. We determined the overall Hamming Loss using Eq. (3.1) by calculating the average for all groups of four labels for the testing audio files. These results are shown in Table 3.5. These values indicate that our classifier is capable of making correct predictions for our multi-labels for the testing audio recordings an average of about 90% of the time.

Test File #1	**Test File #2**	**Test File #3**	**Test File #4**	**Test File #5**	**Test File #6**	**Test File #7**	**Test File #8**
0.1250	0.1375	0.1250	0.1125	0.0875	0.1375	0.1000	0.0875

Table 3.5: Average Hamming Loss Per Testing Respiratory Sound File for the Mel-Spectrogram Based Model

Additional metrics, namely Micro-Averaged Sensitivity (Recall), Specificity, F-score,

and Accuracy values, were also calculated for each quadrant. In the case of Micro-Averaged Specificity, it is important to note that since we are dealing with more than two classes (breathing sound types), we define TN by using the one against all approach. Our model's ability to predict various respiratory sound types in different spatial chest areas is based on the compilation of our test file results for each label. Our results are presented in Table 3.6. A discussion of these results and a comparison to our second set of experimental results and published work that is based on the same database can be found in Sec. 3.5.5.

	SPATIAL LABELS			
METRIC	AL	AR	PL	PR
μA_Sensitivity	0.9467	0.9467	0.8736	0.8322
μA_Specificity	0.9754	0.9815	0.9563	0.9471
μA_F-Score	0.9430	0.9534	0.8750	0.8478
μA_Accuracy	0.9350	0.9520	0.8764	0.8750

Table 3.6: Auscultation Site-Based Label Metrics for the Mel-Spectrogram Dataset

3.5.2 Simultaneous Region of Interest Analysis on Raw Waveform Data

For the experiments with the raw waveform data, training and validation were also based on the same 54 sets of sound recordings. Optimal results have been obtained with a window size of 250 ms and a 20% overlap. During training and validation, macro accuracy, macro F-Score and binary cross-entropy loss were calculated at the end of each epoch for each output label. We also monitored training to make sure that validation loss was always smaller than training loss to prevent any possible overfitting. Once the validation loss did not improve for three epochs as the learning rate was adjusted, training/validation would stop. The final training and validation results are shown in Tables 3.7 and 3.8, respectively.

Testing was performed on the same 8 sound files previously used with the Mel-Spectrogram-based experiment but formatted according to the raw quadrant format. We determined the average Hamming Loss using Eq. (3.1) for all groups of four labels for the testing audio

Label	Accuracy	F-Score	Loss
Label #1 (AL)	0.9023	0.7858	0.2306
Label #2 (AR)	0.8803	0.7533	0.2586
Label #3 (PL)	0.9067	0.8034	0.2246
Label #4 (PR)	0.8961	0.7727	0.2413

Table 3.7: Training Results on Raw Waveform Input

Label	Accuracy	F-Score	Loss
Label #1 (AL)	0.8906	0.7640	0.1732
Label #2 (AR)	0.8841	0.7498	0.1751
Label #3 (PL)	0.8924	0.7729	0.1620
Label #4 (PR)	0.8962	0.7716	0.1457

Table 3.8: Validation Results on Raw Waveform Input

files, and the results are shown in Table 3.9. The Hamming Loss values indicate that our classifier made incorrect predictions about 26% of the time. It is important to seek the lowest value possible for this parameter since it would mean that the better the classifier is at predicting the overall group of breathing sound types for the four auscultation sites simultaneously.

Test File #1	Test File #2	Test File #3	Test File #4	Test File #5	Test File #6	Test File #7	Test File #8
0.2125	0.2125	0.2500	0.2225	0.3000	0.3250	0.2750	0.3250

Table 3.9: Average Hamming Loss Per Testing Respiratory Sound File for the Raw Waveform Based Model

Micro-Averaged Sensitivity (Recall), Specificity, F-score, and Accuracy values, were also calculated for each quadrant, and are shown in Table 3.10. Our model's ability to predict various respiratory sound types in different spatial chest locations is based on the compiled results of our test files for each label. We analyze these results in the following subsections.

	SPATIAL LABELS			
METRIC	AL	AR	PL	PR
μA_Sensitivity	0.7692	0.7769	0.7243	0.7218
μA_Specificity	0.9001	0.9119	0.8921	0.8868
μA_F-Score	0.7408	0.7761	0.7092	0.7196
μA_Accuracy	0.7217	0.7569	0.7246	0.7167

Table 3.10: Auscultation Site-Based Label Metrics for the Raw Waveform Format Dataset

3.5.3 Treating Each Region of Interest Separately and Independently

In order to illustrate the benefit of analyzing all four regions of interest simultaneously and forming data as spatiotemporal, we compared the proposed multi-output spatially-descriptive method with a single-output one. More specifically, we trained the models by using the data for each auscultation site separately for both Mel-Spectrogram and raw waveform data formats. In order to calculate the metrics, a "multi-label" was then formed out of the four separate outputs obtained. In other words, we placed the individual quadrant predictions in the same format as our true multi-label outputs (i.e. AL-AR-PL-PR), for a commensurate comparison of results. The average quadruple-label Hamming Loss values are shown in Tables 3.11 and 3.12 for Mel-Spectrogram and raw waveform data formats, respectively. Based on the increase in Hamming Loss to an average above 40% for both input data formats and comparing Tables 3.11 and 3.12 with Tables 3.5 and 3.9, it is noticeable that there is a dependency between labels of different locations, and the original proposed approach can better capture this dependency by providing a much lower Hamming loss. The increase in Hamming Loss is also accompanied by a drop in the other metrics (shown in Tables 3.13 and 3.14) when four regions of interest are treated separately and independently. When the results in Tables 3.6 and 3.10 are compared to results in Tables 3.13 and 3.14, respectively, a drop is observed in the latter set of tables. Again, this drop can be attributed to the fact that the dependency across different labels cannot be captured when each region is treated individually, and each auscultation site was used separately to train the neural network models with the use of a single label. Thus, the simultaneous and

spatially descriptive analysis of the regions of interest was significantly more robust than the independent region of interest and single output analysis, and our research shows the benefit of using spatiotemporal data, and analyzing multiple regions concurrently with an LSTM.

Test #1	**Test #2**	**Test #3**	**Test #4**	**Test #5**	**Test #6**	**Test #7**	**Test #8**
0.4000	0.3750	0.4250	0.3500	0.4625	0.35	0.375	0.4625

Table 3.11: Average Hamming Loss for the Mel-Spectrogram Based Model when Each Region Is Treated Separately

Test #1	**Test #2**	**Test #3**	**Test #4**	**Test #5**	**Test #6**	**Test #7**	**Test #8**
0.3125	0.3750	0.4000	0.3125	0.3750	0.45	0.525	0.575

Table 3.12: Average Hamming for the Raw Waveform Based Model when Each Region Is Treated Separately

	SPATIAL LABELS			
METRIC	AL	AR	PL	PR
μA_Sensitivity	0.6888	0.6414	0.5776	0.7242
μA_Specificity	0.8724	0.8748	0.8219	0.9080
μA_F-Score	0.6953	0.6559	0.5904	0.7719
μA_Accuracy	0.6590	0.6455	0.5786	0.7250

Table 3.13: Auscultation Site-Based Label Metrics for the Mel-Spectrogram Model when Each Region Is Treated Separately

	SPATIAL LABELS			
METRIC	AL	AR	PL	PR
μA_Sensitivity	0.5570	0.6956	0.6277	0.6575
μA_Specificity	0.8310	0.8584	0.8008	0.8763
μA_F-Score	0.5839	0.6910	0.6384	0.6886
μA_Accuracy	0.5439	0.6012	0.5813	0.6324

Table 3.14: Auscultation Site-Based Label Metrics for the Raw Waveform Model when Each Region Is Treated Separately

3.5.4 Analysis of Correctly Labeling Two Regions of Interest

Even though the Hamming loss provides a statistical way of representing the extent of mislabeling that occurs within a predicted multi-label, it does not specify which regions are correctly classified at the same time. We performed further analysis and experiments to evaluate the accuracy of correctly classifying two lung regions at the same time. Similar to the analysis above, we performed a comparison between the proposed spatiotemporal multi-labeling approach, and classifying each region separately on Mel-Spectrogram data. Tables 3.15 and 3.16 show the average accuracy of correctly classifying two lung regions (AL&AR, PL&PR, AR&PR, AL&PL, AL&PR and AR&PL) at the same time with our original approach and independent region analysis, respectively. As can be seen, the accuracy values are significantly higher when multiple lung regions are analyzed simultaneously with our original proposed approach. Moreover, the accuracy values for correctly labeling AR&PR and AL&PL at the same time are the highest in Table 3.15 indicating the model's ability to detect a stronger correlation for simultaneous analysis of recordings from the same sides of the lungs.

Quadrants	AL&AR	PL&PR	**AR&PR**	**AL&PL**	AL&PR	AR&PL
Average Accuracy	0.74	0.7	**0.86**	**0.79**	0.74	0.78

Table 3.15: Accuracy of Classifying Two Lung Regions at the Same Time by Using Mel-Spectrograms with Proposed Approach

Quadrants	AL&AR	PL&PR	AR&PR	AL&PL	AL&PR	AR&PL
Average Accuracy	0.25	0.45	0.35	0.31	0.42	0.35

Table 3.16: Accuracy of Classifying Two Lung Regions at the Same Time by Using Mel-Spectrograms and Analyzing each Region Separately

3.5.5 Further Discussion of Results

Although training and validation metrics were robust for both types of simultaneous region of interest/spatially descriptive input data, the Mel-Spectrogram format provided better classification results on test data. More specifically, the Mel Spectrogram-based Hamming Loss shows that, on average, our classifier made incorrect predictions for our multi-labels about 10% of the time, whereas the raw waveform-based Hamming Loss almost tripled to approximately 26%. This trend was further supported by all the other metrics. Micro-Averaged Sensitivity (Recall) was greater than 0.90 for all four spatially-labeled locations for the Mel-Spectrogram sound format as compared to a range of values between 0.72 and 0.78 for the raw waveform input format. Micro-Averaged Specificity, F-Score, and Accuracy also showed better results for the Mel-Spectrogram format, as seen in Table 3.6 in comparison to Table 3.10. Thus, our results support the use of Mel-Spectrogram input data with deep learning in classifying sound recordings. This is further supported by Becker et al. [99], who obtained a higher accuracy with spectrogram-based data than raw waveform-based data when classifying audio signals of spoken digits (0-9) in English. Xie et al. [100] also confirmed that the Mel-Spectrogram is the best type of spectrogram to use with a CNN model when classifying bird sounds.

Although the sound recordings of the database were described to be noisy by Rocha et al. [73], our proposed method was successful at classifying four different types of sounds at four different chest auscultation locations simultaneously. Our multi-labeling and simultaneous region analysis technique is unique and offers a spatial aspect to our multi-labels. For existing studies, which have used the same database, the approaches consist of using a single label, which gets assigned a class or a group of classes provided by the annotations from the database. Most studies focused on a smaller number of classes, and have not addressed the spatial aspect as mentioned above, For instance, the results by Perna [92] were based on generalized categories of healthy and unhealthy sounds in one case, and then healthy, chronic and nonchronic diseases in the other case. Accuracy of 83% and F-Score of

88% for the former case, and accuracy of 82% and F-Score of 84% for the latter case were presented, which are lower than the results we obtained with the Mel-Spectrogram format, as we achieved an average accuracy of 91% and an average F-Score of 90%. Since the categories (classes) for our multi-labels were specific to the ones provided by the database and not generalized into broader groups, our results can be considered more robust. Chen et al. [125] focused on three classes of the same database (normal breathing, wheezing, crackling) with the use of an optimized S-transform and a deep Residual Network. The reported classification accuracy was 98.79% and the average Sensitivity/Specificity score was 98.14%. Despite the good performance, the type of classification that was performed is also less specific than mine. Firstly, the class of combined wheezing and crackling sounds was not included. Secondly, their results were based on the detection of a particular sound excluding any spatial auscultation site information. Acharya et al. [80] and Kochetov et al. [91] used all four types of breathing sounds. The best reported averages of Sensitivity/Specificity scores were 71.81% and 67.9%, respectively. Our simultaneous and spatially descriptive analysis of regions of interest with Mel-Spectrogram data outperforms these numbers achieving average scores of 90% and 97%, respectively. Finally, Shuvo et al. [93], achieved exceptional results in classifying the data into both broad lung disease categories (chronic, non-chronic, and healthy) and specific pulmonary disease categories with 98.9% and 98.7% accuracies, respectively. Similarly to previously mentioned research, these categories do not offer a snapshot view of specific types of sounds at different spatial locations, as provided by our model. The classification is not as complex as ours.

Our proposed approach uses LSTM, which is a modified RNN known for its ability to detect label dependencies [101], [43], [102]. The results in Table 3.15 show that the use of our simultaneous and spatially descriptive multi-labeling schema, in conjunction with LSTM, can learn the dependencies across different regions of interest well.

3.6 Conclusion

In this chapter, we have shown that our approach of multi-labeling spatiotemporal data to detect different classes in several regions of interest simultaneously, can be applied to autonomously detect different types of breathing sounds ((normal, wheezing only, crackling only, wheezing and crackling) in audio recordings. Our multi-labels represented four auscultation sites (AL, PL, AR and PR). We presented a CNN and LSTM-based network to classify the pulmonary sounds spatially. We used Mel-Spectrograms as well as raw waveforms to represent the audio data, and compared their performances when using our multi-labeling approach. Moreover, we compared the performance of our approach, which analyzes four regions of interest simultaneously, with that of treating each region separately and independently. Our results support the use of simultaneous and spatially descriptive Mel-Spectrogram input data and labels with deep learning in classifying sound recordings with an average Hamming Loss of 0.10, and average F-Score of 0.90. It was reported that the diagnosis of normal breathing sounds through classical auscultation has an accuracy of about 20% [62], [65]. Our optimal average accuracy for sound type '0' (normal breathing) after testing was 84.8%. Although wheezing and crackling detection, with a diagnosis based on the use of a conventional stethoscope, had higher accuracy values of 85% and 67%, an automated classification of recordings of these sounds would still be beneficial. Our model would offer an accuracy of 88.8% for detecting wheezing (sound type '1') and 86.7% for detecting crackling (sound type '2'). Another possible application of our proposed method is to create an automated tool for detecting various patterns and coughs related to the onset of COVID-19 symptoms. Creating an application which can offer monitoring of breathing sounds can aid in determining the correct moment to offer medical intervention [103], [53]. Finally, autonomous detection of different types of breathing sounds in various chest locations concurrently can offer an improved assessment of the locality, severity and progression of pulmonary disease.

Chapter 4

Classification of fNIRS Finger Tapping Data with Multi-Labeling and Deep Learning

4.1 Introduction

The study of brain activation in Human Computer Interaction (HCI) and Brain-Computer Interfacing (BCI) lends itself well to the importance of detecting concurrent reactions to stimuli in several regions of interest. Motivated by making a contribution to HCI/BCI research, our novel multi-label/multi-class way of annotating input data is applied to Functional Near Infrared Spectroscopy (fNIRS) data. Making strides in the advancement of machines as an aid to people with disabilities and for rehabilitation drives us to continue applying our novel classification method, as described in Chapters 2 and 3, to another form of spatiotemporal data, channel readings from fNIRS probes. Gathering brain data through non-intrusive mediums along with finding robust ways of interpreting the data have been an integral part of this research. For instance, through image processing, a nose tracking cursor and an eye gazing interface give people with impaired motor function the ability

to communicate with others [119], [105]. In another work, a classification model was presented based on the use of Electroencephalogram (EEG) data to distinguish between mental counting and wrist rotations [106]. EEG data was also used by León [17] as she classified combinations of hand and feet movement imagery for robotic arm control. Other examples include utilizing a hybrid set of data such as a combination of EEG and fNIRS to detect different types of motor imagery. Such experiments included classifying the motor imagery related to the force and speed of right hand clenching [108] and motor imagery of left and right hand grasping [109]. Another area of interest in HCI/BCI is to explore the separate ability of fNIRS signals to represent movement and imagery. Three right foot soccer playing motion imageries (i.e. passing, stopping, shooting) were classified by Li et al. [107] with an average accuracy of approximately 79%. Right thumb and little finger physical tapping were distinguished with a validation accuracy of 97.17% by Woo et al. [110]. All the aforementioned studies offer ways of helping those who have compromised motor abilities by providing a possible means of communication with the outside world. Moreover, by increasing our knowledge of the brain and how to best capture and classify its signals gives hope to those who suffer from locked-in syndrome.

Functional magnetic resonance imaging (fMRI) represents the gold standard for brain measurement in cases where it is possible to place participants in the fMRI magnet with minimal motion permitted. Although fNIRS cannot measure deep brain structures like fMRI, it can take comparable measurements of hemodynamic responses across the brain cortex, and it can do so in naturalistic real-world environments due to portability, ease of set up, and decreased sensitivity to a subject's motion [111]. Thus, fNIRS could be a suitable device for BCI applications, where target users can wear the non-invasive device in their naturalistic environments. The motion sensitivity of fNIRS sensors is also smaller than that of EEG sensors and fNIRS can provide greater spatial resolution [112]. Although its acquisition of neural information is restricted to 1 cm below the surface of the brain, fNIRS has been a popular way of collecting brain data signals as it is capable of capturing

hemodynamic information in a non-restrictive and practical way while providing ample spatial and temporal resolutions.

The classification of finger tapping activities with the use of fNIRS data is a popular area of research [112], [113]. Finding a suitable interpretation of data collected by fNIRS sensors as it relates to a subject's physical or mental activity is key to making the data useful for HCI/BCI. In this chapter, we propose an approach that aims at capturing the spatiotemporal nature of the fNIRS data with a robust deep learning algorithm, which combines a Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). Moreover, assigning labels to the data which represent the spatial nature of the probe configuration can yield information about the location of activation. Therefore, by handling the readings acquired by the fNIRS probe channels similar to video frames, we can apply our novel multi-label/multi-class deep learning classifier, first introduced to detect different activity levels in various regions of interest simultaneously in video data [84]. In this case, finger tapping data would be formatted to represent two regions of interest in the brain, namely, the left and right motor cortices. Channel readings, based on the corresponding probe location, would be formatted to represent these two sides of the brain.

The data used in this study was collected at a University in the Western United States. The dataset contains 51 trials of index finger tappings from the right and left hands at two different frequencies. Each trial included sequences of both index fingers tapping at the same frequencies, single index finger tappings and rest periods. The different levels of finger tapping frequency (rest, 80 bpm, and 120 bpm) lend themselves well to a multi-class labeling scheme. Then, applying our novel spatial multi-labeling technique of designating two labels to the two sides of the brain, each of which is assigned a finger tapping frequency level, classification would be based on a multi-label/multi-class schema. This is a novel approach to multi-labeling since in most studies, binary labels are used for the presence (1) or absence (0) of each class in a multi-label [23], [79]. In these cases, the multi-label does not provide spatial information the way that our schema does.

We propose a unique and promising approach to classifying various frequency levels of index finger tapping $simultaneously$ with the use of a multi-label/multi-class labeling schema. Our novel multi-labeling approach provides a concurrent detection of different levels of brain activation in the two sides of the brain.

The spatiotemporal nature of the fNIRS signal acquired during finger tapping trials is captured during the training and validation of a Convolutional LSTM model. Different from our previous work described in Chapters 2 and 3, which focuses on video and recorded sound data, in this work (i) the labels are assigned to fNIRS data formatted to represent the probe configuration for the two sides of the brain; (ii) we employ and generate Shapley Additive Explanation values [138] to help explain the spatial characteristics of our Convolutional LSTM model; (iii) we plot the Shapley values on an image-like background, which represents the fNIRS channel layout used as data input. Applying Shapley values is notable as it can be used to view the structure of deep learning models (and these deep learning 'black boxes' can often be difficult to interpret) and to show the regions of the brain that were most important for different model predictions. Finger tapping has been heavily studied in the brain measurement domain, and there are known brain activations in the contralateral primary motor cortices, based on which finger is being tapped (left finger activates right primary motor cortex, and vice versa). We use our Shapley values to show that the most predictive channels in our deep learning model align with known spatial characteristics of the brain during finger tapping. By using our model explainability techniques on a benchmark finger tapping task, we demonstrate the potential of applying this approach to more complex classification tasks (e.g. classifying types of workload or emotional states from fNIRS data), where the use of Shapley values can help to explain the model, while also having the potential to add to the field of cognitive neuroscience, whereby complex interactions between interconnected brain regions could be identified with an explainability technique.

The organization of this chapter is as follows: We discuss related work in Section 4.2.

We then describe our network model and illustrate the formatting of our dataset, and labeling structure in Section 4.3. We present our experimental results along with further analysis of our proposed method with Shapley Additive Explanations (SHAP) values in Section 4.4. We then conclude the chapter in Section 4.5.

4.2 Related Work

Finger tapping abilities have been studied as a way of determining the progression of Parkinson's Disease [114], [115]. Similarly, other neuromuscular disabilities caused by cerebral palsy and stroke can be better understood with finger tapping exercises [116], [117]. This means that studying the relationship between the brain and finger tapping motions can contribute towards an improved understanding of neuromuscular impairment. Furthermore, by acquiring brain data signals non-intrusively during finger tapping exercises and building a robust classification model, can aid in the fields of HCI and BCI for people with compromised motor function. Training BCI applications on real, or imagined, finger tapping motions has been heavily studied in part because finger tapping has been found to result in consistent patterns of activation in brain areas involved in motor function. Specifically, right finger tapping shows reliable activation in the left primary motor cortex, and vice versa. Tapping both fingers simultaneously will result in both left and right primary cortices being activated (amongst a host of other regions that are implicated in the execution of motor function) [17], [152]. The study outlined in this paper leverages these known correlations between finger tapping and left/right primary cortices, which can be readily measured with the fNIRS modality.

The classification of brain signals which show greatest activation during finger tapping motions, can also lead to establishing a brain mapping based on sensor locations [119] thereby assisting in motor skill rehabilitation [118]. Moreover, the study of finger tapping motor imagery is instrumental in the fields of HCI/BCI [122], [123]. Correctly classifying

brain signals retrieved during motor imagery can provide a means for locked-in patients to communicate with their environments. Since fNIRS signals evoked during finger tapping imagery and execution are correlated [120], [124], an improvement in the interpretation of data collected during finger tapping movement can bring insight into its mental visualization.

Research based on fNIRS has gained popularity over the last several years [125], [126]. Compared to other neuroimaging technologies, such as fMRI and MEG, fNIRS is adaptable to real-world environments due to its portability and ease of use. Compared to EEG signals, fNIRS signals are less susceptible to motion artifact and provide a greater spatial resolution. The acquisition of fNIRS signals is non-intrusive. Using optical wavelengths between 650 nm and 1000 nm produced by the emitter probes, the detector probes detect the reflected light due to changes in oxygenated (OXY) and deoxygenated (DEOXY) hemoglobin concentrations at the cerebral cortex. The changes in hemoglobin oxygenation are a result of neural activity elicited by stimuli.

The study of neural activation due to finger tapping using fNIRS signals has shown interesting results. Bak et al. [127] performed a binary classification of left index finger tapping versus right index finger tapping with an SVM model and obtained an average accuracy of about 83%. Woo et al. [110] used a deep convolutional generative neural network to augment their data and trained a CNN model which produced accuracy results of 92.42% and 97.17% for thumb tapping and little finger tapping classifications, respectively. Nazeer et al. [121] used vector-based phase analysis features and a Linear Discriminant Analysis (LDA) model to distinguish left index finger tapping from right index finger tapping and also these two finger tappings from rest. The results, based on a sample size of 7 were 98.7% for the two-class distinction and 85.4% for the three-class distinction. Siddique et al. [128] used a Bayesian Neural Network to discern left index and right index finger tappings and obtained an average classification accuracy of 86.44% for 30 volunteers.

It is difficult to compare the results of the aforementioned studies since different mod-

els were used and different finger tapping exercises were conducted. All were based on fNIRS data and reported classification accuracies are robust. However, the disadvantage of using shallow learning (i.e. SVM, LDA) is the need to choose features that best fit the nature of the data. Nazeer et al. [121] reported robust classifications results with LDA but with a very small sample size of 7 subjects. Woo et al. [110] reported their robust results based on a slightly larger sample size of 11 subjects but with a deep learning CNN model. The latter study along with research by Trakoolwilaiwan et al. [129] and Wickramaratne at al. [130] have demonstrated that the CNN-based algorithm is suitable for learning patterns in raw fNIRS data, thereby eliminating the need for generating handcrafted features. Additionally, recent research has been conducted to support the use of LSTM networks in the classification of fNIRS data [131], [132]. Although the data was not finger tapping related, improvements in classification results were shown as the algorithm was able to capture the temporal characteristics of the data.

The monitoring of simultaneous activity based on spatiotemporal data has been studied for a variety of applications. For example, a Convolutional LSTM was used to predict the simultaneous demands of different modes of transportation in an urban environment [142]. Also, using magnetoencephalography (MEG) and SVM, four types of simultaneous bilateral hand movements were classified with average accuracies of 75% and 70% for physical and imagined movements, respectively [141]. Classification of concurrent events offers a snapshot view of the output states through time providing additional information related to the correlation of these actions. For example, this provides a means of exploring the neural correlates of finger tapping performed by fingers on different hands. Our interest in detecting simultaneous activity in spatiotemporal data necessitates the use of multi-labels. Multi-labeling has been used in video and other spatiotemporal data classification [31], [43]. However, most multi-labeling schemas in current research are based on selecting one of more descriptors (classes) and then referring to them as a multi-label. For example, the multi-labels in the YouTube-8M database [133] are annotations which

describe the contents of the video. Similarly, in combating noise pollution, environmental sounds, including simultaneous sounds from different sources, are tagged, thereby forming the multi-labels [79]. In the medical field, multi-labels have been used in classifying motor execution and imagery. For example, Olsson et al. [134] classified compound hand movements based on high density surface electromyography (HD-sEMG) recordings using a series of labels that describe the basic movements (i.e.: individual finger movements) needed to attain the final compound movement (i.e.: fist). Therefore, the basic movements defined the individual labels which were used to build the multi-labels. Also, León [17] used multi-labels to represent different combinations of hand and feet motor imagery captured by EEG data. One of the labels would be assigned a '1' based on the motion(s) detected and all other labels would then receive a '0'. The previously mentioned examples utilize a pool of classes to choose from in building multi-labels. This approach is a common type of multi-label/multi-class schema, To the best of our knowledge, the spatially descriptive multi-label/multi-class approach that we employ in our research has not been attempted. In our case, the labels represent spatial regions of interest. Each region of interest gets assigned one of many classes depending on the activity level in the area that is monitored.

Ensuring that labels are balanced is important to avoid classification skew which can impact results. Multi-labeling can magnify the problem of label imbalance [31]. Data augmentation is a technique which increases the number of samples, and helps balance labels [110], [135]. For example, when classifying different levels of activity in surveillance videos [84], as described in Chapter 2, we augmented the video data by creating additional videos based on our original video dataset through rotation of frame quadrants to increase samples of minority labels. Additionally, another approach to label and class balancing is to use resampling techniques, such as Random Undersampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE). SMOTENN [137], an interesting balancing technique, combines both oversampling and undersampling. More specifically, it combines SMOTE

oversampling (for the minority classes) with Edited Nearest Neighbors undersampling (for the majority classes). This balancing technique is utilized in our proposed method.

4.3 Proposed Method

By considering the two sides of the brain as the two regions of interest, it is important to visualize the fNIRS probe layout used in this study which is shown in Fig. 4.1.

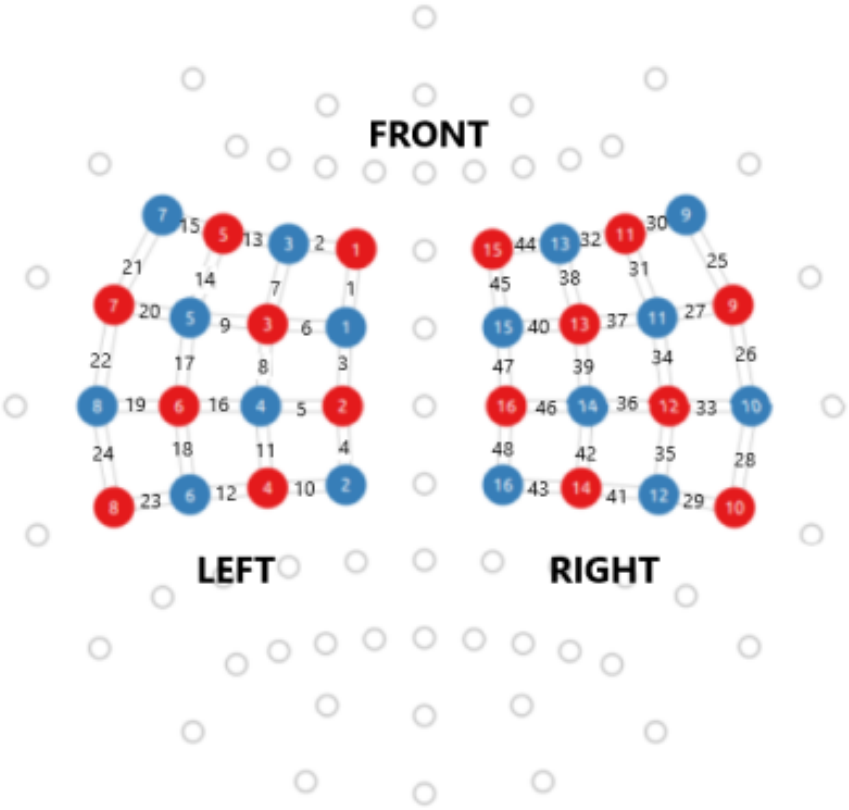


Figure 4.1: fNIRS Probe Layout, with Regions on the Left and Right Primary Motor Cortex Covered

Paying attention to the channels in between the sources (in red) and detectors (in blue), the OXY and DEOXY data channels in the fNIRS readings were separated into the two sections of interest, namely, left primary motor cortex and right primary motor cortex. Then, the channels in each respective region were rearranged to reflect the proximity of detectors to the sources. The final configuration is shown in Table 4.1.

LEFT MOTOR CORTEX						RIGHT MOTOR CORTEX					
CH_1	CH_14	CH_20	CH_6	CH_11	CH_24	CH_45	CH_31	CH_27	CH_40	CH_42	CH_28
CH_2	CH_15	CH_17	CH_3	CH_16	CH_23	CH_44	CH_30	CH_34	CH_47	CH_36	CH_29
CH_7	CH_21	CH_9	CH_4	CH_18	CH_12	CH_38	CH_25	CH_37	CH_48	CH_35	CH_41
CH_13	CH_22	CH_8	CH_5	CH_19	CH_10	CH_32	CH_26	CH_39	CH_46	CH_33	CH_43

Table 4.1: Channel Order Format To Reflect Probe Configuration in the Left and Right Regions of Interest

The two distinct spatial regions each correspond to a label assigned to an index finger as part of our multi-label. With this set-up, we are able to determine the types of activation taking place in each side of the brain simultaneously during index finger tappings.

In our labeling schema, each spatial region corresponds to an index finger which gets assigned one of three possible finger tapping frequencies: rest, 80 bpm, and 120 bpm. Our spatially descriptive multi-labels enable the recognition of concurrent finger tappings by the right and left index fingers simultaneously, and the identification of different rates of tapping through a $multi - class$ descriptor for each label. Our labeling schema, like most multi-labeling schemas, is more prone to label imbalance. That is why we chose SMOTENN, which uses hybrid oversampling/undersampling, to balance our labels. The effect of SMOTENN on the class sample distributions for our two labels is shown in Table 4.2.

Before SMOTENN	**Label #1**	**Label #2**
Class '0'	45.9%	47.1%
Class '1'	27.2%	26.3%
Class '2'	26.8%	26.6%
After SMOTENN		
Class '0'	34.5%	33.8%
Class '1'	32.9%	32.5%
Class '2'	32.6%	33.7%

Table 4.2: Class distribution before and after label balancing with SMOTENN

Our network model, dataset and the labeling structure are described in detail in the following subsections.

4.3.1 Network Model

Since the fNIRS signals acquired during index finger tappings are spatiotemporal, we chose to base our network model on a Convolutional LSTM to be able to detect the spatial and temporal properties of the data. The structure of our network model is shown in Fig. 4.2. The fNIRS input data is formatted into two regions of interest and labeled with our multi-labels. The network contains two 2D convolutional layers, with the first followed by a max pooling layer. Then, following a dropout layer, there is an LSTM layer and a dense layer which leads to a final output layer. An adaptive learning rate which decreased by a factor of 2 to a minimum of 10^{-6} was used along with binary cross-entropy loss and $rmsprop$ optimization. The adaptive learning rate and early stopping monitor validation loss so that adjustments are made to the learning rate if the validation loss does not decrease after two epochs and consequently, training automatically stops if the validation loss has stopped decreasing after three epochs. After training, the model was tested on $unseen$ fNIRS finger tapping .csv files formatted the same way as our input data (detailed next).

As illustrated in Fig. 4.2, the input to our proposed network model is formatted into two sections, wherein the left and right include the channels which capture the changes in hemoglobin concentrations for the left and right motor cortices for each participant. Specifically, the channels for each region of interest, with a rearranged channel order that conforms to the probe layout, were grouped into chunks of 50 samples with a sliding window of 10 samples. Batches of these groups were then used to train and validate the model.

4.3.2 Dataset

Our fNIRS data in this study was collected with a NIRx NirsSport2 device at a sampling rate of 10.2 Hz. We used the standard NIRX montage that covers the right and left primary motor cortices, and changes in hemoglobin concentrations were recorded by the Aurora fNIRS software. Data was bandpass filtered from 0.01 Hz and 0.5 Hz to remove noise. The modified Beer-Lambert Law was applied to convert the light intensities into data repre-

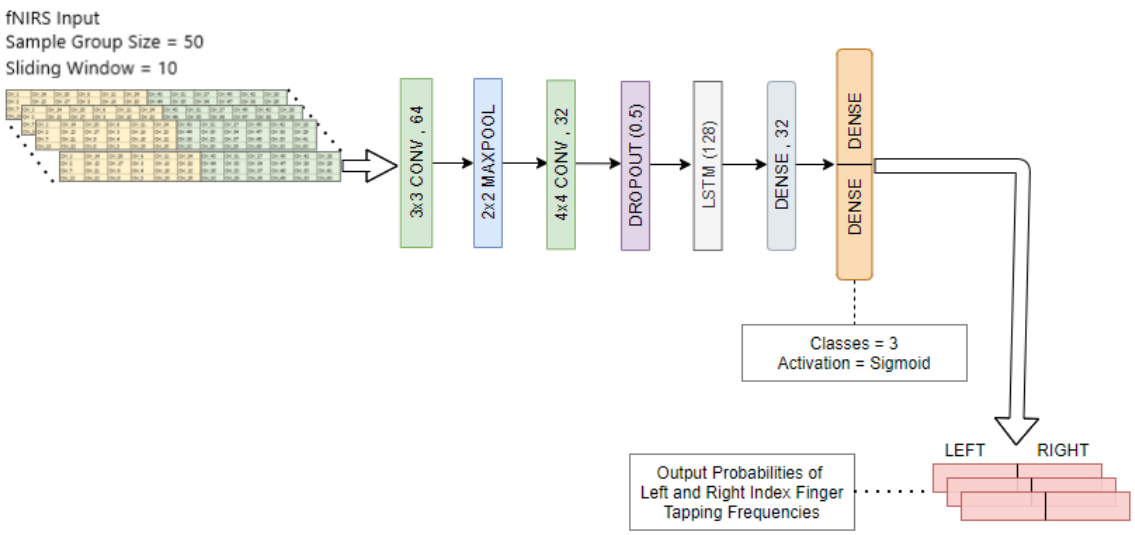


Figure 4.2: Proposed Deep Learning Model Structure

sentative of relative change in OXY and DEOXY hemoglobin. Z-score normalization was applied to each channel. For each sample (OXY and DEOXY), the channels belonging to the left motor cortex of the brain were rearranged and reshaped to correspond to the probe layout. Subsequently, after the channels for the right motor cortex were formatted accordingly, the two sets of data were concatenated as illustrated in Table 4.1. The final format of the comma-separated values (.csv) data files included a layer of 4x12 channel-ordered OXY data followed by a layer of 4x12 channel-ordered DEOXY data.

Each file included sequences of the following finger tapping combinations: both index fingers at rest; one index finger at rest and the other tapping at either 80 bpm or 120 bpm, both index fingers tapping at 80 bpm, and both index fingers tapping at 120 bpm. Therefore, there were seven possible multi-labels. Using two labels to spatially represent the left and right motor cortices, one of three classes characterizing the tapping frequency was assigned to each label. The SMOTENN [137] label balancing algorithm, a combination of SMOTE oversampling (for the minority classes) and Edited Nearest Neighbors undersampling (for the majority classes) was applied to the training and validation data before samples were grouped in sets of 50 with a sliding window of 10 samples. During the grouping process,

the label of the middle sample in each group at each location (region of interest), i.e. the label of the twenty-fifth sample, in this case, was used as part of the multi-label. With a total of 51 finger tapping data files, we randomly chose 46 files for training and validation. We used 80% of the 46 files for training, and kept the remainder for validation. The remaining 5 $unseen$ finger tapping files were then used for testing. This process was repeated seven times for cross-validation.

4.3.3 Proposed Multi-Labeling Structure

Given two spatial regions of interest, namely, the left and right motor cortices, we established a novel approach to multi-labeling. Due to the contralateral relationship between an index finger's tapping motion and motor cortex activation [153], the left index finger was assigned a label which represents the right motor cortex. The right index finger was assigned a label that represents the left motor cortex. The three tapping frequencies in our data were rest, represented with a '0', 80 bpm, represented with a '1', and 120 bpm, represented with a '2'. The definition for each type of finger tapping frequency that was used in our labeling schema is shown in Table 4.3.

LEVEL OF TAPPING	DEFINITION
0	Rest (No Tapping)
1	Index Finger Tapping at 80 bpm
2	Index Finger Tapping at 120 bpm

Table 4.3: Three Different Types of Finger Tapping Frequencies Are to Be Detected in Each of the Two Sides of the Brain.

As we rearranged and reshaped the fNIRS data into our two regions of interest (left and right) for each input file (for both OXY and DEOXY layers), we combined the corresponding labels (left index finger and right index finger) for our multi-labels. For example, suppose that for a particular sample of finger tapping data, the right and left regions of the brain are labeled to have left index finger tapping at 120 bpm (level of tapping = '2'); and

right index finger tapping is at rest (level of tapping = '0'), respectively. The multi-label of $[2, 0]$ would be assigned to the pertinent sample of fNIRS data as shown in Table 4.4.

LABEL #1 Left Index Finger Level of Tapping	LABEL #2 Right Index Finger Level of Tapping
2	**0**

Table 4.4: Example of Our Spatially and Tapping-Level Descriptive Multi-Label Schema

The finger tapping combinations in the fNIRS database resulted in seven possible multi-labels, namely, $[0, 0]$, $[0, 1]$, $[0, 2]$, $[1, 0]$, $[2, 0]$, $[1, 1]$, and $[2, 2]$. For each cross-validation run, testing was performed on the 5 finger tapping files that were not in the training and validation set. After testing, predictions were made and each type of finger tapping level was assigned a probability for the respective region of interest. The class of the finger tapping frequency ('0', '1', or '2') with the highest probability was then chosen as part of the predicted multi-label.

4.3.4 Evaluation Criteria

We report our results based on the micro metrics of F-Score and Accuracy. Additionally, we use the measure of Hamming Loss, a common metric used in multi-label research. Hamming Loss offers an overall look at a classifier's prediction error [54]. It is an instance-based metric since it is based on the entire multi-label prediction for each time period of testing data. On the other hand, micro-averaged F-Score and Accuracy are label-based metrics since the label assigned to each region of interest in the testing results is considered separately [34], [37], [54].

i) Hamming Loss.

For each instance, we compared the group of two predicted classes of finger tapping in our multi-label with the ground truth label assignments. The average Hamming Loss for each

testing finger tapping trial was calculated as shown in Eq. (4.1), where S represents the number of seconds in a testing finger tapping trial, and $p_{i,j}$ and $g_{i,j}$ indicate the predicted level of tapping and the ground truth level of tapping, respectively. Therefore, for each label within a multi-label, a mismatch is assigned a 1. These values are then added and averaged over the product of the number of labels (2) in a multi-label and the time span of the recording in seconds (S).

$$\frac{1}{2S}\sum_{i=1}^{S}\sum_{j=1}^{2}[if(p_{i,j}=g_{i,j},0,1)] \tag{4.1}$$

ii) Micro-Averaged F-Score and Accuracy.

In order to provide a complete set of performance criteria, we also consider label-based (i.e. single side of the brain) metrics, such as Micro-averaged F-Score and Accuracy.

Since our labels can be assigned one of three classes ('0', '1', or '2') to represent three levels of finger tapping frequency, it is best to perform a micro-average of these parameters. For instance, instead of taking an average of the calculated Accuracy for each level of finger tapping, we determine the overall Accuracy for all levels of tapping at once. The calculations of Micro-Averaged F-score (μA_Fscore), and Micro-Averaged Accuracy (μA_Accuracy) are shown in Equations (4.2), and (4.3), respectively.

$$\mu A\_\text{FScore} = \frac{\sum_{i=0}^{2}(2*TP)_i}{\sum_{i=0}^{2}((2*TP)_i+(FP)_i+(FN)_i)} \tag{4.2}$$

$$\mu A\_\text{Accuracy} = \frac{\sum_{i=0}^{2}((TP)_i+(TN)_i)}{\sum_{i=0}^{2}((TP)_i+(TN)_i+(FP)_i+(FN)_i)} \tag{4.3}$$

4.4 Experimental Results

Our goal was to detect different levels of finger tapping in two different spatial sides of the brain simultaneously. Through our novel multi-labeling schema and deep learning-based algorithm, we built a model to be able to perform automated classification of spatiotemporal finger tapping data and determine the simultaneous types of tapping taking place by both index fingers. The training/validation dataset included 46 fNIRS data files based on finger tapping trials. Since the SMOTENN algorithm was applied to this data, the number of instances for different classes, was void of skew and imbalance. After label balancing, approximately 200 minutes of multi-labeled fNIRS finger tapping data was available for training and validation. Optimal results have been obtained with a window size of 5 seconds and a 20% overlap. During training and validation, we monitored training to make sure that validation loss was always smaller than training loss to prevent any possible overfitting. Once the validation loss did not improve for three epochs with an adaptable learning rate, training/validation would end.

After training on 46 trials of fNIRS data, the classifier is then tested on an additional 5 similarly configured finger tapping files captured with the NIRx NirsSport2 device. Results were compared to the database's ground truth label annotations. We determined the overall Hamming Loss using Eq.(4.1) by calculating the average for all groups of two labels for the 5 testing finger tapping files. We cross-validated our model seven times and our results are shown in Table 4.5. These values indicate that our classifier is capable of making correct predictions for our multi-labels for the testing finger tapping data an average of about 80% of the time.

Additional metrics, namely Micro-Averaged F-Score, and Accuracy values, were also calculated for each region of interest (i.e. side of the brain). Results are presented in Table 4.6. A discussion of these results can be found in Sec. 4.4.2.

Cross-Validation Run	1	2	3	4	5	6	7
Average Hamming Loss	**0.185**	**0.209**	**0.235**	**0.217**	**0.161**	**0.211**	**0.188**

Table 4.5: Average Hamming Loss of 5 Testing fNIRS Finger Tapping Files for 7 Cross-Validation Runs

Cross - Validation	**Micro-Averaged F-Score**		**Micro-Averaged Accuracy**	
Run #	Label #1	Label #2	Label #1	Label #2
1	0.819	0.823	0.810	0.818
2	0.803	0.787	0.790	0.786
3	0.775	0.748	0.750	0.752
4	0.793	0.819	0.804	0.812
5	0.838	0.839	0.829	0.841
6	0.760	0.783	0.765	0.759
7	0.827	0.803	0.814	0.811
Average	**0.802**	**0.800**	**0.796**	**0.797**

Table 4.6: Right (Label #1) and Left (Label #2) Sides of the Brain Average Label Metrics for the 5 Testing fNIRS Finger Tapping Files for 7 Cross-Validation Runs

4.4.1 Visualizing Our Proposed Method with SHAP

Motivated by our promising Hamming Loss, Micro-Averaged F-Score and Micro-Averaged Accuracy values, our next goal was to gain a further understanding of our network model with Shapley Additive Explanations (SHAP) values. SHAP, with its basis in game theory [138], is a way of illustrating model interpretation by assigning impact values on learned features (in the case of deep learning) as they relate to a model's predictions [139], [140]. Using the deep explainer which is specialized for neural network models, we generated the SHAP values to visualize how our model handles the channel readings when making predictions. The SHAP values represent a channel's marginal contribution to the output class predictions.

For example, for the prediction of [0,0], meaning that both the left and right index fingers are at rest, the SHAP values show that the channels on the right side (top right) and

also left side (bottom left) have the highest impact values for '0' (rest). Using the channel layout illustrated in Table 4.1 as the background, the channels with the greatest positive effect on the classification result are coded in the brightest color of red in Fig. 4.3.

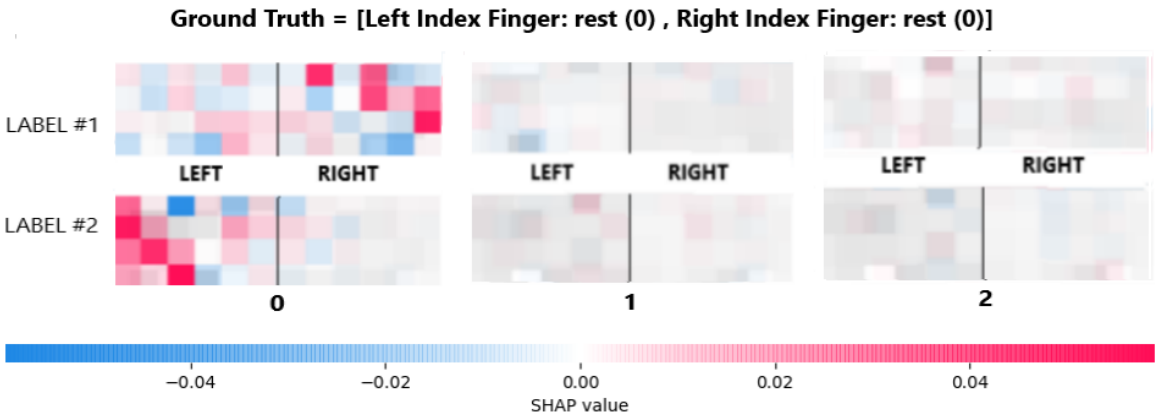


Figure 4.3: SHAP Values for Multi-Label [0,0]

A large body of research has found that real, or imagined, finger tapping has reliably been found to activate the primary motor cortex. Tapping the right finger activates the left primary motor cortex, and vice versa for the left finger. Tapping of both fingers has been found then to activate both regions simultaneously [17], [152].

When comparing these SHAP values to the channel layout in Table 4.1, the specific channels with the largest positive impact can be located in this table. This is shown in Fig. 4.4.

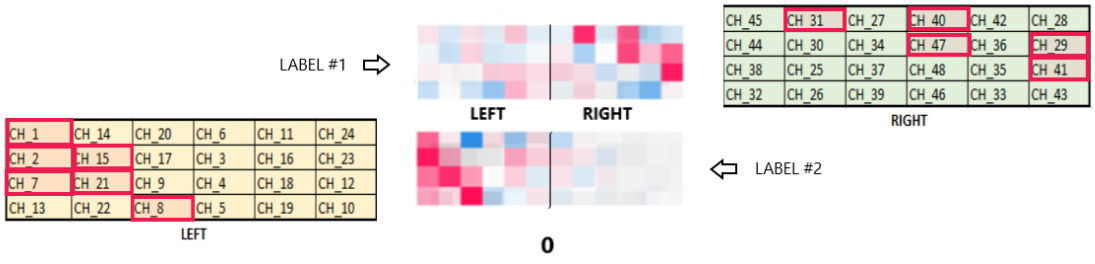


Figure 4.4: Mapping SHAP Values to Specific Channels for [0,0]

Therefore, this allows a visualization of the channels with the highest positive prediction impact on the actual probe layout as shown in Fig. 4.5.

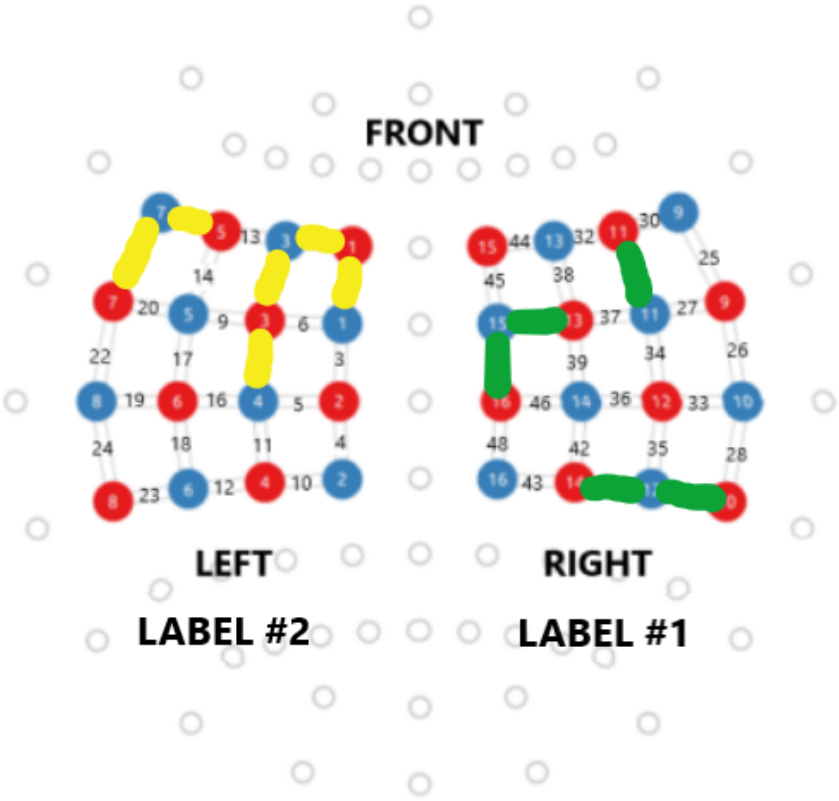


Figure 4.5: Highlighting Channels with Highest Positive SHAP Values on Probe Layout for [0,0]

Subsequently, the SHAP values for the multi-labels [0,1] and [0.2] are shown in Fig. 4.6 and Fig. 4.7, respectively.

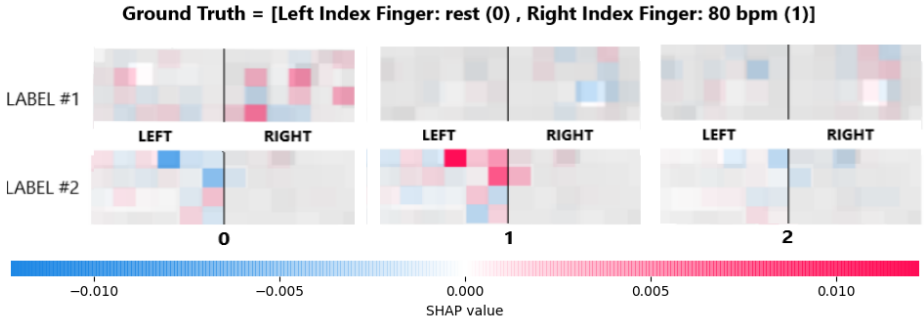


Figure 4.6: SHAP Values for Multi-Label [0,1]

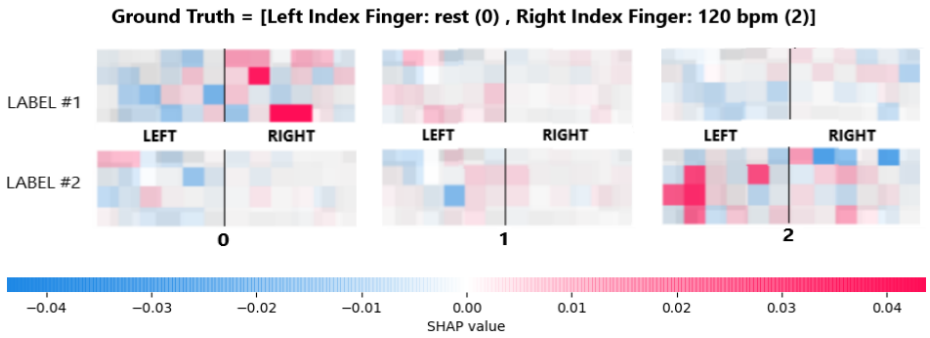


Figure 4.7: SHAP Values for Multi-Label [0,2]

In Fig. 4.6, the multi-label [0,1] represents the simultaneous motion of the right index finger tapping at 80 bpm while the left index finger is at rest. Similarly, in Fig. 4.7, the multi-label [0,2] represents the left index finger at rest while the right index finger is tapping at 120 bpm. In both results, as expected, a greater number of positive SHAP values are shown in the right side channels for a prediction of '0' for Label #1. The second label shows the brightest positive SHAP values in the left side channels for predictions of '1' and '2' for multi-labels [0,1] and [0,2], respectively. It is interesting to note the difference in scale between Fig. 4.6 and Fig. 4.7 which seems to indicate that the activation levels for the simultaneous detection of an index finger at rest while the other one is tapping at 80 bpm is lower than when the tapping is at a higher rate, 120 bpm.

A similar difference in scale persists when comparing the SHAP values for the multi-labels [1,0] and [2,0] are shown in Fig. 4.8 and Fig. 4.9, respectively.

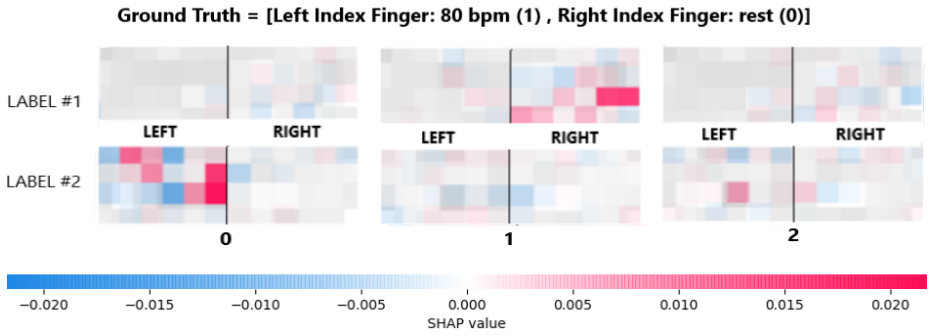


Figure 4.8: SHAP Values for Multi-Label [1,0]

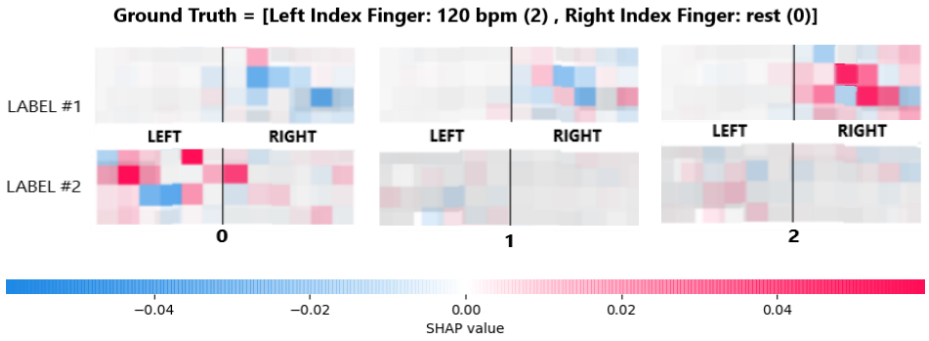


Figure 4.9: SHAP Values for Multi-Label [2,0]

In Fig. 4.8, the left finger is tapping at 80 bpm while the right finger is at rest. In Fig. 4.9, the left finger is tapping at the higher rate, 120 bpm, and the right finger is at rest. Corresponding channels from the two sides of the motor cortex show the brightest SHAP values, as expected.

Finally, the SHAP values for the multi-labels which represent both index fingers tapping at the same rate, namely [1,1] and [2,2], are shown in Fig. 4.10 and Fig. 4.11, respectively. As anticipated, simultaneous activation is seen in the channels representing both sides of the motor cortex for the same classes in the two figures.

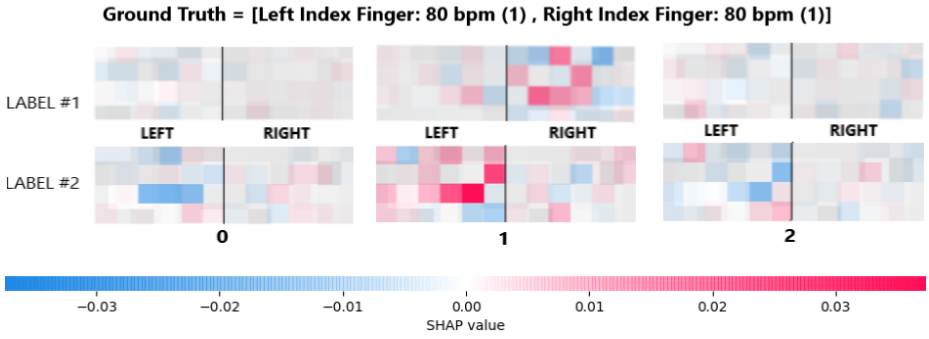


Figure 4.10: SHAP Values for Multi-Label [1,1]

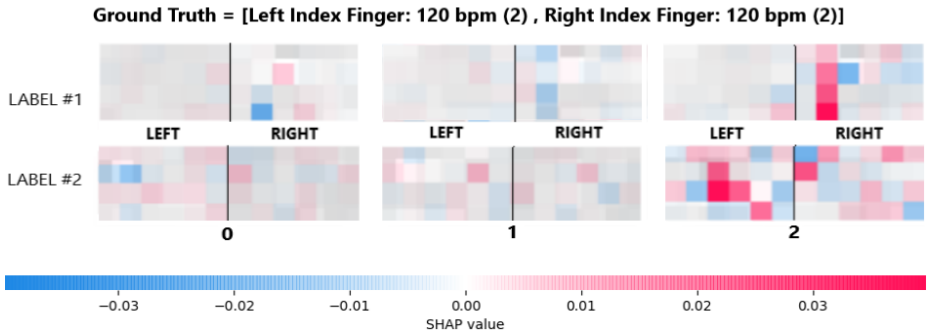


Figure 4.11: SHAP Values for Multi-Label [2,2]

Overall, the SHAP values support the well-established contralateral correlation between the sides of motor cortex activation and finger tapping hand [153]. Also, higher levels of activation are exhibited when one finger is at rest and the other finger is tapping at a class level of '2' as compared to a class level of '1' as seen in Fig. 4.6 through Fig. 4.9. The smaller scale for the resulting 80 bpm SHAP values (Fig. 4.6 & Fig. 4.8) as compared to the 120 bpm (Fig. 4.7 & Fig. 4.9) supports this. Therefore, this study reinforces the use of SHAP values as a way of understanding a deep learning "black box" model. The locations of the brighest positive SHAP values provide an affirmative backing of León's [17] statement that: "During the preparation of a motor task certain cortical and subcortical regions are activated; contrarily, in the course of motor imageries most of the activity is found within the primary motor cortex over the corresponding contralateral hemisphere."

4.4.2 Further Discussion of Results

We have presented the predictions for all seven of our multi-labels using SHAP values and was able to demonstrate that the positive impact values (in red) were located in corresponding sides of the brain. For example, the SHAP values for the multi-label [2,0] are showing the highest activation on the right motor cortex for class '2' and left motor cortex for class '0'. These map correctly to the left index finger tapping label of '2' and the right index finger tapping label of '0'. This supports our model's weight assignments for the learned

features after it completes training and validation.

Our Hamming Loss results show that on average, for the 5 testing files and seven cross-validation runs, correct predictions were made 80% of the time. The corresponding micro-averaged F-score and accuracy were also close to 80% in both cases. This outcome is similar to the accuracy obtained by Bak et al. [127] who obtained an average accuracy of about 83% when classifying fNIRS data based on left index finger tapping versus right index finger tapping with an SVM model. However, their model was meant to distinguish between the two types of tapping separately and not simultaneously, as in our approach with a multi-label. Moreover, classification was binary and not multi-class. Another study as reported by Woo et al. [110] showed that augmentation followed by training a CNN model resulted in classification results of 92.42% and 97.17% for thumb tapping and little finger tapping, respectively. These results are based on a simpler classification than ours. Once again, a binary classifier was used to distinguish the tappings of two different fingers but, in this case, the fingers were on the same hand. Nazeer et al. [121] used a very small sample size of 7 subjects in support of classification accuracy results of 98.7% when distinguishing right index finger tapping from left index finger tapping and 85.4% when also including the state of rest. The latter result is closer to our approach since it comprises of classifying right index finger tapping, left index finger tapping, and rest. However, it does not include different tapping frequencies and multi-labels. Similarly, Siddique et al. [128] discerned left index and right index finger tappings using a binary classifier and obtained an average classification accuracy of 86.44%. In general, it is non-commensurate to compare our accuracy results to the aforementioned research since the models in question were performing binary and single label types of classifications.

Our proposed approach introduces a novel spatial multi-labeling schema which uses two labels to represent the two sides of the brain. As both labels are classified simultaneously, one of three finger tapping frequencies can be assigned to each label. As the SHAP values illustrate, our model is able to learn the correct features of the fNIRS data which

contribute to the class predictions of our multi-labels. Finally, our multi-labeling classification was able to benefit from the use of the LSTM, known for its ability to detect label dependencies [101], [43], [102].

4.5 Conclusion

Providing a means of communicating with the outside world for those who have compromised motor function has played an important role in HCI/BCI research. Gaining a better understanding of how to capture the information in fNIRS brain signals to autonomously detect different levels of simultaneous index finger tapping frequencies can contribute to this endeavor. We have presented a promising approach to multi-labeling spatiotemporal data to detect different classes in two regions of interest $concurrently$. In this case, we have applied our approach to detect three different levels (rest, 80 bpm, and 120 bpm) of finger tapping for both index fingers at the same time. The spatial aspect of our fNIRS data, formatted to reflect probe layout, is captured with CNNs and the temporal one is captured with an LSTM. Our novel multi-labeling technique enables us to classify activity on both sides of the brain simultaneously with our network. Our network's SHAP values support its ability to choose appropriate spatial features (which importantly aligns with the known spatial characteristics of finger tapping on the primary motor cortex) when making predictions. By using Shapley's model explainability technique on these benchmark finger tapping tasks, we demonstrate the potential of applying this approach to more complex classification tasks (e.g. classifying types of workload or emotional states from fNIRS data), where the use of Shapley values can help to explain the model, while also having the potential to add to the field of cognitive neuroscience. There is a need in the AI domain to build transparent and explainable AI, and with this paper we demonstrate one way that this can be done for deep learning on high density fNIRS data.

Chapter 5

Conclusion and Future Work

What began as an interest in finding ways of classifying emotions with webcam video footage and then fNIRS data, developed into researching how to optimize models, which detect simultaneous activity in different regions of interest in different kinds of spatiotemporal data. In Chapter 2, the data was a set of two types of surveillance videos. Video frames were treated as having four regions of interest (i.e. quadrants) and different levels of activity were to be monitored in these quadrants simultaneously. In order to accomplish this, a novel multi-labeling/multi-class system was developed and a model based on a ConvLSTM was designed. In Chapter 3, the spatiotemporal data was sound data, specifically, respiratory sound recordings from different auscultation sites. The sites defined the regions of interest/multi-labels and four contrasting types of sounds were classified simultaneously with a model based on a combination of CNN and LSTM. Finally, in Chapter 4, we circled back to working with fNIRS data. In this case, the spatiotemporal data due to finger tapping stimuli was to be classified to detect the rates of tapping of the two index fingers concurrently. The regions of interest were the two sides of the brain and the model was a CNN/LSTM one.

At the core of this research, spatiotemporal data was handled spatially through multi-labeling. The number of regions of interest (i.e. labels) was defined based on the chosen

spatial resolution and characteristics of the dataset. For example, surveillance video frames were labeled according to a four quadrant resolution. This was a fitting choice because there were four main entry/exit points in the videos. Also, a group of four labels was used to represent the respiratory sound recordings in four auscultation sites. Making the choice of using the data from four sites was mostly based on maintaining a well-balanced dataset. Adding additional data from other ausculation sites and increasing the number of labels would have introduced unwanted skew. Similarly, using two labels to represent the two sides of the brain's motor cortex was the appropriate resolution for the fNIRS finger tapping data. Therefore, the number of regions of interest and multi-labels in our approach can be adapted to best suit the context of the spatiotemporal data. Also, when applicable, data was to be reformatted to fit the desired spatial structure.

Another aspect of our approach, which can be adjusted to befit the setting of a spatiotemporal dataset, is the number of possible classes to be defined. In the case of surveillance videos, since three different levels of activity are to be detected in four spatial quadrants in Chapter 2, three classes are defined to represent these levels. Likewise, to depict three types of abnormal breathing sounds and healthy breathing, four classes are designated in the classification of respiratory sounds in Chapter 3. Finally, in Chapter 4, three finger tapping rates necessitated the use of three classes. In all three studies, it was imperative to maintain label/class balance for optimal classification performance. In the first study, custom data augmentation was performed by rotating and flipping frame quadrants to increase the number of samples which included minority classes. In the second study, the data selected from the Kaggle Respiratory Sounds database [57] (a.k.a. the ICBHI 2017 Challenge Respiratory Sound Database [73]) resulted in a balanced dataset. In the third case, SMOTENN, an algorithm based on a combination of SMOTE oversampling and Edited Nearest Neighbors undersampling, was applied to finger tapping dataset.

All three studies produced simultaneous classification predictions of class assignments for the multi-label outputs. The network models, deep in nature, had varying hyperparam-

eters to adapt to the characteristics of the spatiotemporal datasets. The core structure of the first network was a ConvLSTM. The basis of the other two networks were separate CNN and LSTM layers. These slightly different structures were determined empirically with respect to training and validation accuracy and F-Score measures. The ConvLSTM, which embeds convolutional operators within an LSTM cell, has been successfully used for action recognition/detection in videos [143], [144], [145]. This supports our promising results for the detection of different levels of activity in videos as reported in Chapter 2. The decisions to use CNN/LSTM structures for the studies described in Chapters 3 and 4 were also based on the success of training and validation results. In general, cascades of CNN(s) and LSTM(s) are well suited for audio and fNIRS signal classification [146], [147], [130], [148]. Therefore, the decision to use ConvLSTM or a sequence of CNN(s) and LSTM(s) is dependent on how well the model is able to learn the important characteristics of the spatiotemporal data as it goes through training and validation. This is ultimately determined by whether the internal convolutions embedded in the LSTM cells (ConvLSTM) are able to catch the spatial features of the data while working on extracting the temporal features or whether handling these two processes independently conforms better to the data.

The novel approach introduced in this research can be adapted to many types of spatiotemporal data. A spatially descriptive multi-labeling schema is suitable for applications which benefit from having the ability to detect different levels and types of simultaneous activity. In the first study with videos, the advantage of this type of classification could be seen in security-type systems. Being able to spatially detect different levels of activity concurrently can also be applied to the medical field whereby autonomous classification of various types of abnormal respiratory or heart sounds can assist in the determination of the severity of disease. Finally, simultaneous detection of varied levels of activity in different regions of the brain can aid in finding the neural correlation between the stimulus and how the brain responds to it. Our research can bring benefit to many different types of fields.

Domain experts in security systems, pulmonology along with other medical fields, and neuroscience can specify a preferred number of regions of interest and classes and a customized model can be developed. Our approach can offer further insight into the implications of detecting simultaneous activation in the brain for the neuroscience field. Likewise, providing impactful diagnostic techniques by being able to classify different types of sounds in the lungs or other organ simultaneously could make a contribution to the medical field.

Future work is planned for further exploration into applying this approach to other spatiotemporal datasets. For example, it will be interesting to continue the analysis of sound data but, in this case, cardiac sounds, to classify different types of murmurs which may be difficult to detect through traditional auscultation. Also, the approach presented in this research would be suitable for the investigation of how to determine different cognitive load levels based on fNIRS data gathered while volunteers perform various mental tasks.

Bibliography

[1] N. Sommer, L. Hirshfield, and S. Velipasalar, Our Emotions as Seen through a Webcam, International Conference on Augmented Cognition, Springer; 2014. p. 78–89.

[2] S. Hiwa, K. Hanawa, R. Tamura, K. Hachisuka, and T. Hiroyasu, Analyzing Brain Functions by Subject Classification of Functional Near-Infrared Spectroscopy Data Using Convolutional Neural Networks Analysis, Computational Intelligence and Neuroscience, Hindawi Publishing Corp.; 2016. p. 3.

[3] T. Hiroyasu, K. Hanawa, and U. Yamamoto, Gender classification of subjects from cerebral blood flow changes using Deep Learning, 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE; 2014. p. 229–233.

[4] G. Huve, K. Takahashi, and M. Hashimoto, Brain activity recognition with a wearable fNIRS using neural networks, 2017 IEEE International Conference on Mechatronics and Automation (ICMA), IEEE; 2017. p. 1573–1578.

[5] J. Hennrich, C. Herff, D. Heger, and T. Schultz, Investigating deep learning for fNIRS based BCI, Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE; 2015. p. 2844–2847.

[6] , D. Bandara, L. Hirshfield, and S. Velipasalar, Classification of affect using deep learning on brain blood flow data, Journal of Near Infrared Spectroscopy, Vol. 27, No. 3, SAGE Publishing; 2019, p. 206–219.

[7] G. Sreenu, and M. Durai, Intelligent video surveillance: a review through deep learning techniques for crowd analysis, Journal of Big Data, Vol. 6, Springer; 2019.

[8] F. Jiang, J. Yuan, S. Tsaftaris, and A. Katsaggelos, Anomalous video event detection using spatiotemporal context, Computer Vision and Image Understanding, Vol. 115, No. 3, Elsevier; 2011, p. 323–333.

[9] , A. Babiker, I. Faye, A. Malik, Non-conscious Behavior in Emotion Recognition, 2013 IEEE 9th International Colloquium on Signal Processing and its Applications, 2013, p. 258–262.

[10] M. Guerriero, P. Willett, S. Coraluppi, and C. Carthel, Radar/AIS data fusion and SAR tasking for maritime surveillance, 2008 11th International Conference on Information Fusion, IEEE; 2008, p. 1–5.

[11] A. Renga, M. Graziano, M. D'Errico, A. Moccia, and A. Cecchini, SAR-based sea traffic monitoring: a reliable approach for maritime surveillance, SAR Image Analysis, Modeling, and Techniques XI, Vol. 8179, International Society for Optics and Photonics; 2011.

[12] S. Franconeri, B. Alvarez, and P. Cavanagh, Flexible cognitive resources: competitive content maps for attention and memory, Trends in cognitive sciences, Vol. 17, No. 3, Elsevier; 2013, p. 134–141.

[13] Z. Yuan, and X. Lin, Using fNIRS to identify the brain activation and networks associated with English versus Chinese simultaneous interpreting, Clinical and Translational Neurophotonics, Vol. 10864, International Society for Optics and Photonics, 2019.

[14] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, Event detection and recognition for semantic annotation of video, Multimedia Tools and Applications, Vol. 51, No. 1, Springer; 2011, p. 279–302.

[15] C. Zhao, J. Wang, J. Li, and H. Lu, Automatic group activity annotation for mobile videos, Multimedia Systems, Vol. 23, No. 6, Springer; 2017, p. 667–677.

[16] AA. Liu, Z. Shao, Y. Wong, J. Li, Y. Su, and M. Kankanhalli, Mohan, LSTM-based multi-label video event detection, Multimedia Tools and Applications, vol. 78, Springer; 2019, p. 677–695.

[17] C. Lindig León, Multilabel classification of EEG-based combined motor imageries implemented for the 3D control of a robotic arm, URL: https://tel.archives-ouvertes.fr/tel-01549139, NUMBER: 2017LORR0016, Université de Lorraine, 2017, HAL Id : tel-01549139, version 1.

[18] Y. Zhu, S. Liu, and S. Newsam, Large-scale mapping of human activity using geo-tagged videos, Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM; 2017.

[19] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, Learning to track for spatio-temporal action localization, Proceedings of the IEEE international conference on computer vision, IEEE; 2015, p. 3164–3172.

[20] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE; 2015, p. 2625–2634.

[21] S. Saha, G. Singh, M. Sapienza, P. Torr, and F. Cuzzolin, Deep learning for detecting multiple space-time action tubes in videos, arXiv preprint arXiv:1608.01529, 2016.

[22] Z. Yang, J. Gao, and R. Nevatia, Ram, Spatio-temporal action detection with cascade proposal and location anticipation, arXiv preprint arXiv:1708.00042, 2017.

[23] K. Sozykin, S. Protasov, A. Khan, R. Hussain, and J. Lee, Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks, 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), IEEE; 2018, p. 146–151.

[24] L. Constantine, G. Badaro, H. Hajj, W. El-Hajj, L. Nachman, M. BenSaleh, and A. Obeid, A Framework for Emotion Recognition from Human Computer Interaction in Natural Setting, 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD 2016), Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2016), 2016.

[25] M. Monfort, K. Ramakrishnan, D. Gutfreund, and A. Oliva, A Large Scale Multi-Label Action Dataset for Video Understanding, 2018 Conference on Cognitive Computational Neuroscience, doi: 10.32470/CCN.2018.1137-0.

[26] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei,Every moment counts: Dense detailed labeling of actions in complex videos, International Journal of Computer Vision, volume 126, number 2-4, Springer; 2018, p. 375–389.

[27] J. Ray, H. Wang, D. Tran, Y. Wang, M. Feiszli, L. Torresani, and M. Paluri, Scenes-Objects-Actions: A Multi-Task, Multi-Label Video Dataset, Proceedings of the European Conference on Computer Vision (ECCV), 2018, p. 635–651.

[28] O. Dekel, and O. Shamir. Multiclass-multilabel classification with more classes than examples, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, p. 137–144.

[29] P. Mehta, A.Y. Lee, C. Lee, M. Balazinska, and A. Rokem, Multilabel multiclass classification of OCT images augmented with age, gender and visual acuity data,

bioRxiv, Cold Spring Harbor Laboratory; 2018, p. 316–349.

[30] C. Mercan, S. Aksoy, E. Mercan, L.G. Shapiro, D. L. Weaver, and J.G. Elmore, Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images, Transactions on Medical Imaging Vol. 37(1), IEEE; 2017, p. 316-325.

[31] F. Herrera, F. Charte, A. Rivera, and M. Del Jesus, Multilabel classification, Multilabel Classification, Springer; 2016, p. 17–31.

[32] M. Hoai, Z. Lan, and F. De la Torre, Joint segmentation and classification of human actions in video, CVPR 2011, IEEE; 2011, p. 3265–3272.

[33] Z. Daniels, and D. Metaxas, Addressing imbalance in multi-label classification using structured hellinger forests, Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[34] G. Tsoumakas, E. Mencıa, I. Katakis, S. Park, and J. Fürnkranz, On the combination of two decompositive multi-label classification methods, Workshop on Preference Learning, ECML PKDD, Citeseer volume 9; 2009, p. 114–133.

[35] J. Read, A. Puurula, and A. Bifet, Multi-label classification with meta-labels, 2014 IEEE international conference on data mining, IEEE; 2014, p. 941–946.

[36] D. Costa Júnior, E, Paiva, J. Silva, and R. Cerri, Label Powerset for Multi-label Data Streams Classification with Concept Drift, 2017.

[37] M. Zhang, and Z. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering, Vol 26(8), 2013, p. 1819–1837.

[38] F. Charte, A. Rivera, M. del Jesus, and F. Herrera, Addressing imbalance in multilabel classification: Measures and random resampling algorithms, Neurocomputing Vol 163, Elsevier; 2015, p. 3–16.

[39] F. Charte, A. Rivera, M. del Jesus, and F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, Knowledge-Based Systems Vol 89, Elsevier; 2015, p. 385–397.

[40] A. Shustanov, and P. Yakimov, CNN design for real-time traffic sign recognition, Procedia engineering Vol 201, Elsevier; 2017, p. 718–725.

[41] M. Rad, A. von Kaenel, A. Droux, F. Tieche, N. Ouerhani, H. Ekenel, and J. Thiran, A computer vision system to localize and classify wastes on the streets, International Conference on Computer Vision Systems, Springer; 2017, p. 195–204.

[42] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, Cnn: Single-label to multi-label, aarXiv preprint arXiv:1406.5726, 2014.

[43] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, Cnn-rnn: A unified framework for multi-label image classification, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 2285–2294.

[44] J. Medel, and A. Savakis, Anomaly detection in video using predictive convolutional long short-term memory networks, arXiv preprint arXiv:1612.00390, 2016.

[45] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, Dense-captioning events in videos, Proceedings of the IEEE international conference on computer vision, 2017, p. 706–715.

[46] W. Chu, F. De la Torre, and J. Cohn, Learning spatial and temporal cues for multi-label facial action unit detection, 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE; 2017, p. 25–32.

[47] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, 2015, p. 802–810.

[48] B. Zapata-Impata, P. Gil, and F. Torres, Learning Spatio Temporal Tactile Features with a ConvLSTM for the Direction Of Slip Detection, Sensors, Vol. 19, No. 3, Multidisciplinary Digital Publishing Institute, 2019.

[49] Y. Li, A deep spatiotemporal perspective for understanding crowd behavior,IEEE Transactions on Multimedia, Vol. 20, No. 12, IEEE; 2018, p. 3289–3297.

[50] C. Olah, Understanding lstm networks, http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015.

[51] The VIRAT Video Dataset, http://www.viratdata.org/. Accessed 21 February 2020.

[52] H. Sager, and W. Hoff, Pedestrian detection in low resolution videos,IEEE Winter Conference on Applications of Computer Vision, IEEE; 2014, p. 668–673.

[53] M. Zhai, L. Chen, J. Li, M. Khodabandeh, and G. Mori, Object detection in surveillance video from dense trajectories, 2015 14th IAPR International Conference on Machine Vision Applications (MVA), IEEE; 2015, p. 535–538.

[54] R. Pereira, A. Plastino, B. Zadrozny, and L. Merschmann, Correlation analysis of performance measures for multi-label classification, Information Processing & Management, Vol. 54, No. 3, Elsevier; 2018, p. 359–369.

[55] Y. Dai, Y. Li, and S. Li, Multi-label learning for concept-oriented labels of product image data, Image and Vision Computing, 2019; p. 103821.

[56] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, Multi-label classification of music into emotions, ISMIR, Vol. 8, 2008, p. 325–330.

[57] Respiratory Sound Database, https://www.kaggle.com/vbookshelf/respiratory-sound-database. Accessed 23 March 2020.

[58] J. Lee, J. Park, K. Kim, L. Keunhyoung, and J. Nam, Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification, Applied Sciences, vol.8, Multidisciplinary Digital Publishing Institute; 2018, p. 150.

[59] CNN: Detection of wheezes and crackles, https://www.kaggle.com/eatmygoose/cnn-detection-of-wheezes-and-crackles. Accessed 23 March 2020.

[60] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, Sevgi, Classification of lung sounds using convolutional neural networks, EURASIP Journal on Image and Video Processing, vol. 2017, Springer; 2017, p. 65.

[61] B. Liu, and G. Tsoumakas, Synthetic Oversampling of Multi-Label Data based on Local Label Distribution, arXiv preprint arXiv:1905.00609, 2019.

[62] A. Bohadana, H. Azulai, A. Jarjoui, G. Kalak, and G. Izbicki, Influence of observer preferences and auscultatory skill on the choice of terms to describe lung sounds: a survey of staff physicians, residents and medical students, BMJ Open Respiratory Research Vol 7, Archives of Disease in Childhood; 2020, p. e000564.

[63] E. Andrès, R. Gass, A. Charloux, C. Brandt, and A. Hentzler, Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0, Journal of medicine and life Vol.11, Carol Davila-University Press; 2018, p. 89.

[64] A. Gurung, C. Scrafford, J. Tielsch, O. Levine, and W. Checkley, Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis, Respiratory medicine Vol 105, Elsevier; 2011, p. 1396–1401.

[65] H. Hafke-Dys, A. Breborowicz, P. Kleka, J. Kociński, and A. Biniakowski. The accuracy of lung auscultation in the practice of physicians and medical students, PloS one Vol. 14, Public Library of Science San Francisco, CA USA; 2019, p. e0220606.

[66] E. Gottlieb, J. Aliotta, and D. Tammaro. Comparison of analogue and electronic stethoscopes for pulmonary auscultation by internal medicine residents, Postgraduate Medical Journal Vol. 94, The Fellowship of Postgraduate Medicine; 2018, p. 700–703.

[67] Y. hui Huang, S. jun Meng, Y. Zhang, S. sheng Wu, Y. Zhang, and others. The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods, medRxiv, Cold Spring Harbor Laboratory Press; 2020.

[68] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data, arXiv preprint arXiv:2006.05919; 2020.

[69] G. Sgalla, S. Walsh, N. Sverzellati, S. Fletcher, S. Cerri, B. Dimitrov, D. Nikolic, A. Barney, F. Pancaldi, L. Larcher, and others, Velcro-type" crackles predict specific radiologic features of fibrotic interstitial lung disease, BMC Pulmonary Medicine, Vol. 18, Springer; 2018, p. 103.

[70] B. Zimmerman, and D. Williams. Lung Sounds, StatPearls [Internet], StatPearls Publishing; 2019.

[71] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra. Auscultation of the respiratory system, Annals of thoracic medicine Vol. 10, Wolters Kluwer–Medknow Publications; 2015, p. 158.

[72] J. Proctor, and E. Rickards. How to perform chest auscultation and interpret the findings, Nursing Times Vol. 116, 2020, p. 23–26.

[73] B. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. Kahya, N. Jakovljevic, T. Turukalo, I. Vogiatzis, E. Perantoni, and others. A respiratory sound database for the development of automated classification, International Conference on Biomedical and Health Informatics, Springer; 2017, p. 33–37.

[74] D. Mennitt, K. Fristrup. Influential factors and spatiotemporal patterns of environmental sound levels, INTER-NOISE and NOISE-CON Congress and Conference Proceedings Vol. 250, Institute of Noise Control Engineering; 2015, p. 2029–2040.

[75] H. Phan, O. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, and A. Mertins. Spatio-temporal attention pooling for audio scene classification, arXiv preprint arXiv:1904.03543, 2019.

[76] A. Gupta, G. Tang, and S. Suresh. HeartFit: An Accurate Platform for Heart Murmur Diagnosis Utilizing Deep Learning, arXiv preprint arXiv:1907.11649, 2019.

[77] G. Atluri, A. Karpatne, and V. Kumar. Spatio-temporal data mining: A survey of problems and methods, ACM Computing Surveys (CSUR) Vol. 51, ACM New York; 2018, p. 1–41.

[78] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F. Smolle-Jüttner, H. Olschewski, and F. Pernkopf. Multi-channel lung sound classification with convolutional recurrent neural networks, Computers in Biology and Medicine, Elsevier; 2020, p. 103831.

[79] M. Cartwright, J. Cramer, A. Mendez, Y. Wang, H. Wu,V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, Justin and others. SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context, arXiv preprint arXiv:2009.05188, 2020.

[80] J. Acharya, and A. Basu. Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning, IEEE transactions on biomedical circuits and systems Vol. 14, IEEE; 2020, p. 535–544.

[81] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan. Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases, arXiv preprint arXiv:2002.03894, 2020.

[82] C. Villanueva, J. Vincent, A. Slowinski, Alexander and M. Hosseini. Respiratory Sound Classification Using Long-Short Term Memory, arXiv preprint arXiv:2008.02900, 2020.

[83] A. Yadav, M. Dutta, and J. Prinosil. Machine Learning Based Automatic Classification of Respiratory Signals using Wavelet Transform, 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), IEEE; 2020, p. 545–549.

[84] N. Sommer, S. Velipasalar, L. Hirshfield, Y. Lu, and B. Kakillioglu. Simultaneous and Spatiotemporal Detection of Different Levels of Activity in Multidimensional Data, IEEE Access Vol. 8, IEEE; 2020, p. 118205–118218.

[85] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li. Triple-classification of respiratory sounds using optimized s-transform and deep residual networks, IEEE Access Vol. 7, IEEE; 2019, p, 32845–32852.

[86] R. Pramono, S. Bowyer, and E. Rodriguez-Villegas. Automatic adventitious respiratory sound analysis: A systematic review, PloS one Vol. 12, Public Library of Science San Francisco; 2017, p. e0177926.

[87] S. Zayet, Q. Lepiller, H. Zahra, P. Royer, L. Toko, V. Gendrin, T. Klopfenstein, and others, Clinical features of COVID-19 and influenza: a comparative study on Nord

Franche-Comte cluster, Microbes and infection, Vol. 22, No. 9, Elsevier; 2020. p. 481–488.

[88] E. Furman, A. Charushin, E. Eirikh, S. Malinin, V. Sheludko, V. Sokolovsky, G. Furman, THE REMOTE ANALYSIS OF BREATH SOUND IN COVID-19 PATIENTS: A SERIES OF CLINICAL CASES, medRxiv, Cold Spring Harbor Laboratory Press; 2020.

[89] S. Bernardi, F. Giudici, M. Leone, G. Zuolo, S. Furlotti, R. Carretta, and B. Fabris. A prospective study on the efficacy of patient simulation in heart and lung auscultation, BMC medical education, Vol. 19, No. 1, Springer; 2019, p. 1–7.

[90] D. Pereira, M. Amélia-Ferreira, R. Cruz-Correia, and M. Coimbra. Teaching Cardiopulmonary Auscultation to Medical Students using a Virtual Patient Simulation Technology, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE; 2020, p. 6032–6035.

[91] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto. Noise masking recurrent neural network for respiratory sound classification, International Conference on Artificial Neural Networks, Springer; 2018, p. 208–217.

[92] D. Perna, Convolutional neural networks learning from respiratory data, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE; 2018, p. 2109–2113.

[93] S. Shuvo, S. Ali, S. Swapnil, S. Irtiza, T. Hasan, and M. Bhuiyan, Mohammed, A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram, IEEE Journal of Biomedical and Health Informatics, IEEE; 2020.

[94] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen. Multi-label vs. combined single-label sound event detection with deep neural networks, 2015 23rd European signal processing conference (EUSIPCO), IEEE; 2015, p. 2551–2555.

[95] R. Serizel, N. Turpault, A. Shah, and J. Salamon. Sound event detection in synthetic domestic environments, ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE; 2020, p. 86–90.

[96] N. Baghel, M. Dutta, and R. Burget. Automatic Diagnosis of Multiple Cardiac Diseases from PCG Signals using Convolutional Neural Network, Computer Methods and Programs in Biomedicine. Elsevier; 2020, p. 105750.

[97] F. Nabi, K. Sundaraj, C. Lam, and R. Palaniappan. Characterization and classification of asthmatic wheeze sounds according to severity level using spectral integrated features, Computers in biology and medicine Vol. 104, Elsevier; 2019, p. 52–61.

[98] V. Basu, and S. Rana. Respiratory diseases recognition through respiratory sound with the help of deep neural network, 2020 4th International Conference on Computational Intelligence and Networks (CINE), IEEE; 2020, p. 1–6.

[99] S. Becker, M. Ackermann, S. Lapuschkin, K. Müller, and W. Samek, Interpreting and explaining deep neural networks for classification of audio signals, arXiv preprint arXiv:1807.03418; 2018.

[100] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, Investigation of Different CNN-Based Models for Improved Bird Sound Classification, IEEE Access, Vol. 7, IEEE; 2019, p. 175353–175361.

[101] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, arXiv preprint arXiv:1710.10501; 2017.

[102] B. Zhao, X. Li, X. Lu, and Z. Wang, A CNN–RNN architecture for multi-label weather recognition, Neurocomputing, Vol. 322, Elsevier; 2018, p. 47–57.

[103] A. Belkacem, S. Ouhbi, A. Lakas, E. Benkhelifa, and C. Chen, End-to-End AI-Based Point-of-Care Diagnosis System for Classifying Respiratory Illnesses and Early Detection of COVID-19, arXiv preprint arXiv:2006.15469; 2020.

[104] S. Khan, M.Sunny, M.Hossain, E.Hossain, and M.Ahmad, Nose tracking cursor control for the people with disabilities: An improved HCI, 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), IEEE; 2017, p. 1–5.

[105] O. Prabhune, and P. Rege, Speakingeyes: Enabling paralyzed people to communicate, 2019 IEEE 16th India Council International Conference (INDICON), IEEE; 2019, p. 1–4.

[106] P. Goel, R. Joshi, M. Sur, H. Murthy, A common spatial pattern approach for classification of mental counting and motor execution EEG, International Conference on Intelligent Human Computer Interaction, Springer; 2018, p. 25–35.

[107] Y. Li, X. Xiong, Z. Li, and Y. Fu, Recognition of three different imagined movement of the right foot based on functional near-infrared spectroscopy, Sheng wu yi xue gong cheng xue za zhi= Journal of biomedical engineering= Shengwu yixue gongchengxue zazhi, Vol. 37, No. 2, 2020, p. 262–270.

[108] X. Yin, B. Xu, C. Jiang, Y. Fu, Z. Wang, H. Li, and G. Shi, A hybrid BCI based on EEG and fNIRS signals improves the performance of decoding motor imagery of both force and speed of hand clenching, Journal of neural engineering, Vol. 12, No. 3, IOP Publishing; 2015.

[109] G. Hirsch, M. Dirodi, R. Xu, P. Reitner, and C. Guger, Online Classification of Motor Imagery Using EEG and fNIRS: A Hybrid Approach with Real Time

Human-Computer Interaction, International Conference on Human-Computer Interaction, Springer; 2020, p. 231–238.

[110] S. Woo, M. Kang, and K. Hong, Classification of Finger Tapping Tasks using Convolutional Neural Network Based on Augmented Data with Deep Convolutional Generative Adversarial Network, 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob), IEEE; p. 328–333.

[111] V. Scarapicchia, C. Brown, C. Mayo, and J. Gawryluk, Functional magnetic resonance imaging and functional near-infrared spectroscopy: insights from combined recording studies,Frontiers in Human Neuroscience, Vol. 11, Frontiers; 2017, p. 419.

[112] T. Wilcox, and M. Biondi, fNIRS in the developmental sciences, Wiley Interdisciplinary Reviews: Cognitive Science, Vol. 6, No. 3, Wiley Online Library; 2015, p. 263–283.

[113] A. Zafar, U. Ghafoor, M. Yaqub, and K. Hong, Initial-dip-based classification for fNIRS-BCI, Neural Imaging and Sensing 2019, Vol. 10865, International Society for Optics and Photonics; 2019, p. 108651N.

[114] M. Trager, K. Wilkins, M. Koop, and H. Bronte-Stewart, A validated measure of rigidity in Parkinson's disease using alternating finger tapping on an engineered keyboard, Parkinsonism & related disorders, Vol. 81, Elsevier; 2020, p. 161–164.

[115] R. Krupička, P. Krỳže, S. Net'uková, T. Duspivová, O. Klempíř, Z. Szabó, P. Dušek, K. Šonka, J. Rusz, and E. Ržička, Instrumental analysis of finger tapping reveals a novel early biomarker of parkinsonism in idiopathic rapid eye movement sleep behaviour disorder, Sleep Medicine, Vol. 75, Elsevier; 2020, p. 45–49.

[116] A. Alves-Pinto, S. Ehrlich, G. Cheng, V. Turova, T. Blumenstein, and R. Lampe, Effects of short-term piano training on measures of finger tapping, somatosensory

perception and motor-related brain activity in patients with cerebral palsy, Neuropsychiatric disease and treatment, Vol. 13, Dove Press; 2017, p. 2705.

[117] J. Birchenall, M. Térémetz, P. Roca, J. Lamy, C. Oppenheim, M. Maier, J. Mas, C. Lamy, J. Baron, and P. Lindberg, Individual recovery profiles of manual dexterity, and relation to corticospinal lesion load and excitability after stroke–a longitudinal pilot study, Neurophysiologie Clinique, Vol. 49, No. 2, Elsevier; 2019, p. 149–164.

[118] B. Young, Z. Nigogosyan, L. Walton, J. Song, V. Nair, S. Grogan, M. Tyler, D. Edwards, K. Caldera, J. Sattin, and others, Changes in functional brain organization and behavioral correlations after rehabilitative therapy using a brain-computer interface, Frontiers in Neuroengineering, Vol. 7, Frontiers; 2014, p. 26.

[119] M. Khan, and K. Hong, Active brain area identification using EEG-NIRS signal acquisition, 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), IEEE; 2015, p. 7–11.

[120] M. Peifer, L. Zhu, and L. Najafizadeh, Real-Time Classification of Actual vs Imagery Finger Tapping Using fNIRS, Biomedical Optics, Optical Society of America. 2014, p. BM3A–34.

[121] H. Nazeer, N. Naseer, R. Khan, F. Noori, N. Qureshi, U. Khan, and M. Khan, Enhancing classification accuracy of fNIRS-BCI using features acquired from vector-based phase analysis, Journal of Neural Engineering, Vol. 17, No. 5, IOP Publishing; 2020, p. 056025.

[122] S. Zhang, Y. Zheng, D. Wang, L. Wang, J. Ma, J. Zhang, W. Xu, D. Li, and D. Zhang, Application of a common spatial pattern-based algorithm for an fNIRS-based motor imagery brain-computer interface, Neuroscience Letters, Vol. 655, Elsevier; 2017, p. 35–40.

[123] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer, Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface, NeuroImage, Vol. 34, No. 4, Elsevier; 2007, p. 1416–1427.

[124] A. Batula, J. Mark, Y. Kim, and H. Ayaz, Comparison of brain activation during motor imagery and motor movement using fNIRS, Computational intelligence and neuroscience, Vol. 2017, Hindawi; 2017.

[125] W. Chen, J. Wagner, N. Heugel, J. Sugar, Y. Lee, L. Conant, M. Malloy, J. Heffernan, B. Quirk, A, Zinos, and others, Functional near-infrared spectroscopy and its clinical application in the field of neuroscience: advances and future directions, Frontiers in Neuroscience, Vol. 14, Frontiers; 2020, p. 724.

[126] K. Hong, and M. Yaqub, Application of functional near-infrared spectroscopy in the healthcare industry: A review, Journal of Innovative Optical Health Sciences, Vol. 12, No. 6, World Scientific; 2019, p. 1930012.

[127] S. Bak, J. Park, J. Shin, and J. Jeong, Open-access fNIRS dataset for classification of unilateral finger-and foot-tapping, Electronics, Vol. 8, No. 12, Multidisciplinary Digital Publishing Institute; 2019.

[128] T. Siddique, and M. Mahmud, Classification of fNIRS Data Under Uncertainty: A Bayesian Neural Network Approach, arXiv preprint arXiv:2101.07128, 2021.

[129] T. Trakoolwilaiwan, B. Behboodi, J. Lee, K. Kim, and J. Choi, Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution, Neurophotonics, Vol. 5, No. 1, International Society for Optics and Photonics; 2017, p. 011008.

[130] S. Wickramaratne, and M. Mahmud, A Deep Learning Based Ternary Task Classification System Using Gramian Angular Summation Field in fNIRS Neuroimaging Data, arXiv preprint arXiv:2101.05891, 2021.

[131] U. Asgher, K. Khalil, M. Khan, R. Ahmad, S. Butt, Y. Ayaz, N. Naseer, and S. Nazir, Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface, Frontiers in Neuroscience, Vol. 14, Frontiers; 2020, p. 584.

[132] L. Xu, Y. Liu, J. Yu, X. Li, X. Yu, H. Cheng, and J. Li, Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional near-infrared spectroscopy, Journal of Neuroscience Methods, Vol. 331, Elsevier; 2020, p. 108538.

[133] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, Youtube-8m: A large-scale video classification arXiv preprint arXiv:1609.08675, 2016.

[134] A. Olsson, P. Sager, E. Andersson, A. Björkman, N. Malešević, and C. Antfolk, Extraction of multi-labelled movement information from the raw HD-sEMG image with time-domain depth, Scientific reports, Vol. 9, No. 1, Nature Publishing Group; 2019, p. 1–10.

[135] D. Freer, and G. Yang, Data augmentation for self-paced motor imagery classification with C-LSTM, Journal of Neural Engineering, Vol. 17, No. 1, IOP Publishing; 2020, p. 016041.

[136] P. Kaur, and A. Gosain, Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise, ICT Based Innovations, Springer; 2018, p. 23–30.

[137] S. Kudugunta, and E. Ferrara, Deep neural networks for bot detection, Information Sciences, Vol. 467, Elsevier; 2018, p. 312–322.

[138] L. Shapley, A value for n-person games, Contributions to the Theory of Games, Vol. 2, No. 28, 1953, p. 307–317.

[139] S. Lundberg, and S. Lee, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874, 2017.

[140] Y. Hailemariam, A. Yazdinejad, R. Parizi, G. Srivastava, and A. Dehghantanha, An Empirical Evaluation of AI Deep Explainable Tools, 2020 IEEE Globecom Workshops, IEEE; 2020, p. 1–6.

[141] A. Belkacem, S. Nishio, T. Suzuki, H. Ishiguro, and M. Hirata, Neuromagnetic decoding of simultaneous bilateral hand movements for multidimensional brain machine interfaces, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 26, No. 6, IEEE; 2018, p. 1301–1310.

[142] J. Ye, L. Sun, B. Du, Y. Fu, X. Tong, H. Xiong, Hui, Co-prediction of multiple transportation demands based on deep spatio-temporal neural network, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, p. 305–313.

[143] H. Zhu, R. Vial, and S. Lu, Tornado: A spatio-temporal convolutional regression network for video action proposal, Proceedings of the IEEE International Conference on Computer Vision, 2017, p. 5813–5821.

[144] M. Majd, and R. Safabakhsh, A motion-aware ConvLSTM network for action recognition, Applied Intelligence, Springer; Vol. 49, No. 7, 2019, p. 2515–2521.

[145] A. Sanchez-Caballero, D. Fuentes-Jimenez, and C. Losada-Gutiérrez, Exploiting the ConvLSTM: Human Action Recognition using Raw Depth Video-Based Recurrent Neural Networks, arXiv preprint arXiv:2006.07744, 2020.

[146] J. Zhao, X. Mao, and L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Vol. 47, Elsevier; 2019, p. 312–323.

[147] A. Salau, T. Olowoyo, and S. Akinola, Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model, Advances in Computational Intelligence Techniques, Springer; 2020, p. 1–16.

[148] H. Ghonchi, M. Fateh, V. Abolghasemi, S. Ferdowsi, and M. Rezvani, Spatio-temporal deep learning for EEG-fNIRS brain computer interface, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE; 2020, p. 124–127.

[149] J. Aviles-Solis, C. Jacome, A. Davidsen,R. Einarsen, S. Vanbelle, H. Pasterkamp, and H. Melbye, Prevalence and clinical associations of wheezes and crackles in the general population: the Tromsø study, BMC pulmonary medicine Vol. 10, Springer; 2019, p. 373.

[150] V. I. McLane, D. Emmanouilidou, J. West, and M. Elhilali, Electronic Stethoscope Filtering Mimics the Perceived Sound Characteristics of Acoustic Stethoscope, IEEE Journal of Biomedical and Health Informatics, IEEE; 2020.

[151] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, Multi-label music genre classification from audio, text, and images using deep features, arXiv preprint arXiv:1707.04916, 2017.

[152] A. Schaefer, R. Kong, E. Gordon, T. Laumann, X. Zuo, A. Holmes, S. Eickhoff, B. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional

connectivity MRI, Cerebral cortex, Vol. 28, No. 9, Oxford University Press; 2018, p. 3095–3114.

[153] C. Horenstein, M. Lowe, K. Koenig, and M. Phillips, Comparison of unilateral and bilateral complex finger tapping-related activation in premotor and primary motor cortex, Human brain mapping, Vol. 30, No. 4, Wiley Online Library; 2009, p. 1397–1412.

VITA

NAME OF AUTHOR: Natalie M. Sommer

EDUCATION:

Syracuse University, Syracuse, NY : 7/2013 - present

- Currently enrolled and have reached ABD for PhD in Electrical Engineering.
- Expected Graduation Date: May 2021

Syracuse University, Syracuse, NY : 9/1991 - 5/1995

- Finished doctoral coursework for Electrical Engineering major.
- Recipient of the Dana Teaching Fellowship.

Union College, Schenectady, NY : 9/1987 - 6/1991

- Received a B.S.E.E. and an M.S.E.E. in four years.

EXPERIENCE:

DeVry College of New York, New York, NY : 10/2000 - present

Professor of Electronics Engineering Technology

- Taught (online and onsite): Digital Electronics, Signals and Systems, Electrical Circuits Theory, Signal Processing, Mechatronics, Physics I and II, Calculus I and II and C++ Programming.
- Composed lectures, demonstrations and lab activities for all courses taught. Lab activities involved use of VHDL programming of the Altera Cyclone III FPGA, wiring circuits with discrete components, simulations via MultiSim and LabView, programming with Microsoft Visual C++ and Matlab.
- Worked with Learning Management Systems, eCollege and Canvas, and conducted online lectures and tutoring sessions through Webex.
- Advised the DeVry College of New York IEEE student branch, held elections and helped organize seminars, competitions and meetings.
- Contributed to the development of courses for the new Automation and Electronic Systems Associate Program. Developed courses: Circuit Analysis Fundamentals and Electronic Devices and Systems.
- Participated in meetings with the Industry Advisory Committee to discuss curriculum and continuous improvement of DeVry's engineering programs.
- Assisted with the eTAC of ABET accreditation in 2016.

DeVry College of New York, New York, NY : 3/2019 - 5/2020

Faculty Chair of the Northeast Group

- Managed 85 Northeast Group visiting professors.
- Participated in onboarding of new visiting professors, including interviewing panels and teaching demonstrations.
- Supported the Assistant Dean of Academic Affairs in staffing recommendations and confirmations of visiting faculty for both onsite and online sections.

- Coached faculty in methodologies to improve student learning experience.
- Coordinated observations of all visiting faculty.
- Worked closely with Student Support Advisors in handling academic cases.

DeVry College of New York New York, NY : 1/2004 - 5/2012

Chair of Engineering Technology Programs

- Collaborated with other chairpersons in developing schedules for the Engineering Technology Department.
- Prepared presentations and participated in product knowledge sessions with the Admissions department.
- Worked on the TAC of ABET accreditation visits in 2004 and 2010. In 2010, took a lead in the preparations for the visit.

Oakton Community College, Des Plaines, IL : 7/1996 - 7/2000

Adjunct Physics Instructor/Tutor

- Taught Physics courses at the Associates level.
- Conducted Physics demonstrations and laboratory experiments.
- Tutored students in the tutoring center in Calculus.
- Completed the CRLA Tutor Training Certification.