The Pennsylvania State University
The Graduate School

INTERPRETABLE ARTIFICIAL INTELLIGENCE MODELS TO DETECT CHRONIC AND INFECTIOUS DISEASES

A Dissertation in
Industrial Engineering
by
Maryam Zokaeinikoo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2020

The dissertation of Maryam Zokaeinikoo was reviewed and approved by the following:

Soundar Kumara
Allen E. Pearce and Allen M. Pearce Professor of Industrial Engineering
Dissertation Co-Adviser, Co-Chair of Committee

Prasenjit Mitra
Professor of Information Sciences and Technology
Dissertation Co-Adviser, Co-Chair of Committee

Hui Yang
Professor of Industrial Engineering

Qiushi Chen
Professor of Industrial Engineering

Steven Landry
Department Head and Professor of Department of Industrial and Manufacturing Engineering

Abstract

Deep learning and artificial intelligence methods have revolutionized computational analytics by helping solve complex problems in many application domains, including healthcare and medicine. Deep learning methods comprise of multiple linear or nonlinear layers, enabling them to learn sophisticated features and subtle patterns from high-dimensional input data. However, typical deep neural network methods are often considered black-box models as they do not provide adequate insights into interpreting their predictions. This has posed challenges to the successful implementation of deep learning models in practice, especially in healthcare, where transparency and interpretability of models are critical to their application in practice.

This dissertation uses natural language processing, audio processing, and computer vision techniques along with deep learning to develop accurate and interpretable methods to detect chronic and infectious diseases. Three specific research topics are considered. The first research topic focuses on detecting the onset of Alzheimer's disease using transcript of interviews of individuals who were asked to describe a picture. We developed a hierarchical recurrent neural network (RNN) model for natural language processing using a novel attention over self-attention mechanism to model the temporal dependencies of longitudinal data. We demonstrate the interpretability of the model with the importance score of words, sentences, and transcripts extracted from the three-level neural network model.

The second problem we address seeks to eliminate the need for transcribing interviews by developing an end-to-end interpretable deep learning model for detecting Alzheimer's disease using raw audio interviews of patients. Our methods using both the text and the audio models achieve new benchmark accuracy performances compared to previous works. These artificial intelligence models can help diagnose Alzheimer's disease in a non-invasive and affordable manner, improve patient outcomes, and contain cost.

Third, we focused on detecting the Coronavirus Disease 2019 (COVID-19) from chest X-ray and Computed Tomography images. A novel hierarchical attention neural network model is developed to classify chest radiography images as belonging to a person with either COVID-19, other infections, or no pneumonia. The model's hierarchical structure captures the dependency of features and improves model performance while the attention mechanism makes the model interpretable and transparent. This model can be used in conjunction with or instead of laboratory testing (e.g., where laboratory testing is unavailable) to detect and isolate individuals with COVID-19 and prevent onward transmission to the general population and healthcare workers.

This dissertation effectively illustrates the use of deep learning methods in textual,

audio and visual data in medical informatics. Future work in this domain needs to focus on building integrated techniques and platforms to address the integration of the three modalities in specific problem scenarios.

Table of Contents

List of Figures

List of Tables

Acknowledgments

During my studies in elementary school in Iran, I was fortunate enough to have teachers that found my talents in math. I always remember from my time in elementary school that when I solved a math problem, it gave me the most profound joy and fulfillment. At that time, my mom always wanted me to become a medical doctor. She bought biology books for me to see whether she could change my mind. But my passion for math was strong enough to convince her that pursuing a math degree is the right choice for me. Twenty years later, I am so happy with my choice in high school. I am thankful for all my teachers and professors back in Iran. I am also grateful to my M.Sc. advisor, Dr. Nasser Salamsi, at Sharif University, who initially taught me how to conduct impactful research and solve real-world problems using the power of mathematics and computers.

My four years at Pennsylvania State University had many ups and downs and included the most terrible and the most exciting events in my life. During my first year at the Ph.D. program, I lost my very best friend in Iran. We were very close, and her unexpected death made my heart felt empty. I had the toughest time of my life during the first year of my Ph.D. at Penn State. However, I learned to accept my mistakes, grow, and move forward. I recognized what is most important for my inner self and pursue it regardless of all limitations that the outer world may impose on me. During this challenging time, I learned how understanding students and forgiveness could change their future. No matter what mistakes they have made, they should be supported and guided if they are motivated to move forward. Many times I became disappointed, but eventually, I decided to grow no matter what would happen. Not only I found my passion during the following years of my Ph.D. at Penn State, but also, my personality changed a lot.

I conquered many challenges, but I was fortunate to find Dr. Prasenjit Mitra from the College of Information Sciences and Technology in my second year of the Ph.D. program. Dr. Mitra kindly supported me and guided me to the exciting world of Artificial Intelligence (AI), specifically deep learning, and to conduct impactful research in the healthcare domain. Dr. Mitra gave me the freedom to pursue various projects without objection and has supported me over the past three years. The fruitful ideas from Dr. Mitra made me hopeful about advancing and finishing my Ph.D. dissertation with confidence. I always remember the following quote from him:

"Our responsibility is to knock the door; others will decide whether to open it or not."

I would like to express my deepest gratitude to my academic advisors, Dr. Prasenjit Mitra and Dr. Soundar Kumara, for their continuous support and acceptance to serve

as my co-advisors. I have benefitted a lot from their guidance, advice, and of course, trust in my work. I also like to thank the members of my Ph.D. committee, Dr. Akhil Kumar, Dr. Hui Yang, and Dr. Qiushi Chen, for their helpful comments and suggestions to shape my dissertation to this improved version.

Moreover, I thank my friends for providing support and friendship that I needed. I would like to especially thank Dr. Farhad Imani for supporting me throughout my Ph.D. program at Penn State and guiding me for every step during my Ph.D. studies at Penn State.

I especially thank my mom, dad, and sisters. My parents have sacrificed their lives for my sisters and me, and they have provided unconditional support and care. They believed in our potentials and gave us the freedom to experience life to the fullest without any limitations. I love them so much, and I would not have made it this far without them. Special thanks to the newest addition to my family, Pooyan Kazemian, my husband, and his amazing family, all of whom have been incredibly supportive and caring.

Finally, the most exciting thing that happened to me during the past four years was finding my best friend, soulmate, and husband. I married the best person out there for me. There are no words to describe how much I love him. Pooyan is a great mentor and has unconditionally supported me during my good and bad times. He believed in me and always encouraged me to grow. These past several years have not been a smooth ride, both academically and personally. I sincerely thank Pooyan for being there for me at all times. He completely understood how important my education and career are for me and has got my back in all situations.

Chapter 1
Introduction

The emergence of artificial intelligence (AI) and specifically deep neural networks (deep learning [4]) has become a powerful tool for solving complex problems in computational analytics. Deep learning typically uses a large number of parameters that perform very well when trained on a large amount of data. With advances in hardware resources such as graphics processing unit (GPU), deep learning has performed better than traditional machine learning methods [5–8].

Deep neural networks comprise of multiple layers of non-linear transformations, which ultimately can learn complex features from big and high-dimensional input data. However, the high-level abstraction of features makes it hard to interpret the deep neural networks. These complex models are mostly considered as black-box models. In order to obtain both patients' and clinicians' trust and ensure patient safety, we need interpretable AI models especially in the domain of healthcare. Therefore, the impressive predictive performance of deep learning models cannot compensate for their black-box nature for applying in the healthcare domain. The best predictive model is the one that is interpretable without sacrificing its prediction performance to be actively implemented in everyday clinical practice.

Motivated by the success of deep neural network language models in modeling acoustic signals [9], we propose interpretable models to model clinical natural texts. Unlike traditional machine learning approaches, in which feature engineering is required as a preliminary step, deep learning methods learn useful features directly from the data without any feature engineering interventions. Additionally they are more accurate in comparison to traditional machine learning approaches in computer vision, natural language processing (NLP), and acoustic modeling [5–8].

We propose multiple interpretable deep learning models to detect Alzheimer's disease as a chronic disease from patients' interview transcripts and their corresponding audio

recordings, and COVID-19 as an infectious disease from patients' chest X-ray images. In other words, we develop specific deep learning models to solve healthcare-related problems using natural language processing, audio processing, and image processing. We discuss the research problems addressed in this dissertation below.

Alzheimer's disease: Alzheimer's disease (AD) has been known in the past century as the primary cause of dementia, consequently leading to death [10]. Research is still ongoing in AD prevention, progression, and treatment. Early detection of Alzheimer's is vital to preventing, controlling, and stopping the disease. The symptoms of AD differ from person to person. In its early stages, memory-related neurons are damaged, which causes mild memory loss, specifically in remembering recently learned information for most affected individuals. As the disease progresses, the symptoms gradually worsen such that patients lose their ability to carry on daily-life activities and conversations. In the final stage of AD, patients are more vulnerable to infections due to loss of mobility. Lung infection (pneumonia) is the leading cause of death for people with Alzheimer's disease.

Alzheimer's disease is the sixth leading cause of death in the United States. Depending on the patient's age and other comorbidities, the survival rate varies from 4 to 20 years after the early detection of symptoms. Alzheimer's is typically diagnosed through extensive tests and cognitive tools [11–14]. The degeneration of brain cells can be reflected in a variety of ways in brain scans. However, diagnosis of Alzheimer's disease based only on these scans can lead to mistakes (both false positives and false negatives) because it is often not straightforward to distinguish normal age-related changes in the brain from the abnormal AD-related ones. Therefore, more accurate and reliable diagnosis methods are critical to improving patient outcomes.

Manual diagnosis of Alzheimer's disease is not only error-prone but is also time-consuming. Conducting diagnostic tests and neuropsychological examinations for patients and interpretation of the results by physicians may take several days or weeks [15]. The time required to diagnose AD manually is highly dependent on the physician's experience and knowledge [16–18]. Automatic mathematical tools and algorithms that can detect AD efficiently and accurately are, therefore, precious.

We propose deep learning-based predictive models to diagnose AD in its early stages using patients' speech transcriptions. Previous works on automatic AD detection mainly focused on extracting linguistic features from verbal utterances of healthy and AD subjects. Natural language processing (NLP) techniques are used primarily to extract lexical features from text and combined with machine learning algorithms to identify

probable AD subjects from healthy individuals [19].

In this dissertation, we combine NLP and various deep learning models to detect the onset of Alzheimer's disease based on the text in longitudinal patient interviews, and audio data without the need for invasive or expensive diagnostic methods. We directly apply a well-known deep learning algorithm, the long short term memory networks (LSTM) [20], to effectively capture both long and short term dependencies of words within interview transcripts. We first present six deep neural models based on the long short term memory (LSTM) method, which is designed for long sequences. An important advantage of deep learning is that we do not need to design features very carefully; deep learning algorithms can learn representations, which will then be used for the classification task. Unlike the previous works, we do not need the initial phase of careful feature engineering, which includes finding appropriate lexical features. Instead, we directly feed our model with speech transcriptions. The deep learning models learn representations of the features on their own. The first two models we present are LSTM and bidirectional LSTM (BLSTM). We incorporate an attention layer to both LSTM and BLSTM models to develop new learning methods. Then, we apply our methods on the same dataset as the study by Orimaye, et al., [19] to classify, and discuss the similarities and differences between our results.

To compare our results with the ground truth models, we also implement traditional machine-learning algorithms on the same dataset, including support vector machine (SVM) and random forest (RF), with the same evaluation method. We then try to explain the prediction of these models along with deep neural-based models with the help of local interpretable model-agnostic explanations (LIME) [21].

Due to the blackbox nature of LSTM, it was not possible to gain insights within each transcript to see which words or phrases are the most indicators of memory loss. Next, we proposed novel interpretable hierarchical recurrent neural network (RNN) models and combined it with natural language processing (NLP) to detect the onset of Alzheimer's disease (AD) based on longitudinal patient interview data. The goal is to make the black box neural network model interpretable using an attention mechanism. We evaluate the attention scores our model gives to transcripts, sentences, and words to shed light on how the neural network model made predictions. This hierarchical model is described in detail in Chapter2.

Although we developed a novel interpretable model to detect the onset of AD using patients' transcripts, the input data needs an expert who can transcribe the audio recordings to the specific text format, which can encode the patients' disfluencies during

their speech. The transcribing task is time-consuming and labor-expensive. Motivated by this limitation, for the next project, we aimed to develop an interpretable model for patients' audio interviews. The model can be fed directly with raw audio recordings. It will indicate to clinicians not only whether the patient has Alzheimer's disease but also which parts of their audio interview have the signs of memory loss. In this project, we first implemented the current end-to-end audio models on our Alzheimer's audio dataset. Then, we developed a three level hierarchical model with the help of transfer learning and attention mechanism to make the audio model interpretable. The performance of our audio model beats works on the same Alzheimer's audio dataset. In addition, the model is successful to capture patients language disfluencies which are primary indicator of cognitive impairment due to Alzheimer's disease. This interpretable audio model is described in detail in Chapter 3.

COVID-19: The Coronavirus (SARS-CoV-2) which has led to COVID-19 pandemic in late 2019 motivated us to develop an interpretable model to detect COVID-19 from patients' X-ray images and CT scans. The best way to control transmissions and flatten the curve is to test as many people as possible to prevent further transmission of the virus by quickly identifying and isolating the infected individuals. However, since the current PCR testing kits are limited and have a low sensitivity of 70%, we aimed to develop a computational method that can detect not only COVID-19 patients from other infections and normal individuals but also it can recognize the specific regions of the lung that are affected by the virus. The proposed interpretable model is described in more detail in Chapter 4.

In summary, in an effort to develop interpretable models in computational healthcare, we have covered all three main types of input data which are text, audio and image and proposed interpretable models which have demonstrated both good predictive performance and interpretibility capability (Figure 1.1). This is extremely valuable in healthcare where the transparency of predictive models can play an important role in real world clinical application. The conclusions and future works of this dissertation are discussed in more detail in Chapter 5.

We have divided our research problem into four sub-problems. The first sub-problem is to develop sequence-based deep learning models to predict the onset of Alzheimer's disease based on the most recent visit of patients. Developing an interpretable deep neural network to consider the longitudinal Alzheimer's data is the second sub-problem. The third sub-problem discusses developing an end-to-end deep audio model that takes patients' raw audio speech and tells which parts of audio have memory loss issues. The

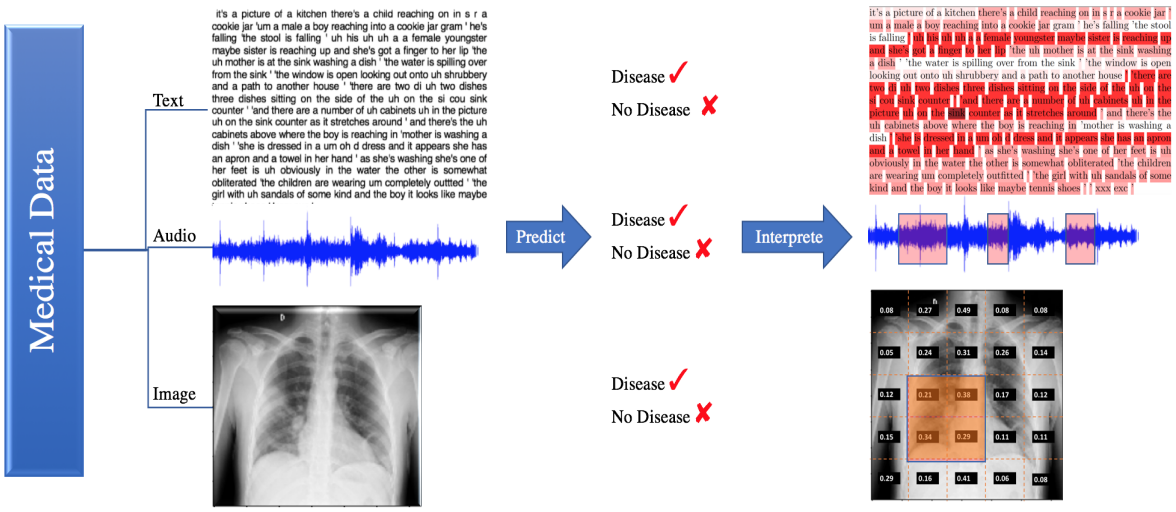


Figure 1.1: Interpretability of predictive models in all three types of medical data

fourth sub-problem focuses on developing an interpretable model based on deep neural networks to detect COVID-19 using patients' chest X-ray images and CT scans.

1.1 Specific Research Questions and Objectives

Deep learning-based models in natural language processing, signal processing, and computer vision have outperformed the traditional machine learning algorithms combined with feature engineering. This dissertation seeks to apply deep learning-based techniques to real-world health data to detect chronic and infectious diseases more efficiently.

This dissertation focuses on the following research questions in chronic disease management:

- Which type of automated methods will be useful to to detect the onset of Alzheimer's disease based on the patient transcripts?
- How to design an interpretable learning structure to model the temporal dependencies of longitudinal patient transcripts?
- How to design an interpretable model to detect the onset of Alzheimer's disease using only the audio recordings as inputs?
- How to develop an interpretable model to detect COVID-19 using patients' chest X-ray images and CT scans as inputs?

To answer these research questions, we have defined the following project objectives:

Objective 1: To explore different deep learning techniques and how they can be applied to textual data. The goal is to explore the state-of-the-art deep learning techniques and move away from the traditional machine learning algorithms that require feature engineering.

Objective 2: To propose and justify a hierarchical deep neural network structure for longitudinal transcripts of Alzheimer's disease to capture the hierarchical dependencies between words, sentences, and longitudinal interview transcripts. To the best of our knowledge, we are the first to implement an attention mechanism to determine the memory loss indicators in words, sentences, and transcripts.

Objective 3: To construct an interpretable hierarchical model using the patients' audio interviews which can detect the signs of memory loss using patients' raw speech.

Objective 4: To develop a new predictive model to detect the COVID-19 from chest X-ray images which can indicates the subtle signs of infected parts of the patients' lung.

This dissertation's overarching goal is to develop interpretable predictive models on three main medical data modalities, including text, audio, and image, to detect chronic and infectious diseases.

1.2 Contributions

Below, We explain the main contributions of this dissertation.

- Achieving a new benchmark accuracy score to predict the onset of Alzheimer's disease based on textual data;
- Identifying main indicators of memory loss in patients' textual data at both sentence-level and word-level;
- Achieving a new benchmark accuracy score to detect Alzheimer's disease using the raw audio interviews of patients;
- Identifying the indicators of memory loss in patients audio speech by developing a novel hierarchical deep audio model; and
- Developing a hierarchical model to detect COVID-19, which not only achieves a new benchmark accuracy score but also can capture the affected areas of the patient's lung very well.

In summary, we develop interpretable models in this dissertation. These models not only describe the relationships between their input and output but also the knowledge extracted from our models aligns very well with the physician's medical knowledge. In the following chapters, we explain these interpretable models and their applicability in the medical domain.

This dissertation is organized as follows. Chapter 2 describes the predictive and interpretable models to detect the onset of Alzheimer's disease from patients' textual interview data. Chapter 3 explains how we designed the mechanism of our interpretable audio model to detect Alzheimer's disease using patitents' audio interview data. We also developed an interpretable model on medical images to detect COVID-19 patients from other viral/bacterial infections and normal subjects. This model is illustrated thoroughly in chapter 4. Finally, we wrapped up the results of this dissertation with conclusion and future works in Chapter 5.

Chapter 2
Detection of Alzheimer's Using Natural Language Processing

Alzheimer's disease (AD) is a progressive neurodegenerative disorder in which cognitive abilities (including memory and language) and executive function deteriorate gradually. In the past century, AD has been known as the primary cause of dementia [22], consequently leading to death [10]. Alzheimer's disease is the sixth leading cause of death in the United States. James et al. [23] found that AD may cause 500,000 annual deaths in the United States, with the mortality rate five or six times higher than official estimates. AD is typically diagnosed through extensive tests and cognitive tools [11]. To diagnose AD, imaging methods, such as positron emission tomography (PET) scan and magnetic resonance imaging (MRI), and invasive methods, such as cerebrospinal fluid analysis, are employed. The degeneration of brain cells can be reflected in a variety of ways in brain scans. However, a diagnosis of AD based only on these scans can lead to mistakes (both false positives and false negatives) because we cannot often easily distinguish between normal age-related changes in the brain and abnormal AD-related ones. We need for more accurate and reliable diagnostic methods to improve patient outcomes [24].

Attempts to develop neuropsychological tests using a series of cognitive tests containing a set of questions and images have been made to detect the early signs of AD with various accuracy levels [25]. These attempts produced screening tools, such as the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA). They consist of questions and cognitive examinations to assess patients' cognitive abilities. However, the quality of the assessment depends on the physicians' experience and ability to distinguish between different categories of the disease [26]. Sometimes physicians need to combine MMSE with other cognitive tests, which makes it cumbersome and complicated to diagnose AD [27]. The National Institute on Aging and the Alzheimer's

Association have called for better approaches to diagnose AD in a non-invasive way [28]. It is estimated that early detection of Alzheimer's disease, even partially, can result in a significant $7 trillion cost saving compared to the status quo [29]. Automatic mathematical tools and algorithms that can detect AD early and accurately are therefore extremely valuable.

Speech is a valuable source of clinical information, which has been proven to be a reliable indicator of cognitive status [30]. The nerve cells that control cognitive ability and speech processing gradually deteriorate in individuals with AD [31]. Thus, the linguistic deficit captured by verbal utterances can be an indicator of Alzheimer's disease [32]. In recent years, artificial intelligence (AI) has been widely used to build prediction models with relatively high accuracy by capturing non-linear and complex patterns in the data [33–35]. Lately, AI methods have been proposed to detect AD by combining signal processing, machine learning, and natural language processing (NLP), which employed either recorded narrative speech [36], or recorded scene descriptions [32]. This fact motivated us to implement deep learning techniques to improve text-based AD diagnosis.

2.1 Related Works

Previous works to detect the early onset of AD using language as their input data are mainly dependent on extracting linguistic features from transcripts [19]. The main problems with feature-based methods are not only that the quality of prediction is highly dependent on the quality of features but also some intricate features may not recognizable by existing methods [4]. In addition, language evolution may also affect linguistic features' extraction methods. Recently, deep learning models has beaten other feature-based machine learning methods in speech recognition [8, 37, 38] and achieved promising results for various tasks in natural language processing including sentiment analysis [39] and natural language understanding [40].

Several studies used lexical features to detect AD from Dementiabank dataset [19, 41]. Wankerl et al. [42] proposed statistical approaches toward detecting AD using n-gram models. They evaluated their approach on DementiaBank dataset and achieved an accuracy 77.1%. Orimaye et al. proposed deep language models using decomposed higher-order n-grams N dimensional vectors as discrete inputs on the Dementiabank dataset [43]. Their experimental results show that deep neural networks sufficiently learn linguistic markers with reasonable accuracy on this small AD clinical dataset (area under ROC curve for their best model is 83%). These models, however, are not interpretable,

and do not provide insight into linguistic deficits that can indicate the onset of AD. Clinicians are hesitant to incorporate non-interpretable models into clinical practice. Karlekar et al. applied three neural models based on convolutional neural network (CNNs), long short term memory recurrent neural networks (LSTM-RNNs), and their combination to distinguish patients with AD from control patients based on documents in Dementiabank [44]. Their best model without feature engineering was the combined CNN-RNN model, which achieved an accuracy of 84.9%. More recently, Chen et al. [45] proposed a network based on attention mechanism by combining CNN and GRU modules to capture linguistic deficits of AD patients from DementiaBank dataset. They the cross-validation accuracy of 97% in detecting AD subjects from control subjects. Fritch et al. [46] improved the statistical approach proposed by Wankerl et al. [42] by proposing a neural language model based on LSTM cells and evaluating perplexity of their model. They obtained an accuarcy of 85.6% for the binary classification involved in identifying AD individuals from normal ones. Pan et al. [47] proposed a hierarchical attention-based neural0-network models that can capture the dependencies of the components within transcriptions. They used automatic speech recognition (ASR) to automatically transcribe and segment data. They achieved an F-score of 84.43% on manual and automatic transcripts from the DementiaBank dataset. Chien et al. [48] proposed a convolutional recurrent neural network model (CRNN) and obtained a performance of 83.8% in terms of the area under the receiver operating characteristic curve. Kong et al. [49] applied hierarchical attention (HAN) network to the DementiaBank dataset without extra feature engineering. Combining with demographic feature (age),they achieved an accuracy of 86.9%.

In this dissertation, we develop a novel three-level attention-based RNN model that does not require feature-engineering, is interpretable, and achieves an outstanding accuracy.

2.1.1 Main Contributions

The contributions of this work are four folds: (1) We developed a new three-level hierarchical structure to capture the hierarchical dependencies between words, sentences, and longitudinal interview transcripts to detect the onset of Alzheimer's disease; (2) To the best of our knowledge, we are the first to develop an attention over self-attention (AoS) mechanism that can prevent information loss by considering the relation of words, sentences, and transcripts within the sequence, and to demonstrate its implementation to confront a pressing healthcare problem; (3) Our model is interpretable, which addresses

a common shortcoming of black-box neural network models and provides valuable insight into language deficit that heralds the onset of Alzheimer's disease; (4) Numerical experiments on the DementiaBank dataset with data augmentation indicate that our novel AoS model achieves a high accuracy (mean cross-validation accuracy=98%) and outperforms other models of similar nature (that do not need feature engineering) developed on the same dataset.

2.2 Supervised Text-based Sequence Classification

Sequence classification is the task of assigning labels to the sequence of inputs over space or time. The challenging part of this task is that the text inputs vary in length and sequence dependencies exist between text symbols. Recently, deep learning approaches have been leveraged for text-based sequence classification. Convolutional neural networks (CNN) have been applied for semantic modeling of sentences [50, 51]. Recurrent neural networks (RNN) have been widely used to model sequences. Since documents are sequences of sentences and sentences are the sequence of words, RNN architectures can be employed to solve sentiment classification tasks [52, 53].

The main drawback of RNN is the vanishing/exploding gradient problem while working with long sequences, which means that the gradients can get very small or very large [54]. Long short term memory networks (LSTM) are developed by modifying the RNN network architecture to overcome the RNN limitations. LSTM models have been proven to be effective in capturing long-term dependencies [20].

These neural network language models have the following in common [34]:

- The models use a dictionary for word encodings as their inputs;
- The softmax activation function is applied to have normalized probability values in the output vector; and
- Cross entropy is used as the loss function for training.

In this work, we demonstrate that LSTM-based models can accurately distinguish probable AD subjects from healthy individuals.

2.3 Word Embeddings

Word embedding is a method in natural language processing that maps words and phrases from the set of vocabulary to vectors of real numbers. This method captures the context of words such that the words with similar or close semantic meanings have close vector representations. Word embedding converts words to the features so they can be fed to our neural networks.

2.3.1 One-hot-encoding

The most straightforward method to convert a word to a vector is called one-hot-encoding. Let N be the total number of unique words in our training dataset. We establish an order from zero to N for all words. The vector for the i-th word is defined as all zeros except for a 1 in the position i. Although this method is simple and easy to use, its performance is highly dependent on the size of the encoded vector. For instance, if we have 1 million words in our vocabulary, the dimension of its corresponding one-hot-encoded vector will be 1 million. However, his method does not capture the semantics of words as accurately as other recent methods. In 2013, the word2vec method was proposed for efficient word representation in vector space and changed the text vectorization field [37].

2.3.2 Word2vec

Word2vec is a technique to produce fixed-size vectors to represent words using surrounding words of that word such that the words that have close semantic meaning have close vectors. The idea of Word2vec comes from the intuition of how humans can understand the meaning of words by their adjacent words. For instance, consider the following sentence:

"I like eating X."

We may not know what X is, but we know that X is something that we can eat and is likely delicious. The human brain comes to this conclusion using the nearby words of "like" and "eating." Word2vec can generate word embeddings with two methods using neural networks: Skip Gram and Common Bag Of Words (CBOW).

CBOW model: The continuous bag-of-word model takes the word's context and predicts the word corresponding to the context. In our previous example, consider the input word "eating"; the goal is to predict "X." By feeding a neural network with the one-hot-encoding of the input word, we can learn the representation of the target word.

Figure 2.1 demonstrates the network structure for CBOW Model [55].

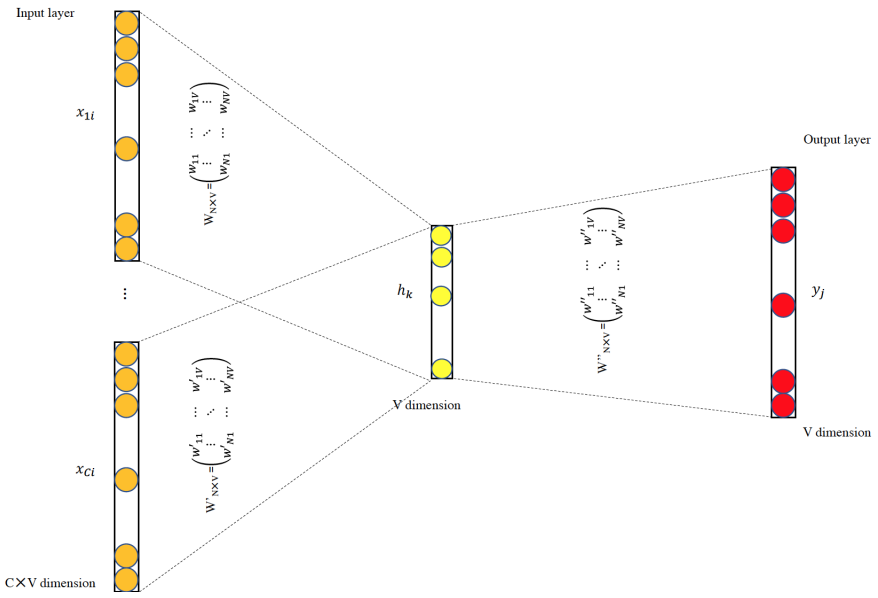


Figure 2.1: The continuous bag-of-word model

In this figure, the input is a one-hot-encoded vector of size N, and the output is the result of the hidden layer of size V and the softmax activation function. Given enough data, Word2vec can learn the meaning of words based on past appearances.

Skip-gram model: The Skip-gram model architecture is designed to predict the context words (surrounding words) given the target word (center word). In fact, Skip-gram is the reverse function of the CBOW model [55] (Figure 2.2). There is one hidden layer of size V which computes the dot product between the weight matrix and the input vector of size N. The output layer computed the dot product between the hidden layer and the output layer. Then, the Softmax function will be applied to obtain the probability of the words of the target context.

2.3.3 GloVe

GloVe (Global Vectors) [56] is another word embedding method designed based on the matrix factorization technique [57]. GloVe creates a matrix that counts the frequency of co-occurrence of words within some contexts. Then, the feature embedding of words is obtained by converting the initial matrix of (words x context) to a lower-dimensional matrix using the factorization method [56]. The resulting word representations perform very well on word analogy tasks [58].

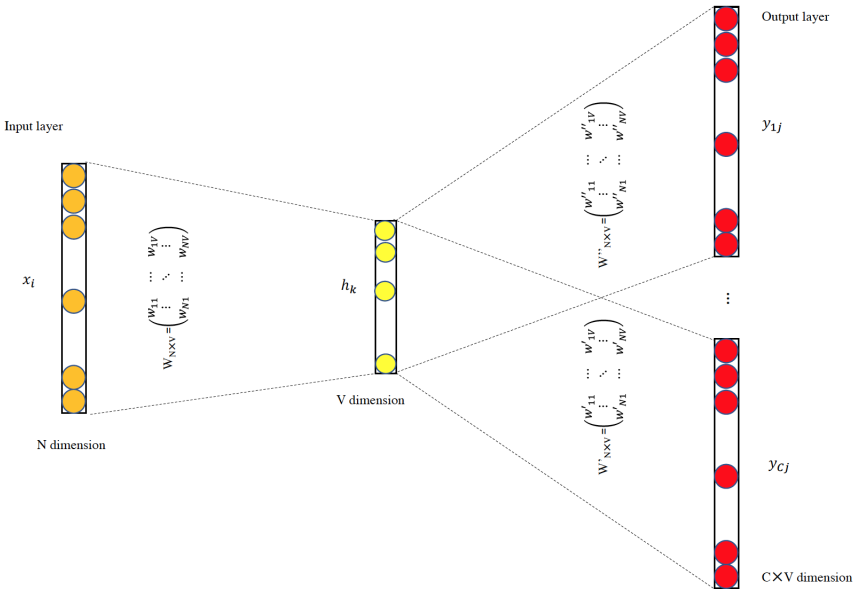


Figure 2.2: The Skip-gram model

2.3.4 ELMO

ELMO (Embeddings from Language Models) [59] is a deep contextualized word representations, which is different from traditional word embedding representations such as Word2vec and GloVe. Depending on the context containing the word, ELMO generates different word embedding vectors. The word representations are transferred from the internal states of a pre-trained deep bidirectional language model (biLM). ELMO has proved its performance in various NLP tasks, including sentiment analysis [59].

2.3.5 Transformers

Transformer models including BERT [60] and GPT-2 [61] are the most recent word embedding approaches. The main advantage of transformers is that they are not sequentially designed like ELMO. Transformer models are built using the attention mechanism [62], which enables them to process all the words of a sequence in parallel. Vaswani et al. [62] explained the internal mechanism of transformers using the attention mechanism.

2.4 Prediction models to detect Alzheimer's disease

2.4.1 Dataset

To evaluate our models, we use the DementiaBank clinical dataset. Data were collected longitudinally by the University of Pittsburgh School of Medicine for the study of communication in dementia. The dataset contains transcripts of the participants' interviews with possible Alzheimer's disease (AD) and other related dementia. Participants were asked to describe everything happening in a Cookie-Theft picture (Figure 2.3). The descriptions were then used to detect language disorders due to AD. The audios of participants' interviews were transcribed to Codes for the Human Analysis of Transcripts (CHAT) format. The CHAT transcription format is a tool that helps automatically transcribe audio files.

*INV: this is the picture.
*PAR: mhm. [+ exc]
*INV: just tell me everything that you see happening in that picture.
*PAR: +< alright. [+ exc]
*PAR: there's &um a young boy that's getting a cookie jar.
*PAR: and it [//] he's &uh in bad shape because &uh the thing is fallin(g) over.
*PAR: and in the picture the mother is washin(g) dishes and doesn't see it.
*PAR: and so <is the> [//] the water is overflowing in the sink.
PAR: and the dishes might <get falled [+ed] over if you don't> [//] fell [//] fall over there [/] there if you don't get it.
*PAR: and it [//] there [//] it's a picture of a kitchen window.
*PAR: and the curtains are very &uh distinct.
*PAR: but the water is &flow still flowing.
*INV: okay thank you very much.
@End

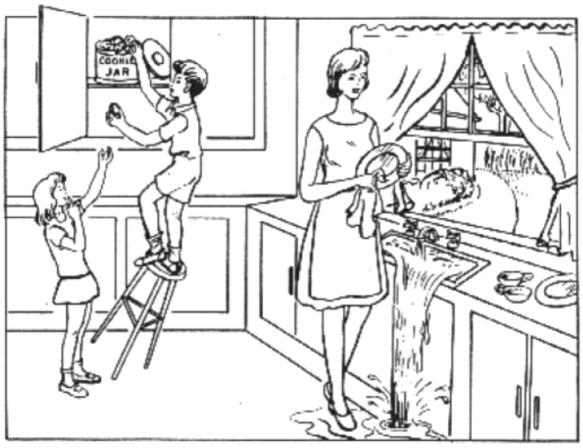


Figure 2.3: Sample interview and the Boston Cookie-theft picture (from the DementiaBank dataset [1]), designed to elicit language deficit that contributes to the diagnosis of AD.

Table 2.1 presents some CHAT disfluency transcription codes from the CHILDES manual that can be helpful for our analysis [63]. The DementiaBank dataset contains 99 healthy subjects and 169 probable AD subjects. In this initial phase, we include the 99 healthy subjects as well as the first 99 AD subjects so that we have a balanced dataset. In order to directly compare our results with Orimaye et al. [19], we choose the dataset for this experiment in the same way as Orimaye et al. did. So, we only use the last visit transcripts of both healthy and probable AD subjects.

Table 2.1: CHAT disfluency codes

Disfluency	Code
whole word repetition	follow word with [/]
multiple whole word repetition	[x 'number of repetitions']
phrase repetition	<> [/]
word revision	[//]
phrase revision	<> [//]
pause	(.) or (..) or (...)
filled pause	&-
unintelligible words	xxx

2.4.2 Data Pre-processing

We removed the interviewer questions as well as the initial information of each text document. We trained our models on the interviewees' textual description of the Cookie-Theft picture.

2.4.3 Classification Models

2.4.3.1 Deep Learning Models

Recurrent neural networks (RNN): Traditional deep neural networks cannot capture the dependencies of sequential data, e.g., words of a sentence or scenes of a movie []. Derived from deep neural networks, recurrent neural networks (RNN) [] are designed in a recurring fashion, in which the hidden states have all the previous information of the input sequences. This feature makes them suitable for many applications, such as natural language processing. The basic structure of an RNN is shown in Figure 2.4.

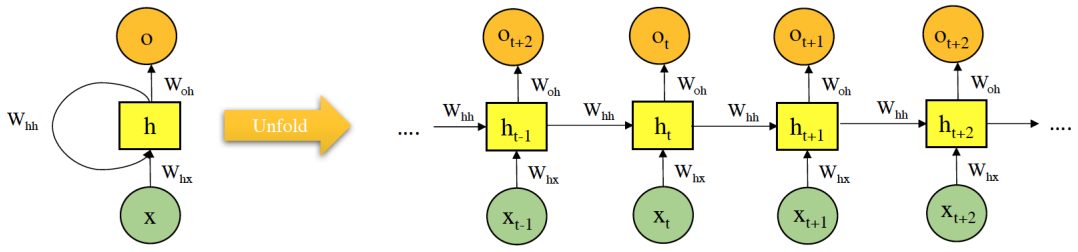


Figure 2.4: The unfolded structure of RNN

The unfolded RNN structure shows the sequential nature of the network. The length

of the network depends only on the length of the sequence. For instance, there will be ten states for a sentence with ten words. The basic functions that define figure 2.4 are as follows:

$$h_t = \sigma(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \tag{2.1}$$

$$o_t = W_{oh}h_t + b_o \tag{2.2}$$

In the above equations, x_t is the input at time step t. h_t is the hidden state at time step t. σ is a non-linear function. W_{hx} is the matrix of weights between the input and hidden layer, and W_{hh} is the matrix of weights between adjacent hidden layers. The vectors b_h and b_o are biases of the network. The final output (label) of the network is computed as:

$$y_t = softmax(o_t) \tag{2.3}$$

Depending on the task, we may get the output at each time step from the network (sequence to sequence) or the final label (sentiment analysis).

Unlike the traditional neural networks that compute different parameters for each hidden layer, RNN shares the parameters of U, W and V across all time steps. This feature significantly reduces the number of parameters compared to a traditional neural network.

Learning long-range dependencies across many time steps can lead to the vanishing or exploding gradients [33, 54]. Depending on whether the weight matrix between hidden states $W_{hh} < 1$ or $W_{hh} > 1$ and the type of activation function (logistic, sigmoid, or ReLU) the vanishing or exploding of gradients may occur. Pascanu, Mikolov, and Bengio [64] provide a more detailed mathematical explanation about the specific conditions under which this problem happens in an RNN.

Long Short Term Memory networks (LSTMs): LSTMs overcome the issue of vanishing gradients in recurrent neural networks [54]. In this type of network, the hidden nodes in an RNN are replaced with memory cells to capture long term dependencies of a sequence [54]. The term "Long-Short-Term-Memory" comes from the idea that LSTMs can capture both long-term memory (by learning weights similar to RNNs) and short-term memory (using its novel internal activation gates).

The internal structure of an LSTM consists of the input gate i_t, forget gate f_t, output

gate o_t, and cell activation vectors c_t. These gates are computed as the functions of the input x_t and previous hidden state h_{t-1} in the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2.4}$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.5}$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.6}$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{2.7}$$
$$h_t = o_t \tanh(c_t) \tag{2.8}$$

where, σ is the logistic sigmoid function.

The recurrent layer consists of the input, output, and forget gates, which play critical roles in remembering or forgetting information within the LSTM network.

Bidirectional RNN: In addition to LSTM, which only keeps track of past information, Bidirectional RNN (BRNN) [65] is one of the most widely used RNN derivatives to model sequences. The BRNN structure is constructed based on two layers of hidden nodes such that both are connected to input and output. The first hidden layer has recurrent connections to the past time while the second layer captures the future dependencies. Figure 2.5 illustrates the structure of a BRNN in detail.

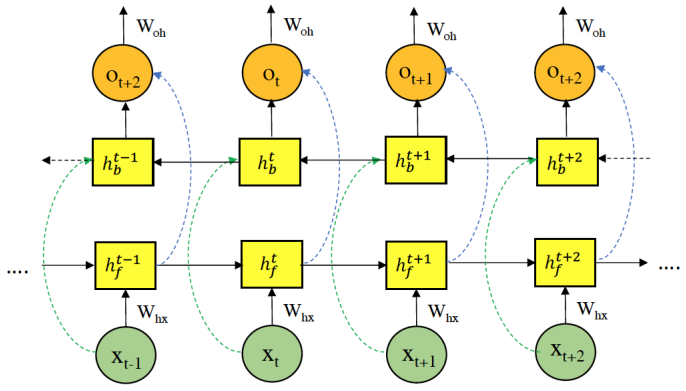


Figure 2.5: The Bidirectional RNN (BRNN) structure

Given a fixed sequence length, the structure of a BRNN is formulated as follows:

$$h_f{}^{(t)} = \sigma(W_{h_f x}x^{(t)} + W_{h_f h_f}h_f{}^{(t-1)} + b_{h_f}) \tag{2.9}$$

$$h_b{}^{(t)} = \sigma(W_{h_b x} x^{(t)} + W_{h_b h_b} h_b{}^{(t+1)} + b_{h_b}) \tag{2.10}$$

$$y^{(t)} = \text{softmax}(W_{y h_f} h_f{}^{(t)} + W_{y h_b} h_b{}^{(t)} + b_y) \tag{2.11}$$

where, $h_f{}^{(t)}$ and $h_b{}^{(t)}$ are parameters for hidden layers of forward and backward networks, respectively.

Attention Bidirectional LSTM: Attention-based neural networks have been performing well in speech recognition and machine translation [6, 66]. In this work, we implement an attention layer right after the BLSTM layer (Figure 2.6) to make a weighted sum of BLSTM output vector (H) that is used to produce the sentence representation (r) as follows:

$$M = tanh(H) \tag{2.12}$$

$$\alpha = softmax(w^T M) \tag{2.13}$$

$$r = H\alpha^T \tag{2.14}$$

where w is a BLSTM-trained weight vector, and T indicates the transpose of the vector. Attention helps the network recognize which parts of the sequence play an important role in text classification. Figure 2.6 demonstrates the structure of Attention Bidirectional LSTM.

In all models, we used deep LSTM-based structures that receive the input as text and return a binary output of 0-1 to classify healthy and probable AD subjects. The first step of our procedure is to break each interview transcripts into tokens and obtain 100-dimensional word embeddings by using the pre-trained GloVe model [56]. The word embedding vectors are considered as the features of input transcripts without any additional efforts on lexical feature extraction and fed directly into the document composition of the LSTM network. Technically, we add an embedding layer to encode text documents to real-valued vectors. Also, we truncated and padded input sequences to equalize the length of the input vectors. we chose the average of 250 words per transcript.

We incorporated two LSTM (BLSTM) layers with the dimensions of 64 and 32 units in all models. The batch size is 32, and we trained our LSTM based models for ten epochs with the validation set as 10% of the training set. The drop out rate is 0.25 in all our experiments.

We then apply the softmax function in the last layer to generate the conditional probabilities for each category of the binary classification. In other words, using the

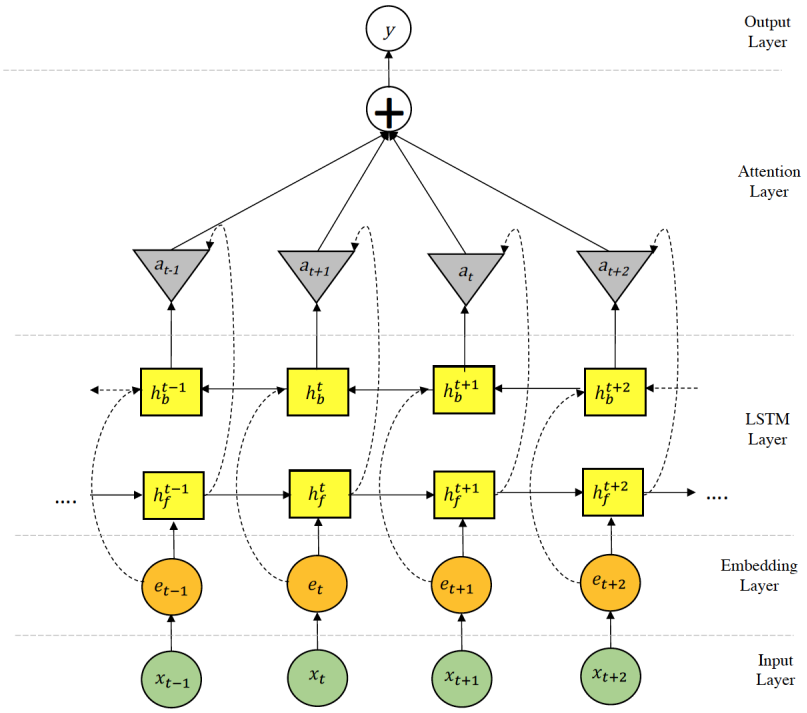


Figure 2.6: Attention Bidirectional LSTM structure

softmax function in the outer layer of the neural network uses a multinomial distribution, which returns the probability of the class c_j as follows:

$$P(c_j|x_j) = \frac{exp(z_j)}{\sum_{k=1}^{n} exp(z_k)} \tag{2.15}$$

where z_j is the output of the neural network for the class label j before applying the softmax function and n is the total number of classes. The loss function is defined as the binary cross-entropy between the actual distribution of text samples and their predicted distribution. We use the well-known ADAM optimizer inside the body of the LSTM network [67].

2.4.3.2 Traditional Machine Learning Models

To compare the performance of our deep learning models with the ground truth, we have implemented two models of traditional machine learning algorithms, including support vector machine (SVM) and random forest (RF). To extract the numerical features from the text transcriptions, we used the well-known TF-IDF [68] (term frequency-inverse document frequency) vectorizer, which measures the importance of each word to a

transcript in a corpus for both unigrams and bigrams.

2.4.4 Experiments

The performance of algorithms in predicting the onset of AD was evaluated using a cross-validation technique. We segmented our balanced dataset including 99 healthy and 99 probable AD subjects into 11 folds in which 18 interview transcripts (9 transcripts for each class of AD and healthy) exist for testing and the remaining of 180 interview transcripts for training and validation. The classification results are measured using 'accuracy,' which is a standard metric to measure the overall textual classification performance [69, 70]. We conducted extensive experiments to find out the best network architecture, which resulted in the highest accuracy. The ultimate evaluation metric is measured using the average accuracy among 11-fold cross-validations.

2.4.5 Results

Table 2.2 compares the results of the proposed six different classification algorithms including LSTM, BLSTM, attention-based LSTM (Att-LSTM), attention-based BLSTM (Att-BLSTM), SVM and RF. The results are reported in terms of accuracy for each fold and the average accuracy among 11 folds. Our results demonstrate that the attention bidirectional LSTM (Att-BLSTM) algorithm achieves the highest mean accuracy of 94.4%.We also represent the summary of the Table 2.2 in Figure 2.7. The figure illustrates the Box and Whisker plot for 11-fold cross-validation. Whiskers extend to data points that are less than 1.5 x The interquartile range (IQR) away from 1st/3rd quartile. As the figure shows, Attention-BLSTM outperforms the other algorithms by 4% or more.

Compared to other related works, Karlekar et al. (2018) [44] implemented the CNN-LSTM with an accuracy of 84.9% with the same untagged dataset as ours. They also achieved an accuracy of 91.1% with POS-tagged data. Orimaye et al. (2016) [71] attained the accuracy of 87.5% with only 38 transcripts using deep neural networks. We achieved a high accuracy performance of 94.4% using the same dataset as Orimaye et al. [71] did. However, our model is not interpretable (black-box). Thus, there is a need to develop both accurate and interpretable models that can provide meaningful insights within the data. In the next section, we propose interpretable deep learning model to detect AD.

Table 2.2: Accuracy for 11-fold cross validation

Fold	LSTM	BLSTM	Att-LSTM	Att-BLSTM	SVM	RF
1	55.6%	61.1%	77.8%	88.9%	88.9%	88.9%
2	61.1%	72.2%	55.6%	77.8%	66.7%	88.9%
3	94.4%	100%	100%	94.4%	77.8%	77.8%
4	100%	100%	88.3%	100%	83.3%	78.9%
5	100%	100%	94.4%	100%	88.9%	88.3%
6	100%	100%	100%	100%	83.3%	88.9%
7	100%	100%	100%	100%	77.8%	77.8%
8	100%	100%	100%	100%	72.2%	77.8%
9	100%	100%	100%	100%	83.3%	77.8%
10	94.4%	94.4%	88.9%	88.9%	72.2%	72.2%
11	83.3%	72.2%	88.9%	88.9%	83.3%	61.1%
Average	**89.9%**	**90.9%**	**89.9%**	**94.4%**	**79.8%**	**80.3%**

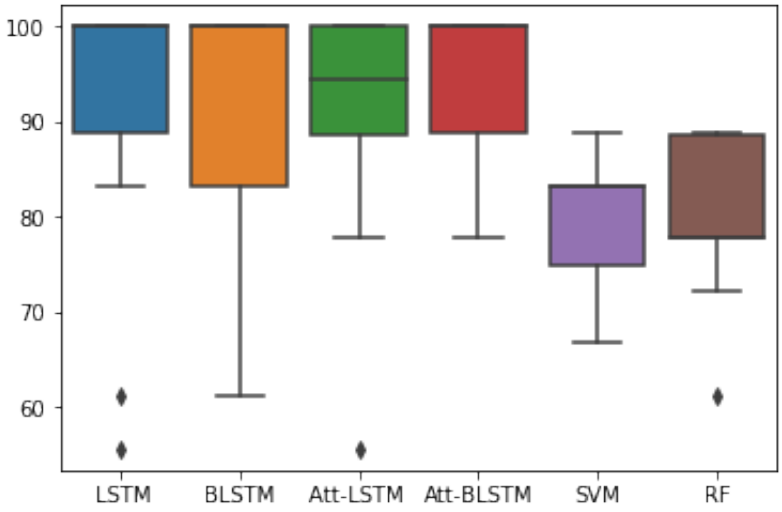


Figure 2.7: 11-fold cross validation accuracy result

2.4.6 Local Interpretable Model-Agnostic Explanations (LIME)

Explaining predictions is necessary to attract users' trust and encourage them to apply these techniques effectively. Understanding the rationale behind the black box model predictions can help users to trust them. Ribeiro et al. [21] proposed a method called LIME, which provides insights into the neighborhoods of a single sample.

LIME is model-agnostic, which means that it can be applied to interpret any machine learning algorithm. It uncovers model behavior by perturbing the input and learns the

relationship between input and output, which is understandable to humans. This method provides local interpretability to understand which features play important roles in the final prediction. This section provides the local interpretable model-agnostic explanations (LIME) for two models, including SVM and LSTM.

To extract insights from our models, we choose two positive (AD subject) and negative (healthy subject) samples and investigate the effect of the most important words in the final prediction for these samples. Figures 2.8 and 2.9 represent the most important words to predict a healthy interview transcript for SVM and LSTM algorithms, respectively. The blue color represents the words that can be indicators of a healthy sample, and orange represents the indicators of memory loss. SVM (Figure 2.8) fails to predict the healthy interview transcript with the prediction probability of 51%, while LSTM correctly predicts the same healthy sample with the prediction probability of 61%.

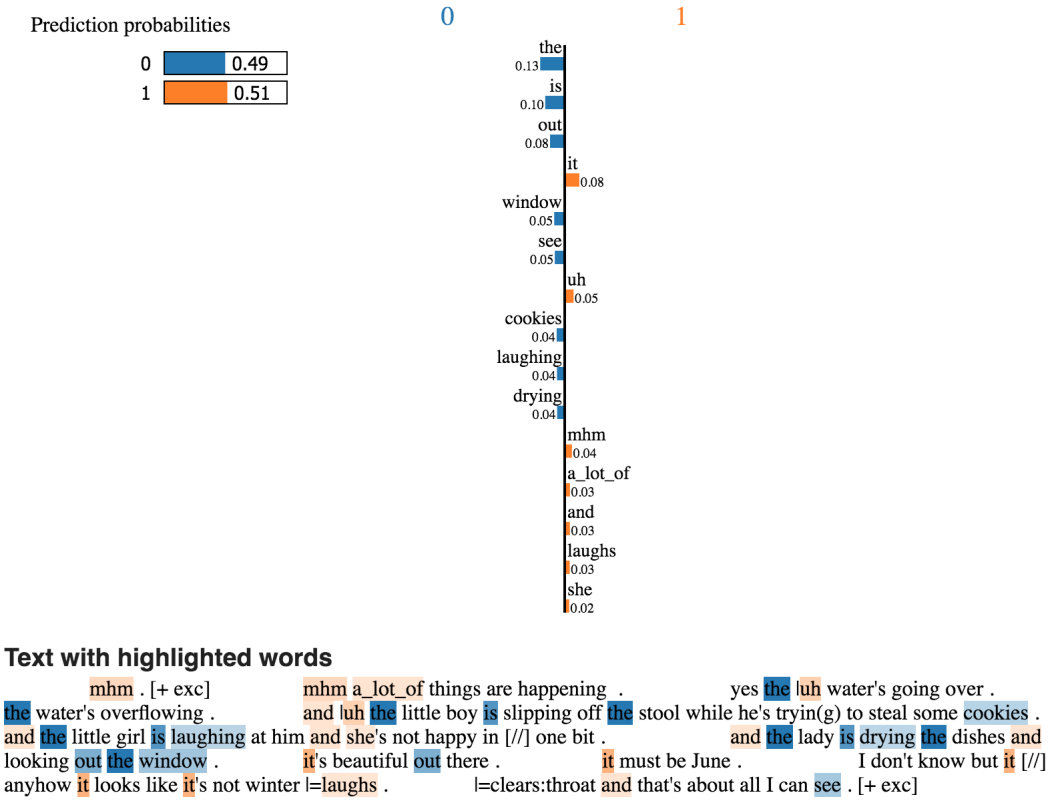


Figure 2.8: SVM text weights for a false positive predicted sample

Also, we present the LIME implementation for a probable AD subject as well. Figures 2.10 and 2.11 represent the most important words for AD prediction using SVM and LSTM, respectively. Again, the LSTM algorithm performs well in terms of prediction with a high prediction probability of 84% (Figure 2.11), while SVM falsely predicts the

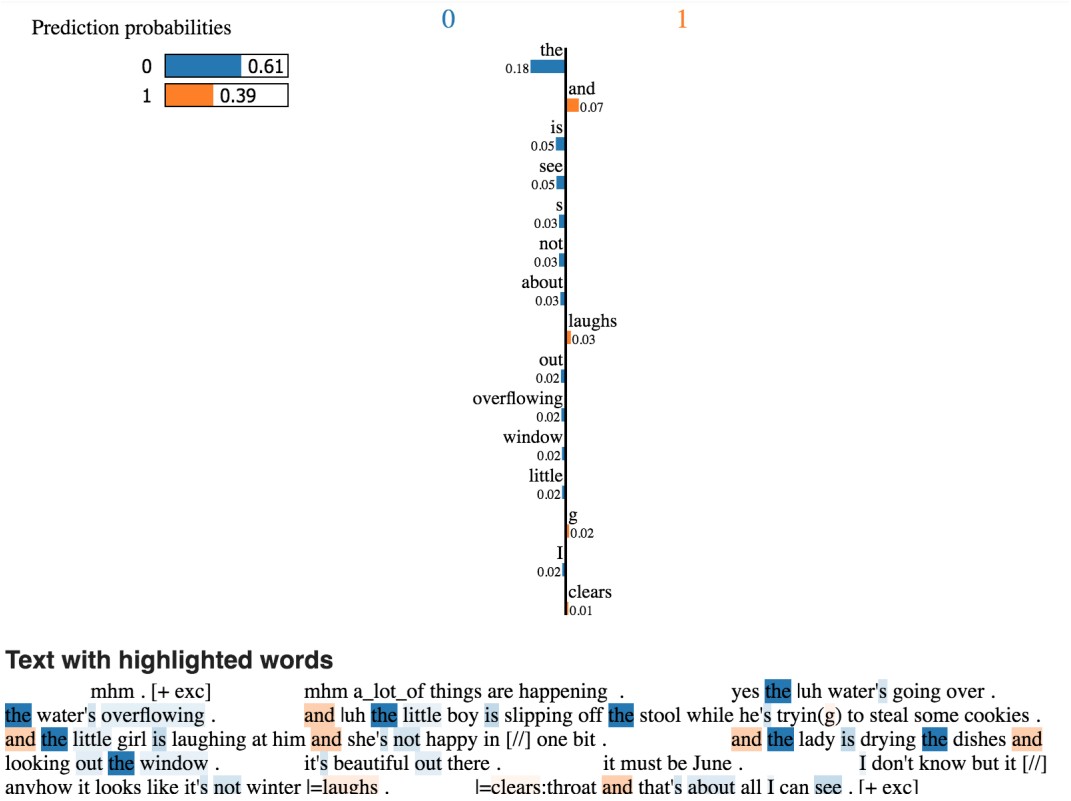


Figure 2.9: LSTM text weights for a true negative predicted sample

AD sample as a healthy one with a probability of 84%.

LIME works reasonably stable when it is used to explain linear classification models; However, they may fall to capture the importance of features for more complex models (i.e., a neural network classifier) [72]. Figure 2.11 confirms that LIME could not obtain the most relevant words that are indicators of AD and memory loss issues. Thus, we design a powerful interpretable model to capture the global importance of words, sentences, and transcripts in a hierarchical structure using attention mechanism, which performs very well based on our results for the same DementiaBank dataset.

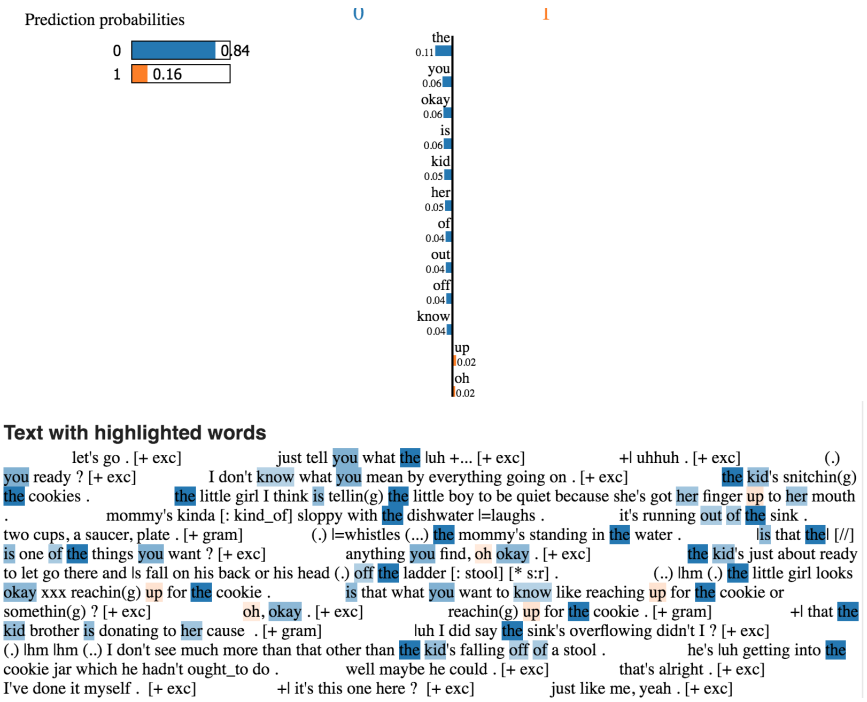


Figure 2.10: SVM text weights for top words in a false negative predicted sample

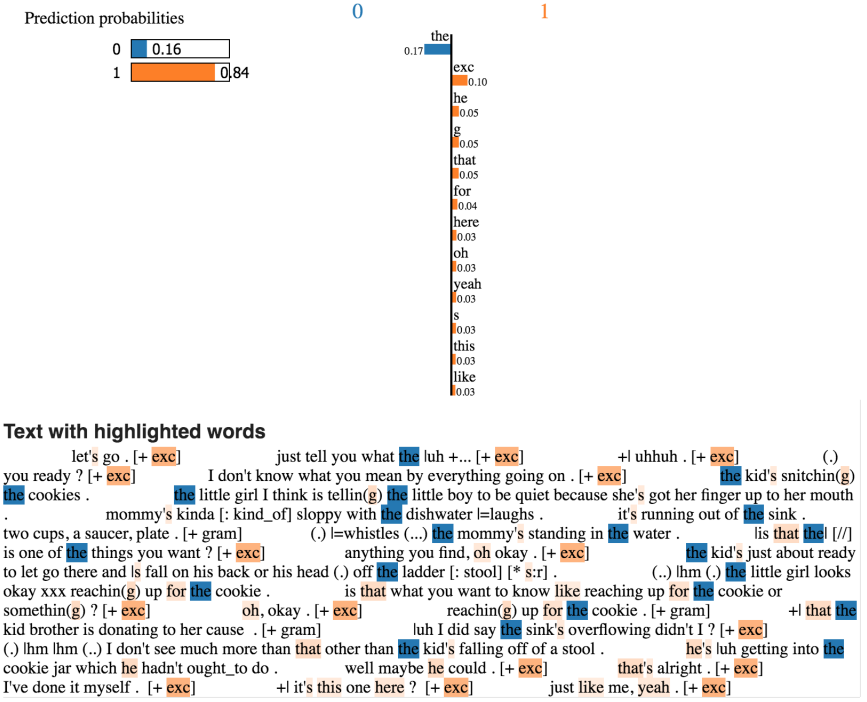


Figure 2.11: LSTM text weights for top words in a true positive predicted sample

2.5 An Interpretable Hierarchical Attention Model

In this section, we combine NLP and hierarchical deep learning models to detect the onset of Alzheimer's disease in patients based on the longitudinal patient interview data from the DementiaBank dataset [1]. These patients were asked to describe a Cookie-Theft picture (Figure 2.3). We propose a three-level hierarchical recurrent neural network with an attention mechanism to develop a powerful and interpretable model that can explain the most important lexical memory-related patterns without the need for any feature engineering.

Our interpretable deep neural network model can help diagnose Alzheimer's disease with high accuracy, obviating the need for expensive imaging tools and invasive methods. While we prove our system's efficacy for Alzheimer's disease, we believe that this type of modeling framework can also be employed to detect other degenerative neurological disorders.

2.5.1 Related Works

In recent years, extracting meaningful insights from black-box deep neural networks has attracted many researchers. For sequential input data, which we consider in this study, researchers have proposed interpretable RNN-based structures for disease diagnosis and prediction [73–75]. They achieved promising results by implementing an attention mechanism to discover where the model concentrates on (attention weights) when making predictions. Ma et al. proposed a diagnostic model based on various attention mechanisms (the Dipole model). The attention layer of their model explains the importance of the RNN model's hidden states as their corresponding attention weights [76]. Sha and Wang proposed a gated recurrent unit (GRU) RNN-based hierarchical attention (GRNN-HA) model. The GRNN-HA model is quite similar to the Dipole model, except that it has a hierarchical structure in which the bi-directional RNN (BRNN) and the attention mechanism in the lower layer encode every medical event within a visit. In the upper layer, the BRNN and the attention mechanism capture the dependencies among the sequence of medical visits [75]. Choi et al. proposed the Reverse Time Attention model (RETAIN), which processes the input in reverse-time order unlike the BRNN structure in the Dipole and the GRNN-HA models [73]. In this section, we extend the theory of interpretable deep neural models by developing novel hierarchical attention over self-attention (AoS) mechanism that sheds light on the importance of each word, sentence, and document in making predictions (Section 2.5.2). The main advantage of our AoS

model is that it prevents loss of information by considering the relation of components (words, sentences, and documents) within the sequence. Numerical results demonstrate that our new AoS model surpasses the regular hierarchical attention-based neural model in terms of accuracy and achieves a new benchmark accuracy for detecting the onset of Alzheimer's using longitudinal interview transcripts.

2.5.2 Methods

In this section, we first describe the components of our models. Then, we explain how we incorporated them in a three-level hierarchical structure to capture both the sequential dependencies and the importance of elements at each level.

2.5.2.1 Basic Components of the Model

We begin by presenting the basic model components:

- **Word Embedding Layer**: A word embedding layer maps each word from the vocabulary set to a high-dimensional vector space using a pre-trained word embedding model. This method captures the context of words such that the words with similar or close semantic meanings have close vector representations. In this study, we used the pre-trained word vectors, GloVe, to obtain the vector representation for each word [56].

- **Contextual Embedding Layer**: Chung et al. [77] found that Gated Recurrent Unit (GRU) [78] cells outperform LSTM cells when working with small datasets. Since our DementaBank dataset is relatively small, we used a GRU-based recurrent neural network (RNN) layer [78] on top of the component embeddings from previous layers to capture the temporal dependencies within every component categories (words, sentences, and transcripts) at each level.

  We placed a GRU in both forward and backward directions (i.e., bidirectional) to capture more information from both past and future utterances, then concatenated the outputs of the two GRUs. Therefore, if the input vector has d dimension, the output of this layer will be $2d$-dimensional.

2.5.2.2 Attention over Self-Attention Mechanism (AoS)

Suppose we have a sentence (S) constructed from a sequence of n word embedding representation vectors (v_i).

$$V = (v_1, v_2, ..., v_n) \tag{2.16}$$

where (v_i) is a u dimensional word embedding for the i-th word in the sentence. In the first layer, a bidirectional GRU is applied to encode the embedding representations as $E^1 = (e_1^1, e_2^1, .., e_n^1)$. Then, the self-attention structure is employed to extract different aspects of the sequence into a vector representation. The self-attention mechanism takes the vector of input (E), and output the vector of weights as α [79]:

$$\alpha = \text{Softmax}(w_2 \tanh (W_1 (E^1)^T)) \tag{2.17}$$

Here, W_1 is the weight matrix with the shape of $d - by - 2u$ (u is the GRU dimension), which is going to be learned in our model. w_2 is a vector of parameters with size d (d is a hyperparameter).

In the second layer, each component is concatenated with the self attentive representation in order to keep the relation information [80].

$$\hat{e}_t = \Sigma_{j=1}^{n} \alpha_j^t e_j^1 \tag{2.18}$$

$$e_t^2 = [v_t, \hat{e}_t] \tag{2.19}$$

Each e_t^2 captures the relationship between v_t and other words in the sentence.

We then apply an attention layer on top of the vector representation $E^2 = (e_1^2, e_2^2, ..., e_n^2)$ to capture the global attention scores (β) of each component (word).

$$\beta = \text{Softmax}(w_3^T \tanh E^2) \tag{2.20}$$

Here, w_3 is the trained weight vector that is going to be learned in this network.

2.5.2.3 Three-level Hierarchical Attention over Self-attention Network

In this section, we first extend the two-level structure of Yang et al. [81] by developing a three-level hierarchical attention mechanism that considers the dependencies between words, sentences, and documents for longitudinal interview transcripts of patients. Then, we propose a novel attention over self-attention (AoS) mechanism to capture the importance of the components constructing the three levels (words, sentences, and documents) of our algorithm.

Recall that our goal is to detect the onset of Alzheimer's disease for each participant of the study using interview transcripts. Suppose that each participant has a sequence of at most N transcripts t_i such that each transcript contains L sentences. We denote each sentence as s_{ij}, which represents the j^{th} sentence in the i^{th} transcript of this patient. Each sentence s_{ij} contains T words such that w_{ijk} represents the k^{th} word in the j^{th} sentence of the i^{th} transcript. In the first level of our algorithm, we use GloVe to obtain a low dimensional representation vector for each word, w_{ijk} as follows:

$$x^1_{ijk} = W_{emb} w_{ijk} + b_{emb}, k \in [1, T], \tag{2.21}$$

where W_{emb} is the embedding matrix obtained through the pre-trained GloVe embedding.

Then we encode each word with GRU-BRNN as h_{ijk} and its hidden representation u_{ijk} as follows:

$$h^1_{ijk} = [\overrightarrow{\text{GRU}}(x^1_{ijk}); \overleftarrow{\text{GRU}}(x^1_{ijk'})], k \in [1, T], k' \in [T, 1], \tag{2.22}$$

$$u^1_{ijk} = \tanh(W^1_w h^1_{ijk} + b^1_w), k \in [1, T]. \tag{2.23}$$

The self-attention weights are computed as follows:

$$\alpha^1_{ijk} = \text{Softmax}({u^1_{ijk}}^T u^1_w), k \in [1, T]. \tag{2.24}$$

The last step in this encoding level is to encode each sentence as a weighted sum of h_{ijk} with the attention scores,

$$s^1_{ij} = \Sigma_k \alpha^1_{ijk} h^1_{ijk}, k \in [1, T]. \tag{2.25}$$

Then, we concatenate the calculated sentence embedding s_{ij} with the word embedding in order to avoid information loss.

$$x^2_{ijk} = [s^1_{ij}, x^1_{ijk}]. \tag{2.26}$$

Since we aim to determine the contribution of each word within each transcript to the overall prediction, we applied a global attention layer on top of the concatenated vector representations (x^2_{ijk}) from the self-attention layer to obtain the attention scores α^2_{ijk} for each word w_{ijk} in this level.

$$\alpha^2_{ijk} = \text{Softmax}(w^T_w \tanh(x^2_{ijk})), k \in [1, T], \tag{2.27}$$

Here, w_w is the trained weight vector and its superscript, T, indicates the transpose of the vector. The attention helps the network to recognize which parts of the sequence play an important role (importance score) in text classification. We use those importance scores to interpret our classification model.

The last step in this level is to encode each sentence as a weighted sum of h^2_{ijk} with the attention scores,

$$s^2_{ij} = \Sigma_k \alpha^2_{ijk} x^2_{ijk}, k \in [1, T]. \tag{2.28}$$

In the second level of this structure, we encode each sentence representation obtained from the first level s^2_{ij} applying GRU-BRNN to incorporate both future and past information within a transcript. Then, we calculate the sentence-level attention α^1_{ij} with the sentence-level context vector u^1_s as follows:

$$h^1_{ij} = [\overrightarrow{\text{GRU}}(s^2_{ij}); \overleftarrow{\text{GRU}}(s^2_{ij'})], j \in [1, L], j' \in [L, 1], \tag{2.29}$$

$$u^1_{ij} = \tanh(W^1_s h^1_{ij} + b^1_s), j \in [1, L], \tag{2.30}$$

$$\alpha^1_{ij} = \text{Softmax}({u^1_{ij}}^T u^1_s), j \in [1, L], \tag{2.31}$$

$$t^1_i = \Sigma_j \alpha^1_{ij} h^1_{ij}, j \in [1, L]. \tag{2.32}$$

We again concatenate the calculated transcript encoding t_i with the sentence encoding obtained from previous level (s^2_{ij}) to maintain the relation of each sentence within the

transcript.

$$s_{ij}^3 = [t_i^1, s_{ij}^2], j \in [1, L]. \tag{2.33}$$

Next, we compute the global attention score for each sentence within each document via the following equation:

$$\alpha_{ij}^2 = \text{Softmax}(w_s^T \tanh(s_{ij}^3)), j \in [1, L]. \tag{2.34}$$

As the last step in this level, we compute the transcript representation with the extracted attention scores as follows:

$$t_i^2 = \Sigma_j \alpha_{ij}^2 s_{ij}^3, j \in [1, L]. \tag{2.35}$$

In the third level, we repeat the same process for the transcript representation t_i^2 computed from the second level to determine the attention score for each transcript in our longitudinal dataset as follows:

$$h_i^1 = [\overrightarrow{\text{GRU}}(t_i^2); \overleftarrow{\text{GRU}}(t_{i'}^2)], i \in [1, N], i' \in [N, 1], \tag{2.36}$$
$$u_i^1 = \tanh(W_t^1 h_i^1 + b_t^1), i \in [1, N], \tag{2.37}$$
$$\alpha_i = \text{Softmax}(u_i^{1T} u_t^1), i \in [1, N], \tag{2.38}$$
$$p^1 = \Sigma_i \alpha_i^1 h_i^1. \tag{2.39}$$

The representation vector for each patient is obtained as:

$$t_i^3 = [p^1, t_i^2], i \in [1, N]. \tag{2.40}$$

Then, we apply the global attention mechanism on top of each transcript encoding as follows:

$$\alpha_i^2 = \text{Softmax}(w_t^T \tanh(t_i^3), i \in [1, N]. \tag{2.41}$$

We obtain the patient representation via:

$$p^2 = \Sigma_i \alpha_i^2 t_i^3, i \in [1, N]. \tag{2.42}$$

Finally, we use p^2 obtained from the last level to build a binary classifier as:

$$g = \sigma(W_y p^2 + b_y) \tag{2.43}$$

The architecture of this algorithm is illustrated in Figure 2.12.

2.5.3 Experiments

2.5.3.1 Dataset

We use the same DementiaBank clinical dataset [1] (Section 2.4.1). However, we consider the whole longitudinal data from DementiaBank including 99 healthy subjects and 169 subjects with probable AD, with annual follow-up visits up to 5 years (i.e., a maximum of 5 visits per individual). The distribution of visits for participants of this study is shown in Table 2.3.

Table 2.3: Distribution of visits per individual type

Visit	AD	Healthy
1	99	99
2	60	74
3	24	45
4	11	17
5	3	8

2.5.3.2 Model Configuration and Training

In our model, first, we broke down each transcript into sentences and tokens. Then, we set aside one-tenth of the training set for validation and obtained the 100-dimensional word embeddings by using the pre-trained GloVe model [56] on both training and validation sets.

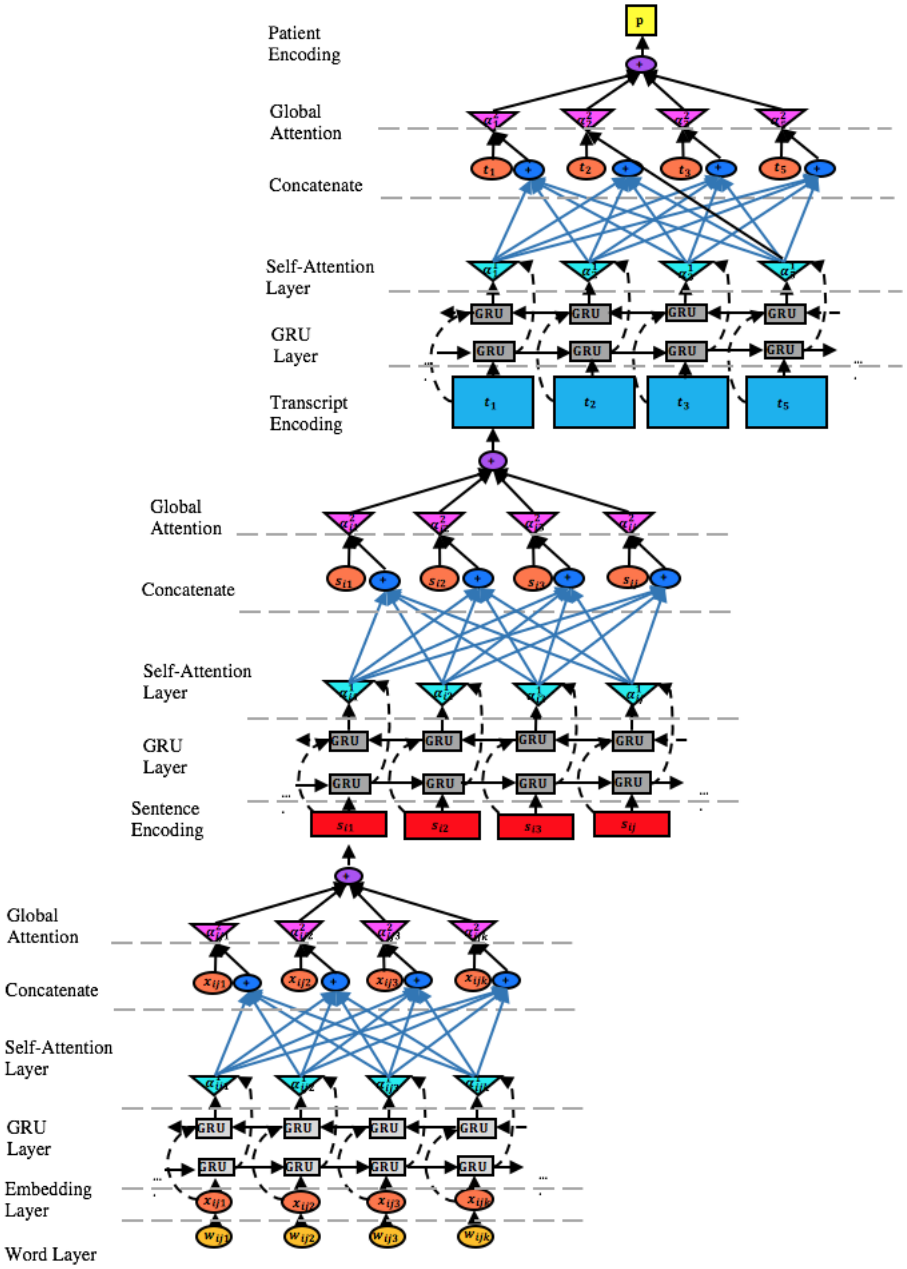


Figure 2.12: Hierarchical attention over self-attention (AoS) structure

The word, sentence, and transcript context vectors were set to have a dimension of 100 and were initialized at random. We set the GRU dimension to 50 for all word, sentence, and transcript levels; hence, the bidirectional GRU has 100 dimensions. For model training, we used a mini-batch size of 10 subjects with the same number of transcripts of five per patient and the same number of 25 sentences per transcript. For those subjects who did not have five visits, we filled the corresponding visit representation vectors with

'nan' so that the embedding matrix gives a big negative number to every element of these vectors.

We used the Adaptive Subgradient (Adagrad) optimizer to train the hierarchical model with an initial momentum parameter of 0.1. We also used 'categorical cross entropy' as the loss function given our goal of binary classification (AD or healthy subjects). Finally, we evaluated our model using the 'accuracy' metric.

2.5.4 Results

To evaluate our model, we conducted three sets of validation experiments. In the first set, we used the balanced dataset from 99 healthy subjects and the first 99 probable AD subjects to provide the same validation setting as Orimaye et al. [19, 43]. We performed 10-fold stratified cross-validation. In each fold, we divided the dataset into 90% (training) and 10% (testing). The validation set was 10% of the training set. Each subject had multiple interview transcripts. We used stratified cross-validation to ensure that each set contains the same ratio of healthy to AD subjects. To evaluate the effect of the GloVe embedding dimension on our model's final performance, we also have tested different GloVe word embeddings using dimensions of 100, 200, and 300. In addition, we compared the performance of our AoS deep neural network model against simpler models, including Support Vector Machine (SVM) and Random Forest (RF). We extracted natural language features based on unigrams and bi-grams of the transcripts. Then, we fed those extracted features into the SVM and RF algorithms.

Table 2.4 shows the results of 10 fold stratified cross-validations for the balanced dataset that consists of 99 healthy and 99 AD subjects. We can see that our three-level hierarchical AoS deep neural network model achieved the best mean accuracy of 0.96 with a standard deviation (SD) of (0.06) across 10-fold cross-validation using a 300-dimensional GloVe word embedding.

Also, the mean (SD) accuracy for SVM was 0.77 (0.09) and for RF was 0.80 (0.10), lower than that of our AoS model, which was 0.96 (0.06) (Table 2.4).

In our second set of experiments, we further investigated the performance of our proposed AoS model on the whole unbalanced DementiaBank dataset, including 169 probable AD subjects and 99 healthy subjects. Tables 2.5, 2.6 and 2.7 present the results of 10 fold stratified cross-validation method using 100, 200 and 300 dimensional GloVe embeddings, respectively. We can see that the best mean accuracy result for an unbalanced DementiaBank dataset is 0.96, which is obtained using 300-dimensional GloVe embedding.

Table 2.4: Testing accuracy scores for 10-fold sratified cross-validation using balanced dataset (99 AD- 99 healthy)

Fold	AoS (GloVe 100)	AoS (GloVe 200)	AoS (GloVe 300)	SVM	RF
1	0.75	0.80	0.80	0.85	0.90
2	0.95	0.90	1.00	0.80	0.85
3	1.00	0.95	1.00	0.80	0.80
4	1.00	1.00	1.00	0.75	0.80
5	1.00	0.95	1.00	0.95	0.95
6	0.95	0.90	0.95	0.70	0.75
7	0.95	0.95	1.00	0.75	0.75
8	0.85	0.95	0.95	0.75	0.85
9	0.95	0.95	0.95	0.60	0.60
10	0.95	0.95	1.00	0.72	0.78
Mean	**0.93**	**0.93**	**0.96**	**0.77**	**0.80**
SD	**0.08**	**0.05**	**0.06**	**0.09**	**0.10**

Table 2.5: Results for 10-fold sratified cross-validation using unbalanced dataset: GloVe 100 (169 AD- 99 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.74	0.86	0.71	0.77	0.86
2	0.96	1.00	0.94	0.97	0.99
3	0.96	1.00	0.94	0.97	0.99
4	0.96	1.00	0.94	0.97	0.99
5	0.93	0.94	0.94	0.94	0.96
6	0.96	1.00	0.94	0.97	0.98
7	0.96	1.00	0.94	0.97	0.98
8	0.96	0.94	1.00	0.97	1.00
9	0.96	0.94	1.00	0.97	0.90
10	0.92	0.94	0.94	0.94	0.99
Mean	0.93	0.96	0.93	0.94	0.96
SD	0.07	0.05	0.08	0.06	0.05

To improve the performance of our model, we got help from text augmentation techniques. In our third sets of experiments, we applied the recent easy data augmentation (EDA) technique on DementiaBank dataset which proved to boost text classification performance [82]. EDA uses four operations of synonym replacement, random insertion, random swap, and random deletion to augment the text data.

Table 2.6: Results for 10-fold sratified cross-validation using unbalanced dataset: GloVe 200 (169 AD- 99 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.74	0.92	0.65	0.76	0.86
2	0.96	0.94	1.00	0.97	1.00
3	0.96	0.94	1.00	0.97	0.99
4	0.93	1.00	0.88	0.94	0.99
5	0.96	1.00	0.94	0.97	0.96
6	0.96	1.00	0.94	0.97	1.00
7	1.00	1.00	1.00	1.00	1.00
8	0.96	0.94	1.00	0.97	0.90
9	0.96	0.94	1.00	0.97	0.99
10	1.00	1.00	1.00	1.00	1.00
mean	0.94	0.97	0.94	0.95	0.97
SD	0.07	0.03	0.11	0.07	0.05

Table 2.7: Results for 10-fold sratified cross-validation using unbalanced dataset: GloVe 300 (169 AD- 99 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.78	0.87	0.76	0.81	0.90
2	1.00	1.00	1.00	1.00	1.00
3	0.93	0.89	1.00	0.94	1.00
4	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00
6	0.96	1.00	0.94	0.97	1.00
7	0.96	1.00	0.94	0.97	1.00
8	1.00	1.00	1.00	1.00	1.00
9	0.96	0.94	1.00	0.97	0.99
10	1.00	1.00	1.00	1.00	1.00
mean	0.96	0.97	0.96	0.97	0.99
SD	0.07	0.05	0.07	0.06	0.03

We augmented the DementiaBank dataset using EDA [82] to build a balanced dataset. We performed one augmentation for the AD group and two augmentations for a healthy group with a changing rate of 10% ($\alpha = 0.1$). We used the augmentation output of 338 healthy and 338 AD subjects to evaluate our AoS model's performance. Tables 2.8, 2.9 and 2.10 present the result of 10 fold cross-validation for 100, 200 and 300-dimensional GloVe embedding using augmented DementiaBank dataset. We can see that the mean accuracy

has improved to 0.98, which is a new benchmark for non-feature-engineering-based models developed on the DementiaBank dataset.

Table 2.8: Results for 10-fold sratified cross-validation using augmented balanced dataset: GloVe 100 (338 AD- 338 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.88	0.88	0.88	0.88	0.96
2	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00
4	0.99	1.00	0.97	0.99	1.00
5	1.00	1.00	1.00	1.00	1.00
6	0.99	1.00	0.97	0.99	1.00
7	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00
10	0.99	1.00	0.97	0.98	1.00
Mean	0.98	0.99	0.98	0.98	1.00
SD	0.04	0.04	0.04	0.04	0.01

Table 2.9: Results for 10-fold sratified cross-validation using augmented balanced dataset: GloVe 200 (338 AD- 338 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.81	0.80	0.82	0.81	0.90
2	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00
10	0.99	1.00	0.97	0.98	1.00
Mean	0.98	0.98	0.98	0.98	0.99
SD	0.06	0.06	0.06	0.06	0.03

To evaluate the interpretability of the model, we extracted attention scores on the three levels (transcripts, sentences, and words). On the transcript level, the model gave different attention score patterns to the transcripts of healthy subjects compared to the

Table 2.10: Results for 10-fold sratified cross-validation using augmented balanced dataset: GloVe 300 (338 AD- 338 healthy)

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.78	0.88	0.65	0.75	0.91
2	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00
10	0.97	1.00	0.94	0.97	1.00
Mean	0.97	0.99	0.96	0.97	0.99
SD	0.07	0.04	0.11	0.08	0.03

transcripts of individuals with Alzheimer's disease. Figure 2.13 depicts the attention score for the five transcripts for both healthy subjects and those with AD. In this boxplot, the middle line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles (i.e., the first and the third quartiles), respectively. The whiskers extend to 1.5 times the inter-quartile range (IQR) from the hinge. Figure 2.13 suggests that for healthy subjects, the model paid more attention to the earlier transcripts, whereas all transcripts were of roughly the same importance for individuals with AD. For healthy subjects, the median attention score for transcript 1 was 0.34 and for transcript 5 was 0.07. On the other hand, for individuals with AD, the median attention scores for transcripts 1 and 5 were 0.20 and 0.18, respectively. Note that, compared to healthy subjects, individuals with AD received lower median attention scores on the first two transcripts and higher median scores on transcripts 3 to 5.

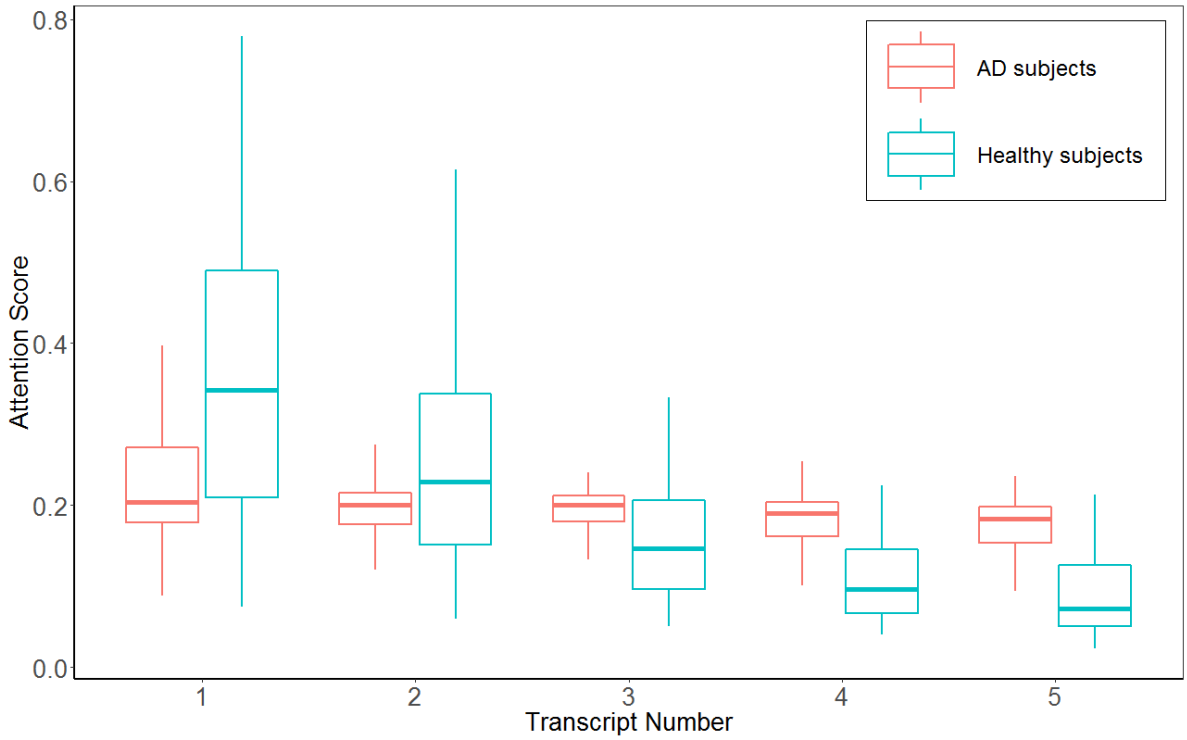


Figure 2.13: Boxplot for transcript-level attention scores

In the sentence level, the model automatically recognizes those sentences that can help to distinguish subjects with AD from healthy individuals. Tables 2.11a and 2.11b present ten sentences of each healthy and AD individuals that had the highest attention scores within the transcript with the highest attention score.

At the word level, we decided to provide example interview transcripts for a healthy subject and an individual with AD (Figure 2.14). In this figure, we have highlighted the words [83] that received a high attention score (> 0.08) in dark color and those that received a medium attention score ($0.04-0.08$) in a light color. Words with a low attention score (< 0.04) were not highlighted in Figure 2.14. This type of analysis can help inform the words that most notably indicate the onset of AD and is of paramount importance to the interpretability of the neural network model. The codes for all experiments are available online at: `https://github.com/marynik66/AD-NLP`

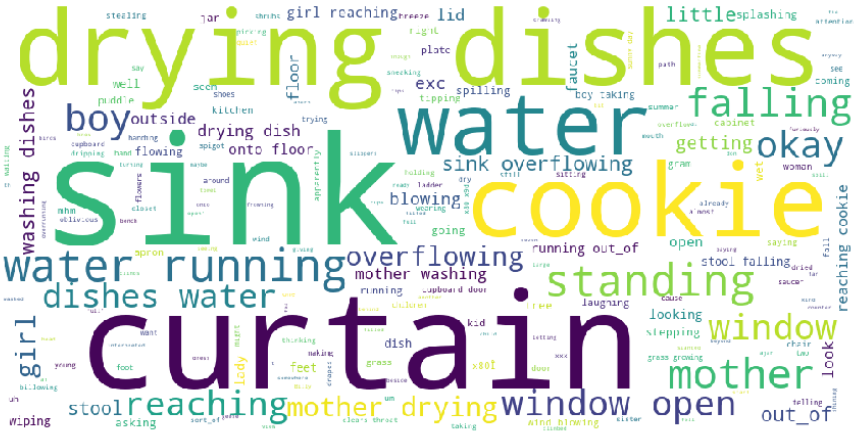
’the woman is washing dishes ’ she’s wearing an apron ’water is pouring out of the sink ’ ’the boy is taking cookies from the cookie jar ’ ’and his sister or young girl is asking for some for her ’ ’the lid is off the cookie jar ’ ’the cupboard door is open ’ ’he is on a three legged stool which is falling uh which is tilted ’ ’there are curtains in the kitchen ’ there’s a tree outside the window ’there are shrubs ’ there’s a walk ’there are two cups on the uh sink counter ’ there’s a plate there ’and as i said the water is overflowing ’ ’it is already on the floor ’ ’there are cupboards above the kitchen counter and below ’ ’and the window is above the kitchen sink ’ ’and the mother has a short sleeve dress on has short hair and is drying the dishes by hand ’ ’curtains on the window that you can see outside the window to the other side of the house gram ’ ’you can see the grass ’ ’you can see the handles on the cupboards ’ ’anything else exc ’

(a) A healthy individual

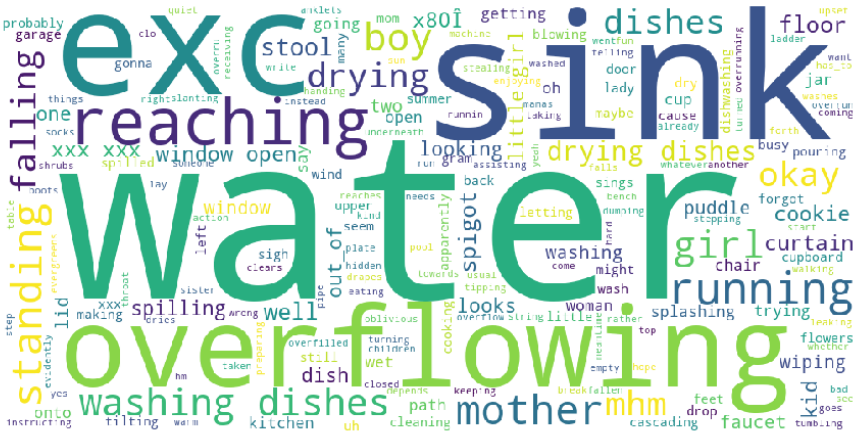
it’s a picture of a kitchen there’s a child reaching on in s r a cookie jar ’ um a male a boy reaching into a cookie jar gram ’ he’s falling ’the stool is falling ’ uh his uh uh a a female youngster maybe sister is reaching up and she’s got a finger to her lip ’the uh mother is at the sink washing a dish ’ ’the water is spilling over from the sink ’ ’the window is open looking out onto uh shrubbery and a path to another house ’ ’there are two di uh two dishes three dishes sitting on the side of the uh on the si cou sink counter ’ ’and there are a number of uh cabinets uh in the picture uh on the sink counter as it stretches around ’ and there’s the uh cabinets above where the boy is reaching in ’mother is washing a dish ’ ’she is dressed in a um oh d dress and it appears she has an apron and a towel in her hand ’ as she’s washing she’s one of her feet is uh obviously in the water the other is somewhat obliterated ’the children are wearing um completely outfitted ’ ’the girl with uh sandals of some kind and the boy it looks like maybe tennis shoes ’ ’ xxx exc ’

(b) An individual with AD

Figure 2.14: Sample transcripts for a healthy individual and an individual with Alzheimer’s disease. Words with high attention score are highlighted in dark blue (healthy) and dark red (AD) and those that received medium attention score are highlighted in light color.



(a) A healthy individual



(b) An individual with AD

Figure 2.15: Word cloud with attention score >0.03

2.5.5 Discussion

In this study, we developed a new three-level hierarchical recurrent neural network (RNN) model and combined it with natural language processing (NLP) to detect the onset of Alzheimer's disease (AD) based on longitudinal patient interview data. We made the black box neural network model interpretable using novel attention over the self-attention (AoS) mechanism. The primary advantage of our AoS model compared to the traditional attention models is that it can prevent information loss by considering the relation of words, sentences, and transcripts within the sequence.

We tried several experiment settings combined with GloVe embedding dimensions $(100, 200, 300)$ and different input data structure (Table 2.13). The hierarchical AoS

model using augmented DementiaBank dataset (AoS-Aug) obtained an overall accuracy of 98%, higher than other similar models developed on the same dataset (Table 2.12). Since our results of using GloVe embedding are very promising (accuracy=98%), we haven't tried other state of the art word embedding methods such as ELMO [59] and BERT [60].

One reason for this outstanding performance is that the AoS model is able to capture the semantic meaning of the words as it can incorporate the relation of the components. We evaluated the attention scores our model gave to transcripts, sentences, and words to shed light on how the neural network model made predictions.

Table 2.12: Comparison of AD detection methods on the interview transcripts of DementiaBank

Model	Accuracy	Precision	Recall	F1	AUC
Wankerl et al. [42]	0.77	-	-	-	-
Fritsch et al. [46]	0.86	-	-	-	-
Chen et al. [45]	0.97	-	-	-	-
Pan et al. [47]	-	0.84	0.85	0.84	-
Fraser et al. [32]	0.82	-	-	-	-
Kong et al. [49]	0.87	0.86	0.90	0.88	-
Orimaye et al. [43]	-	-	-	-	0.83
Karlekar et al. [44]	0.91	-	-	-	-
Orimaye et al. [19]	-	-	-	-	0.93
AoS-Aug (Ours)	**0.98**	**0.99**	**0.98**	**0.98**	**1.00**

Among the maximum of five visits each person could have, the model gave higher attention scores to the earlier visits (Figure 2.13) for healthy subjects and relatively stable scores to the transcripts of individuals with Alzheimer's disease. This can be explained with the degree of language deficit related to AD individuals. If a patient suffers from AD, not only does he/she not remember the same picture the following year, the language deficit may have also been exacerbated due to progression of AD, thus making the later visits as important as the early visits in terms of importance in detecting the disease. On the other hand, for healthy individuals, since they can remember the same picture in the following visits, those visits provide little additional information to the model. So, the model gives lower attention scores to later visits.

At the sentence level, we explored the sentences with the highest attention score

(S_Score) in the transcript with the highest attention score (T_score) for healthy and AD individuals (Tables 2.11a and 2.11b). Unlike the sentences in Table 2.11a, those in Table 2.11b are affected by language deficit, which implies that those individuals may have suffered from memory loss and AD. Also, the sentences in Table 2.11b have more of CHAT disfluency codes (such as those summarized in Table 2.1), which correspond to word repetition, word revision, phrase revision, filled pause, and pause disfluencies.

Finally, at the word level, we can see that the transcript sample of the individual with AD (Figure 2.14(b)) contains particular words that may suggest memory loss, such as 'uh', 'um', 'oh', 'exc', 'maybe', and 'xxx' (unintelligible words), as well as some irrelevant words to the picture (e.g., youngster) (Figure 2.15).

Although our numerical results are promising, they are limited by the size of our dataset. While DementiaBank is one of the best publicly available datasets of its kind, it is a relatively small dataset. Deep learning methods train better on large amounts of data. However, it is common in healthcare to work on smaller datasets, given the complexity of obtaining patient-level data. While we conducted cross-validation to address this issue, further validation of our method on larger datasets is needed to evaluate the performance of our deep learning model.

In summary, we developed unique hierarchical attention over self-attention (AoS) deep recurrent neural network model and demonstrated an important application of it to detect the onset of Alzheimer's disease using longitudinal interview data. We have demonstrated that by employing deep learning, we can unlock the patterns within the natural text that signal memory loss due to Alzheimer's disease. Furthermore, we have illustrated that a three-level AoS mechanism on words, sentences, and transcripts can shed light on how the model comes up with the predictions. This is extremely valuable, in particular for healthcare applications, as transparency of a predictive model is key to its adaptability to everyday clinical practice.

,

Table 2.11: Top 10 sentences with highest attention scores within the transcript with highest attention

(a) Healthy individuals

No.	Sentence	S_Score*	T_Score**
1	&uh the mother's drying dishes but the water is overflowing onto the floor.	0.13	0.41
2	the wind is blowing the curtains .	0.12	0.40
3	the mother is drying dishes and has the water turned on .	0.12	0.41
4	mother is drying dishes and the tap water is overflowing the sink and running on the floor.	0.12	0.41
5	&uh the mother is drying the dishes as the sink faucet has filled the sink bowl and is running over onto the floor.	0.11	0.54
6	mother [//] the reason the water's flowing out over the sink is because the water is running furiously &um and I'm looking out through the window.	0.11	0.42
7	at the point she's drying dishes <the water> [//] perhaps from the noise the water is spilling over the sink and onto the floor.	0.11	0.76
8	the mother's spilling the water and also drying the dishes.	0.11	0.77
9	'the boy is reaching for cookies and the stool is falling over.	0.11	0.37
10	the mother's drying the dishes, frowning but not turning off the faucet.	0.11	0.64

* Sentence attention score
** Transcript attention score

(b) AD individuals

No.	Sentence	S_Score*	T_Score**
1	and she's havin(g) problems because the sink's running over and she's standing in a puddle of water, some empty dishes on the counter.	0.10	0.31
2	she's drying [//] washing and drying dishes.	0.10	0.47
3	the tea cloth is drying the dishes.	0.10	0.58
4	well he's reachin(g) for the cookie &=laughs but he's handing the cookie to her", 'in the [/] the meantime the stool is falling over.	0.10	0.27
5	and mama's drying the dishes as usual for mamas (.)	0.10	0.25
6	&hm &hm (..) I don't see much more than that other than the kid's falling off of a stool.	0.10	0.27
7	and the [//] I guess it's the mother is drying dishes .	0.09	0.47
8	but the water is &flow still flowing .	0.09	0.40
9	<and the mother is> [//] well she's spilling her water which is not very good but she's doing [//] washing dishes and drying them.	0.08	0.21
10	the water is overflowing from the faucet into the sink onto the floor, <while she wipes> [//] &uh while she dries &uh a dish.	0.07	0.49

* Sentence attention score
** Transcript attention score

Table 2.13: Comparison between different experiment settings of our AoS model using 10-fold cross-validation; mean(SD)

Model	Data	GloVe	Accuracy	Precision	Recall	F1	AUC
AoS	Balanced (99-99)	100	0.93 (0.08)	0.93 (0.07)	0.94 (0.11)	0.93 (0.08)	0.98 (0.04)
AoS	Balanced (99-99)	200	0.93 (0.05)	0.95 (0.05)	0.91 (0.10)	0.93 (0.06)	0.98 (0.03)
AoS	Balanced (99-99)	300	0.96 (0.06)	0.96 (0.07)	0.97 (0.07)	0.97 (0.06)	0.98 (0.03)
AoS	Unbalanced (169-99)	100	0.93 (0.07)	0.96 (0.05)	0.93 (0.08)	0.94 (0.06)	0.96 (0.05)
AoS	Unbalanced (169-99)	200	0.94 (0.07)	0.97 (0.03)	0.94 (0.11)	0.95 (0.07)	0.97 (0.05)
AoS	Unbalanced (169-99)	300	0.96 (0.07)	0.97 (0.05)	0.96 (0.07)	0.97 (0.06)	0.99 (0.03)
AoS (Augmented)	Balanced (338-338)	100	0.98 (0.04)	0.99 (0.04)	0.98 (0.04)	0.98 (0.04)	1.00 (0.01)
AoS (Augmented)	Balanced (338-338)	200	0.98 (0.06)	0.98 (0.06)	0.98 (0.06)	0.98 (0.06)	0.99 (0.03)
AoS (Augmented)	Balanced (338-338)	300	0.97 (0.07)	0.99 (0.04)	0.96 (0.11)	0.97 (0.08)	0.99 (0.03)

Chapter 3
Detection of Alzheimer's Using Audio Processing

3.1 Introduction

Language impairment happens early in Alzheimer's disease years prior to the major symptoms of AD [84]. In fact, AD deteriorates neurons and their connections in regions of the brain that are involved in memory and language [85]. Recent studies have identified certain vocal features that may signal the early signs of Alzheimer's disease [86, 87]. Meilan et al. [87] found acoustic speech parameters that are indicators of AD early stages and their relation with linguistic deficit of AD patients. Their study justifies the need for developing an acoustic diagnostic test for AD which can ultimately lead to saving costs (i.e, time, resources) for both patients and healthcare providers.

Several studies have attempted to develop tools for automatic detection of AD with linguistic information [32, 42, 45]. However, they are not able to achieve the same performance level in other languages. On the other hand, preparing linguistic datasets often needs extra tasks such as transcribing the audio recordings, which is labor expensive and time-consuming.

Furthermore, the voice of AD patients contains valuable information and clues of degenerative cognitive ability [88] that textual transcripts may not be able to capture. Muscle weakness caused by brain problems due to dementia changes a person's vocal cord or throat, ultimately leading to weak, hoarse, scratchy voice as the disease progresses [89].

For these reasons, we focus on non-linguistic approaches to detect Alzheimer's disease automatically using only the audio recordings of patients. Since deep neural networks have been successful in speech recognition and linguistic tasks [90, 91], we aim to develop and employ several deep neural network architectures to detect the onset of AD.

3.2 Related Works

A number of studies attempted to automatically detect Alzheimer's disease using linguistic information [32, 42, 45]. However, linguistic models are hard to be applied in different languages. Some studies used both linguistic and acoustic features of patients' speech to detect the onset of AD ([19, 92–94]). Roark et al. [95] derived markers from both the audio and transcript of a spoken narrative recall task to detect mild cognitive impairment (MCI) automatically. Chakraborty et al. [96] investigated different types of acoustic features and feature selection rules to identify the dementia stage. They achieved an accuracy of 82% using score-level fusion. Recently, Warnita et al. [97] developed a deep learning model to detect Alzheimer's disease using gated convolutional neural networks by extracting acoustic features from patients' audio recordings. They achieved an accuracy of 73.1% using a 10-fold cross-validation approach. Liu et al. [98] proposed a simulation model that collects new speech dataset for Alzheimer's disease. They used different machine learning models for their experiments, and show that a model based on logistic regression obtained the best performance.

We focus on a non-linguistic approach by developing end-to-end deep neural networks using only speech audio of the patients without any preprocessing requirement such as feature engineering.

Recently, developing very deep convolutional neural networks (CNNs) such as Alexnet [5], Resnet [99], VGG [100] with the presence of large image datasets (e.g, ImageNet [101]) has significantly improved the performance of image classification tasks. In recent years, the application of CNNs has been extended to automatic audio processing tasks like environmental sound classification [100,102,103], music genre classification [104] and music tagging [105]. The common method for automatic audio classification is to transform the audio waveforms into two-dimensional spectro-temporal representations, which are then used as inputs to two-dimensional convolutional neural networks (2D CNN) for supervised classification [106,107]. The spectrograms are two-dimensional representations of audio waveforms that shrink the high dimensionality of the original audio waveforms. Mel-frequency cepstral coefficients (MFCC) are power spectrograms that frequently use representations for audio speech emotion recognition and classification [106]. Early works on audio classification tasks generally extract MFCC features and use a classification algorithm, such as support vector machine (SVM), for classification [108]. The emergence of big audio training datasets made it possible to train deep CNNs similar to image classification [106].

3.2.1 One-Dimensional CNN

A one-dimensional CNN (1D CNN) can directly model audio waveforms in time frameworks. 1D CNN structure is mainly used to develop end-to-end learning methods without feature engineering on raw audio waveforms [99, 109]. Zeghidour et al. [110] proposed an end-to-end 1D CNN architecture using raw waveforms, which learns features by trainable filterbanks instead of regular Mel-filterbanks. Dai et al. [2] proposed very deep 1D CNN models with raw audio waveforms as their inputs, which achieved the best accuracy of 72% on the UrbanSound8k dataset. Recently, Abdoli et al. [3] proposed an end-to-end 1D CNN architecture for environmental sound classification. They achieved a higher accuracy level of 89% using Gammaton filterbanks.

3.2.2 Transfer Learning

Transfer learning is a method that can transfer the knowledge (trained weights) obtained from a reasonably large training dataset to a problem with smaller training data [111]. The main idea behind this method is that the initial layers in deep neural networks are learning the generic characteristic of data, which can be transferred to a similar dataset and learning more specific features in the deeper layers which are not transferable to the new problem.

3.2.3 Our Work

In this chapter, we have employed several architectures based on deep neural networks. First, we tried the end-to-end one-dimensional convolutional neural networks with the whole interview audio length as its input. We then improved the performance of the same model by segmenting the audio interview into two seconds' and five-seconds' waveforms.

Finally, we propose a novel interpretable audio recognition model based on audio transfer learning. Our model detects Alzheimer's disease with an outstanding prediction performance and recognizes the memory loss and language changes over the course of the disease. In this novel model, the raw audio waveforms of individuals' interviews are used without any extra hand-crafted preprocessing methods, which can facilitate the audio recognition problem.

3.2.4 Main Contributions

This chapter's main contributions are as follows: (1) we developed a new hierarchical deep model using transfer learning to predict the onset of AD using the raw audio interviews of patients. (2) We achieved the highest accuracy of 90% compared to similar works on the same dataset (Warnita et al. [97]). (3) Our model is interpretable, which can capture the cognition deficiencies in patients' speech very well.

3.3 Dataset

In this study, we used the audio sets of Pitt Corpus [1] in the DementiaBank dataset, where patients were asked to describe incidents happening in a picture, namely, the Cookie Theft Picture of the Boston Diagnostic Aphasia Examination. Pitt Corpus contains both text and audio of narrative speech from 169 AD subjects and 99 healthy subjects. Data were collected longitudinally, every year with at most five subsequent years. There are 309 AD interview visits for probable AD subjects and 243 visits for healthy individuals that are collected from multiple visits.

In the first phase of this analysis, we are using the total patients visits without considering the dependencies of annual visits (309 AD and 243 healthy audio interview visits). In the second phase, however, we consider longitudinal visits for each individual (169 AD and 99 healthy audio interview visits) and develop an interpretable hierarchical audio model using longitudinal data.

3.4 Methods

3.4.1 1D Convolutional Neural Networks using Raw Audio Dataset

Dai et al. [2] proposed a very deep one-dimensional convolutional neural network (CNN) using the raw audio waveforms as the input to their networks (Figure3.1). They proposed five CNN structures that vary in depth and weight layers. The main idea to develop such deep models with fewer parameters in each layer is to use very small receptive fields for all the layers except the first one. Depending on the sampling rate of audio waveforms (e.g., 8000Hz), the audio input may have a large number of samples, making the learning layer-expensive to start with small receptive fields in the first convolution layer. Dai et al. [2] chose the size of the receptive in the first layer such that it can cover 10-millisecond

duration — the same window size as the audio MFCC feature extracting methods.

They used the UrbanSound8k dataset to classify ten urban environmental sounds [112]. The dataset contains 8732 audio recordings with a maximum length of four seconds. They have normalized and down-sampled the audio waveforms to 8 kHz using as the input to their designed CNN structures (32000 samples per 4-second audio waveform). The highest accuracy level they achieved was 71.8% for their deep model with 18 convolution layers.

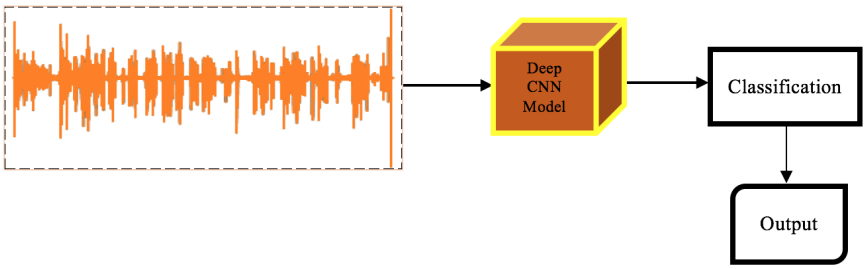


Figure 3.1: Audio classification using Dai et al. [2] model

We start our experiments using the M5 model from Dai et al. [2] which has a small network (five convolution layers) to evaluate the performance of the 1D CNN on our Alzheimer's audio recordings. The details of convolutional layers of this model are presented in Table 3.1. To replicate the same structure, we down-sampled our interview audio recordings (DementiaBank) from 22050 to 8000 samples per second. Then, we truncated and padded all audio waveforms to 90 seconds long ($90 \times 8 = 720000$ samples per audio). Compared to Dai et al.'s implementation [2], where they used 4-second audio wavelengths, our input contains audio recordings of waveforms with 90 seconds length. We employed the M5 model with the entire audio recordings from the DementiaBank dataset (309 AD and 243 healthy audio recordings).

We implemented an 11-fold stratified cross-validation to evaluate the performance of M5 model. In each fold, we trained on 450 (or 451) visits, validated on 51 visits, and tested on the remaining visits. The batch size was 20, and the number of epochs was 10. The results of our implementation are presented in Table 3.2.

Table 3.1: 1D CNN structure [2]. Input dimension: (720000,1)

Layer	Kernel size	Stride	Dimension
Conv1	80	4	128
Maxpool1	4	-	128
Conv2	3	1	128
Maxpool2	4	-	128
Conv3	3	1	256
Maxpool3	4	-	256
Conv4	3	1	512
Maxpool4	4	-	512
Conv5	3	1	512
Global average pooling	1	-	512

Table 3.2: Testing accuracy for 11-fold cross-validation on 90-second audio waveframes

Fold	Accuracy
1	0.59
2	0.59
3	0.56
4	0.70
5	0.64
6	0.78
7	0.44
8	0.76
9	0.78
10	0.56
11	0.44
Mean	0.62

According to Table 3.2, the mean accuracy result among the 11 folds is 62%. Since the wavelength of our audio inputs is pretty large (90 seconds) for the M5 structure, which was initially implemented on smaller input wavelengths (4 seconds), we now improve the accuracy of the M5 model by segmenting the audio recordings into smaller audio waveforms. Section 3.4.2 illustrates the proposed method.

3.4.2 1D CNN on Small Audio Frames

Motivated by the approach taken by Abdoli et al. [3], we extended the 1D CNN implementation on the small audio waveforms. We first split the 90-second audio recordings into disjoint frames with different lengths of 0.5, 1, 2, 3, and 5 seconds for all 309 AD and 243 healthy audio recordings. To maximize the use of audio information, we conducted a separate experiment by segmenting the audio waveforms in a way that successive audio frames are overlapped with a fixed percentage of the frame length (Figure 3.2). Using these overlapped frames increases the number of input frames for 1D CNN. In this experiment, we split all the audio waveforms of both AD and healthy subjects into 2-second wavelength frames, considering one second overlapped length.

We labeled all the AD audio frames as one and healthy ones as zero. We evaluated the performance of M5 model on these small audio waveframes with different lengths using 11-fold stratified cross-validation. In each fold, we used 10% of data for testing and 90% for training/validation. The validation set was 10% of the training/validation set. Since we have an unbalanced dataset, the stratified cross-validation ensures that each fold has the same proportion of each class. The batch size was 32, and the number of epochs was 10. Table 3.3 waves summarizes the 11-fold cross-validation results on the test set.

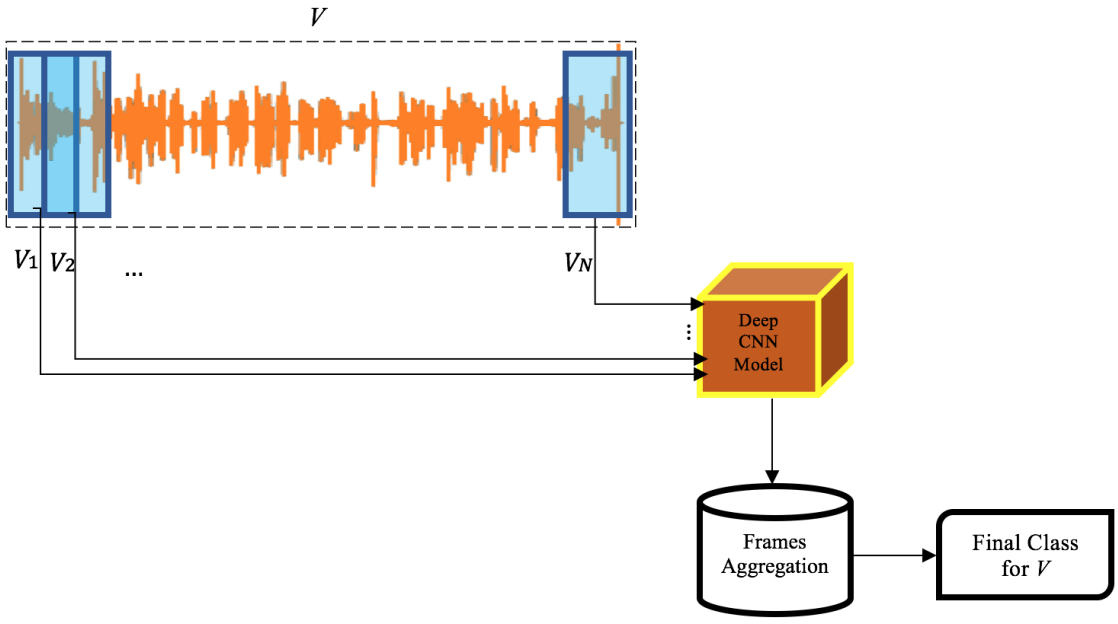


Figure 3.2: Audio frames aggregation [3]

In Table 3.3 we can see that the average accuracy level among 11 folds for the dataset with 2-second disjoint waveforms has the highest value (0.63). Thus, using overlapped audio frames (2-sec(1-sec)) helped to improve the performance of the M5 model.

Table 3.3: Testing accuracy for 11-fold cross-validation on different audio frame lengths

Fold	0.5-sec	1-sec	2-sec	3-sec	5-sec	2-sec (1-sec)
1	0.52	0.51	0.51	0.44	0.57	0.48
2	0.54	0.56	0.51	0.49	0.47	0.52
3	0.52	0.55	0.53	0.56	0.51	0.56
4	0.51	0.53	0.62	0.58	0.43	0.56
5	0.52	0.62	0.58	0.64	0.49	0.50
6	0.56	0.59	0.50	0.54	0.63	0.71
7	0.54	0.72	0.55	0.58	0.71	0.65
8	0.62	0.64	0.77	0.66	0.51	0.70
9	0.64	0.67	0.79	0.79	0.82	0.83
10	0.65	0.66	0.72	0.66	0.57	0.83
11	0.69	0.72	0.83	0.76	0.74	0.83
Mean	0.57	0.62	0.63	0.61	0.59	0.65

3.4.3 Aggregation of audio frames

Since each visit's audio recording V is divided into N waveforms, $V_1, V_2, ..., V_N$, we need to have an aggregation method to obtain the final classification result for each V from 1D CNN prediction of its constructing waveforms (Figure 3.2). To achieve the final prediction for each audio recording V, we conducted the thresholding rule such that the average of predictions as labeled "1" (or AD) should be greater or equal to the threshold value. This thresholding rule is denoted as the following equations:

$$y = \frac{1}{N}\Sigma_j p_j, j \in [1, N], \tag{3.1}$$

$$y \geq T. \tag{3.2}$$

where p_j is the 1D CNN prediction for the $j = 1, ..., N$ frame of the audio recording V and N is the number of frames constructing V. T is the predefined threshold value.

We employed 11-fold cross-validation to evaluate the result of aggregating the 1D CNN prediction outcomes on small audio frames. The thresholding value (T) was varied from 0.1 to 0.9 in 0.1 increments. Table 3.4 reports the mean results of 11-fold cross-validation for the overall prediction value of each audio interview recording at different frame lengths using the thresholding rule.

We can see that $T \geq 0.7$ gives the best accuracy results for all experiments with different frame lengths. The experiments with the 2-second frame with 1-second overlapped frames achieved the highest mean accuracy of 71% at a threshold of 0.7. Furthermore, the 2-second disjoint frames achieved a mean accuracy close to that (67%). These results indicate that the frames shorter than 2 seconds cannot capture the disfluencies in the speech of patients with AD adequately. Also, if we increase the frame size from 2 seconds to 3 seconds or more, the total number of frames will decrease, which may impact the model's performance due to the reduced number of frames.

Table 3.4: Results for mean of 11-fold cross validation using varying threshold values

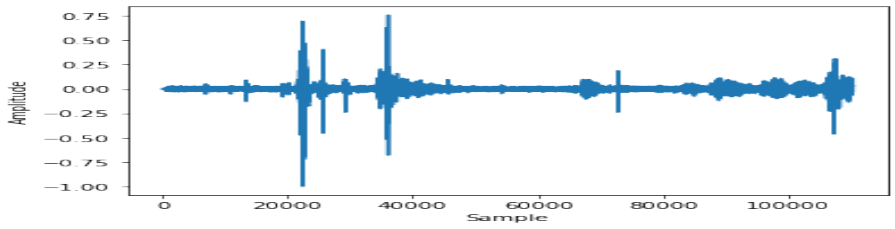
Frame Length	T=0.1	T=0.2	T=0.3	T=0.4	T=0.5	T=0.6	T=0.7	T=0.8	T=0.9
0.5-sec	0.57	0.57	0.58	0.57	0.57	0.56	0.59	0.63	0.59
1-sec	0.57	0.57	0.59	0.62	0.61	0.62	0.65	0.68	0.68
2-sec	0.58	0.60	0.64	0.65	0.66	0.66	0.67	0.65	0.61
3-sec	0.57	0.58	0.61	0.61	0.63	0.64	0.65	0.62	0.59
5-sec	0.56	0.58	0.59	0.60	0.61	0.62	0.63	0.59	0.54
2-sec(1-sec)	0.61	0.66	0.68	0.67	0.67	0.69	0.71	0.69	0.61

Due to the small size of Alzheimer's dataset in Dementiabank, we are motivated to get help from transfer learning methods that are trained on the larger image or audio datasets. Transfer learning models can ultimately improve deep neural networks using limited data. We employ the audio-based transfer learning approach [106], which is pre-trained on a large audio dataset by using the raw audio frames and extracting the mel-spectrogram features [113].

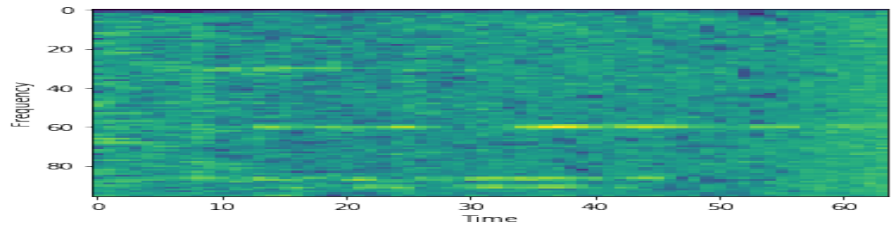
3.4.4 VGGish Transfer Learning

To overcome the small size of our dataset, we applied transfer learning using a pre-trained VGGish model [106], which is trained on the Audioset dataset [113]. The VGGish model is trained on 2 million human-labeled 10-second YouTube-8M and published by Google [106].

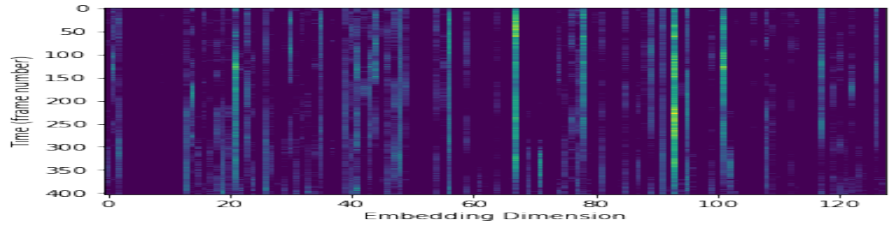
We used VGGish to generate audio feature embeddings from audio recordings. First, the raw audio recordings are resampled to 16 kHz mono, and a spectrogram is generated using the Short-Time Fourier Transform with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window. This spectrogram is used to compute a log mel spectrogram with 64 mel bins covering the range 125-7500 Hz. The final audio features contain 64 mel bands and 96 frames of 10 ms each. The log mel spectrogram tensors are fed to the VGGish model as inputs. The output of the VGGish model is 128-dimensional feature vector for every 1 second time bin (Figure 3.3).



(a) Audio wave



(b) Mel-frequency cepstral coefficients (MFCCs)



(c) VGGish feature embedding

Figure 3.3: An example of 5 second frame of Dementiabank audio wave and its corresponding MFCC and trained VGGish feature embedding (sampling frequency=22050)

Once audio features are extracted using a pre-trained Vggish model, we can use a classifier to detect the onset of AD using only the audio recordings of subjects (Figure 3.4).

This section proposes a novel interpretable hierarchical deep neural network model for the Alzheimer's audio classification problem. Our model not only achieves a new benchmark accuracy performance for detecting AD but can also tell clinicians which

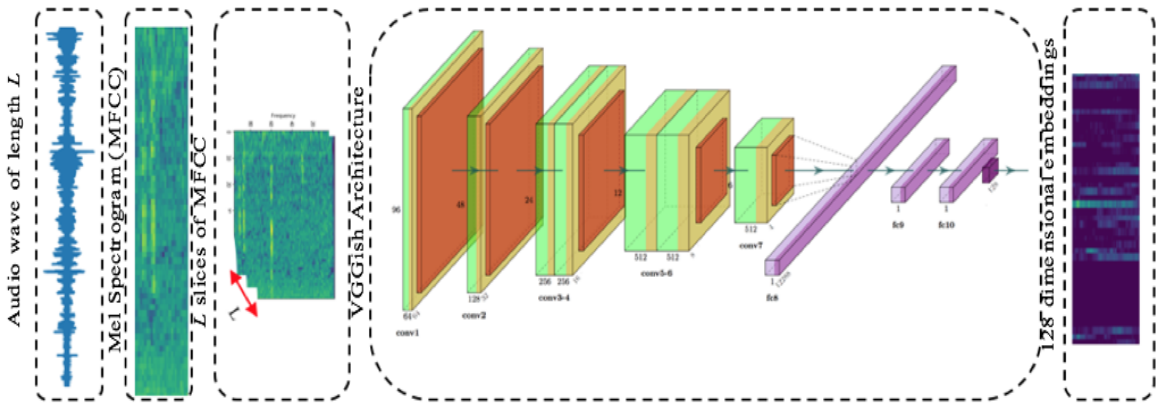


Figure 3.4: VGGish transfer learning feature extracting mechanism

parts of the audio contain signals of memory loss and linguistic deficit due to Alzheimer's disease.

To compare the performance of our hierarchical deep neural network model with other models, we first implemented the deep neural network models including bidirectional gated recurrent neural networks (BGRU), attention-based bidirectional gated recurrent neural networks and the support vector machine (SVM) on the audio feature embeddings obtained from VGGish transfer learning model.

3.4.5 Bidirectional GRU Using VGGish Feature Embeddings

We fed the outputs of VGGish embeddings to bidirectional GRU (BGRU) to take advantage of the presence of both past and future information in an audio recording. We set one BGRU layer with a size of 50. The batch size is set to 32. The last layer is set to have two classes with the Softmax activation function. Table 3.5 presents the results for 11-fold cross validation using BGRU.

3.4.6 Attention based Bidirectional GRU Using VGGish Feature Embeddings

We incorporated the attention layer just after the BGRU structure to consider the importance of audio sequence components for the sake of interpretability and better prediction outcomes. Table 3.6 presents the results for 11-fold cross validation using Attention based BGRU.

Table 3.5: 11 fold cross-validation using Bidirectional GRU

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.54	0.64	0.29	0.40	0.54
2	0.52	0.58	0.29	0.39	0.65
3	0.65	0.67	0.67	0.67	0.72
4	0.57	0.60	0.50	0.55	0.69
5	0.72	0.75	0.65	0.70	0.75
6	0.82	0.89	0.74	0.81	0.93
7	0.87	0.84	0.91	0.87	0.94
8	0.93	0.92	0.96	0.94	0.99
9	0.91	0.88	0.96	0.92	0.98
10	0.87	0.81	0.96	0.88	0.96
11	0.84	0.81	0.91	0.86	0.96
Mean	**0.75**	**0.76**	**0.71**	**0.72**	**0.83**

Table 3.6: 11-fold cross validation results using Attention based Bidirectional GRU

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.50	0.52	0.58	0.55	0.62
2	0.65	0.70	0.58	0.64	0.66
3	0.63	0.63	0.71	0.67	0.75
4	0.76	0.81	0.71	0.76	0.85
5	0.83	0.89	0.74	0.81	0.87
6	0.96	0.96	0.96	0.96	0.98
7	1.00	1.00	1.00	1.00	1.00
8	0.98	1.00	0.96	0.98	0.98
9	0.96	0.96	0.96	0.96	1.00
10	0.98	0.96	1.00	0.98	1.00
11	0.98	0.96	1.00	0.98	0.97
Mean	**0.84**	**0.85**	**0.84**	**0.84**	**0.88**

3.4.7 Support Vector Machine (SVM) Using VGGish Feature Embeddings

To compare the deep learning models with traditional machine learning algorithms, we decided to employ a linear Support Vector Machine(SVM) algorithm on VGGish pre-trained audio embeddings to test whether the prediction can have better performance. The results for 11-fold cross validation using SVM are in Table 3.7.

Table 3.7: 11-fold cross validation results using SVM

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.63	0.77	0.42	0.54	0.64
2	0.50	0.53	0.42	0.47	0.50
3	0.48	0.50	0.33	0.40	0.48
4	0.65	0.79	0.46	0.58	0.66
5	0.54	0.56	0.43	0.49	0.54
6	0.71	0.75	0.65	0.70	0.71
7	0.76	0.73	0.83	0.78	0.75
8	0.71	0.68	0.83	0.75	0.71
9	0.53	0.53	0.70	0.60	0.53
10	0.51	0.52	0.48	0.50	0.51
11	0.60	0.59	0.70	0.64	0.60
Mean	**0.60**	**0.63**	**0.57**	**0.58**	**0.60**

3.4.8 Hierarchical Deep Audio Model

The VGGish transfer learning model outputs 128-dimensional feature embedding for every second of input audio waveforms. To capture the dependencies between waveforms, we employed a Gated Recurrent Unit (GRU)-based recurrent neural network (RNN) layer [78]. Since we have access to both past and future audio waveforms, we used bidirectional GRU (BGRU) to capture more information and obtain better accuracy performance. Figure 3.5 demonstrates how longitudinal input raw audio recordings go through the VGGish model to extract audio embeddings for the proposed hierarchical structure.

To recognize which parts of the audio plays an important role in the final classification, we used the attention layer right after the BGRU layer. Attention-based neural networks have been performing well in speech recognition and machine translation [6, 66].

The attention layer makes a weighted sum of BGRU output vector (H) that is used to produce the audio encoding representation (r) as follows:

$$M = tanh(H) \tag{3.3}$$

$$\alpha = softmax(w^T M) \tag{3.4}$$

$$r = H\alpha^T \tag{3.5}$$

where w is a BGRU-trained weight vector, and T indicates the transpose of the

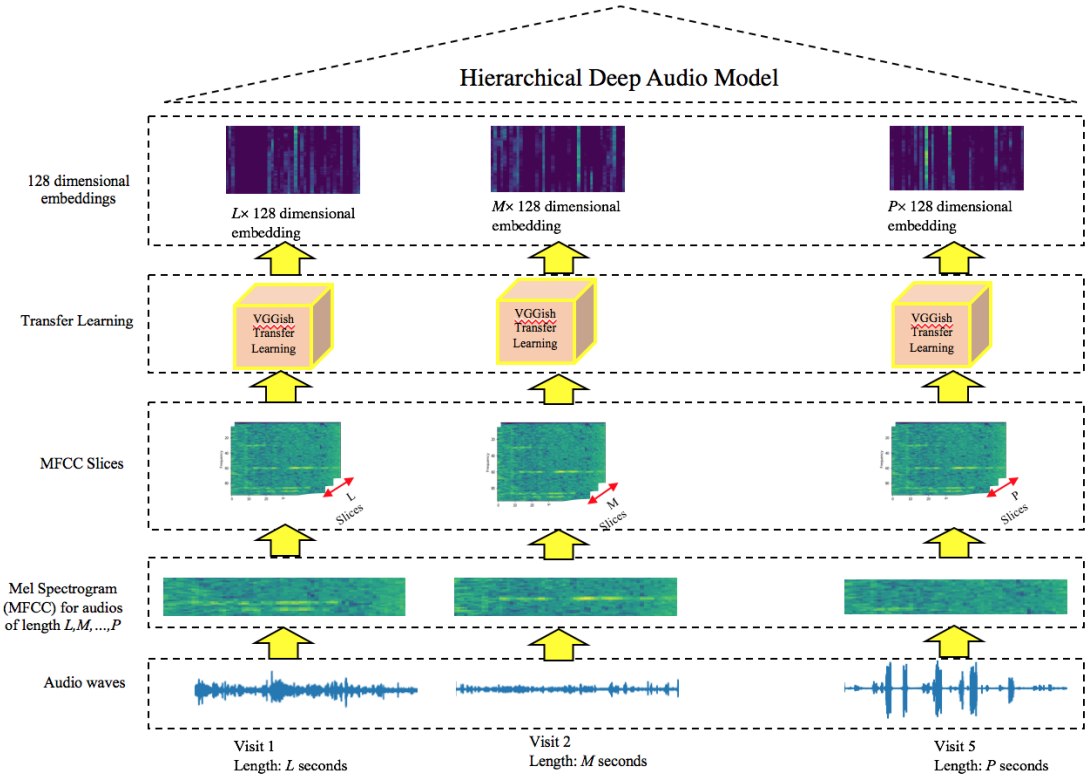


Figure 3.5: VGGish transfer learning on longitudinal audio inputs

vector.

The first level of our hierarchical model captures the dependencies between one-second audio waveforms to build five-second audio encodings. The second level models the dependencies between five-second audio encodings obtained from the previous level, making the entire audio encoding. Finally, the dependencies between subsequent audio interviews for each patient are captured in the third level.

The final goal is to detect the onset of Alzheimer's disease for each individual using audio interview recordings. Suppose that each individual has a sequence of at most L follow-up interview visits v_i such that each audio interview length is L seconds. We segment each audio interview visit v_i into audio chunks as e_{ij}, which represents the j^{th} audio chunk in the i^{th} visit for this patient. Each chunk e_{ij} has the length of T, which means that it contains T one-second audio chunks such that c_{ijk} represents the k^th one-second audio chunk in the j^{th} audio chunk (e_{ij}) in the i^{th} audio interview visit.

Each level of our three level hierarchical deep audio model is illustrated in details as follows.

3.4.8.1 First Level

In the first level of our algorithm, We use VGGish to obtain a 128-dimensional feature embedding vector for every one-second audio chunk, c_{ijk}, via

$$x_{ijk} = W_{(VGGish)}c_{ijk} + b_{(VGGish)}, k \in [1, T], \tag{3.6}$$

Then we encode each audio chunk with length T using GRU-BRNN as h_{ijk} and its hidden representation u_{ijk} as follows:

$$h_{ijk} = [\overrightarrow{\text{GRU}}(x_{ijk}); \overleftarrow{\text{GRU}}(x_{ijk'})], k \in [1, T], k' \in [T, 1], \tag{3.7}$$

$$u_{ijk} = \tanh(W_w h_{ijk} + b_w), k \in [1, T]. \tag{3.8}$$

The attention weights are computed as:

$$\alpha_{ijk} = \text{Softmax}({u_{ijk}}^T u_c), k \in [1, T]. \tag{3.9}$$

The last step in this encoding level is to encode each sentence as a weighted sum of h_{ijk} with the attention scores,

$$e_{ij} = \Sigma_k \alpha_{ijk} h_{ijk}, k \in [1, T]. \tag{3.10}$$

Figure 3.6 illustrates the layers for the first level of our hierarchical algorithm.

3.4.8.2 Second Level

In the second level of this structure, we encode each visit representation using e_{ij} obtained from the first level applying GRU-BRNN to incorporate both future and past information within a transcript. Then, we calculate the attention scores for the components of this level α_{ij} with the context vector u_e as follows:

$$h_{ij} = [\overrightarrow{\text{GRU}}(e_{ij}); \overleftarrow{\text{GRU}}(e_{ij'})], j \in [1, L], j' \in [L, 1], \tag{3.11}$$

$$u_{ij} = \tanh(W_e h_{ij} + b_e), j \in [1, L], \tag{3.12}$$

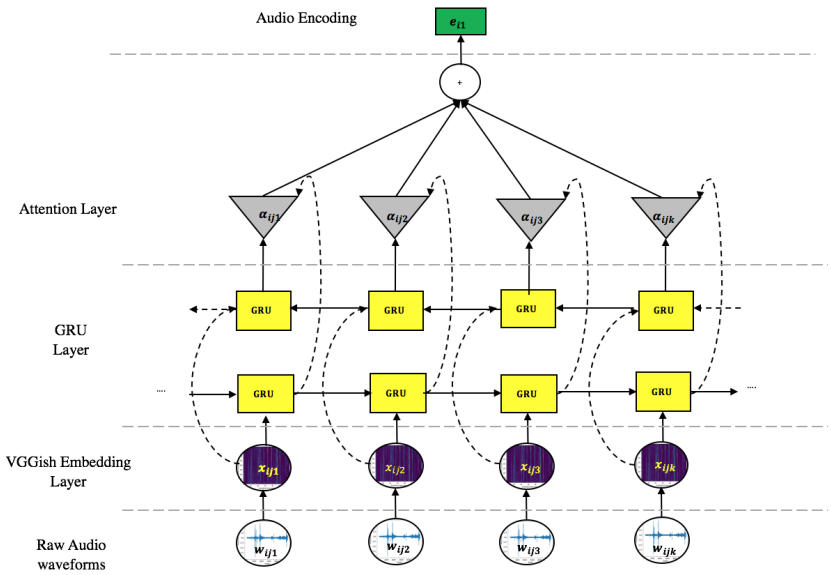


Figure 3.6: First level layers

$$\alpha_{ij} = \text{Softmax}({u_{ij}}^T u_e), j \in [1, L], \tag{3.13}$$

$$v_i = \Sigma_j \alpha_{ij} h_{ij}, j \in [1, L]. \tag{3.14}$$

Figure 3.7 represents the components of the second level.

3.4.8.3 Third Level

In the third level, we repeat the same process for the patient representation P, which is going to be computed by v_i obtained from the second level (Figure 3.8). This level determines the attention scores for each visit in Dementiabank longitudinal dataset as:

$$h_i = [\overrightarrow{\text{GRU}}(v_i); \overleftarrow{\text{GRU}}(v_{i'})], i \in [1, N], i' \in [N, 1], \tag{3.15}$$

$$u_i = \tanh(W_v h_i + b_v), i \in [1, N], \tag{3.16}$$

$$\alpha_i = \text{Softmax}({u_i}^T u_v), i \in [1, N]. \tag{3.17}$$

$$\tag{3.18}$$

We obtain the patient representation via:

$$P = \Sigma_i \alpha_i h_i . i \in [1, N]. \tag{3.19}$$

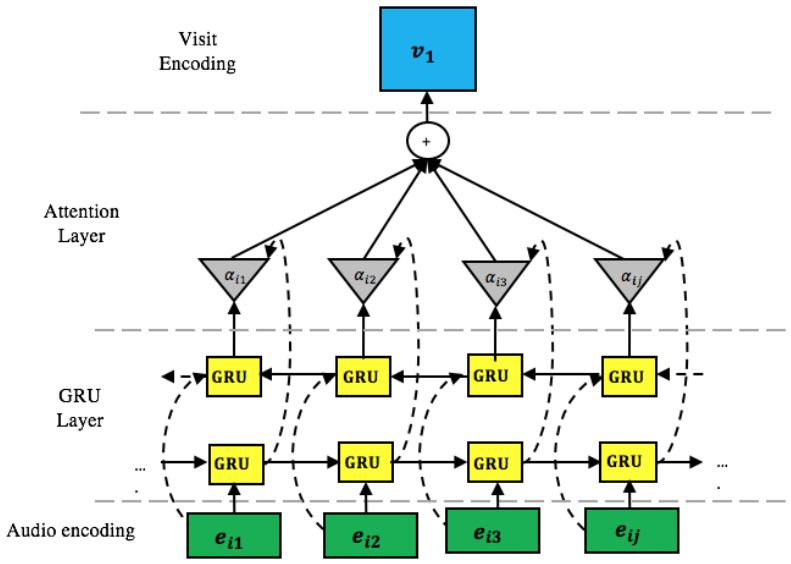


Figure 3.7: Second level layers

Finally, we use P obtained from the last level to build a binary classifier as:

$$d = \sigma(W_d P + b_d) \tag{3.20}$$

Figure 3.9 demonstrates how these three levels interacts with each other in a hierarchical structure.

3.5 Experiments

We developed an interpretable hierarchical deep audio model to detect the onset of Alzheimer's disease. Since our input data contains variable-length audio recordings, we need to transform data such that each audio recording has the same length. Thus, We truncated and padded each audio interview recordings to 120 seconds. We applied both truncating and padding zeroes at the end of one-dimensional audio tensors. Furthermore, we divided each audio input of length 120 seconds to 24 audio chunks of length five seconds.

Initially, we broke down each audio recording into one-second chunks and applied VGGish transfer learning to obtain 128-dimensional audio feature embedding representa-

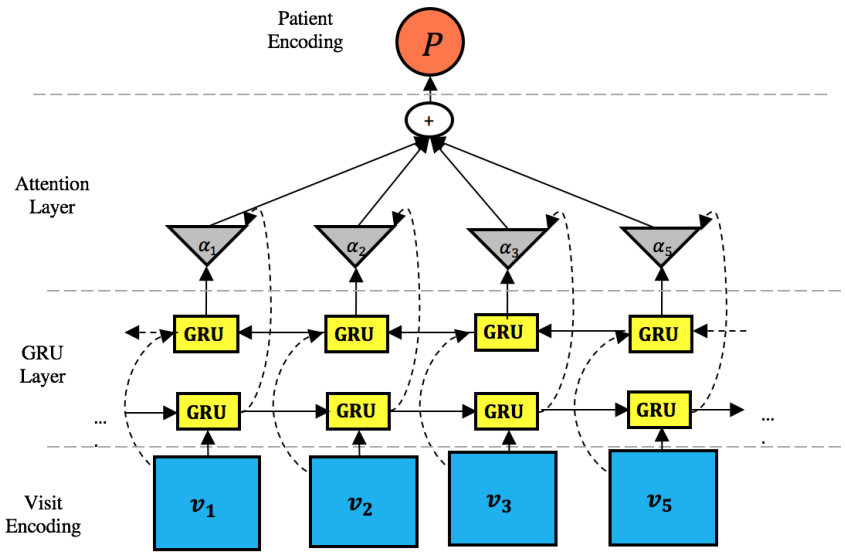


Figure 3.8: Third level layers

tions. Then, we encode five-second audio chunks and the whole visit (120 seconds) audio recording subsequently.

We set the GRU dimension to 50 for all three levels so that the bidirectional GRU will have 100 dimensions. The batch size is set to 10 patients with the same number of five visits per patient and 24 audio chunks of length 5 seconds per visit. The number of epochs is 20.

We applied the Adaptive Subgradient (Adagrad) optimizer to train the hierarchical model. Since we have the binary classification in the final level (AD or healthy subjects), We used 'categorical cross-entropy' as the loss function with the 'accuracy' as our performance metric.

3.6 Results

The Dementiabank dataset contains 99 healthy subjects and 169 subjects with probable Alzheimer's disease. Each subject has at most five annual follow-up visits. So, to evaluate our model, we conducted 11-fold stratified cross-validation. In each fold, we randomly split the input data to the training set and testing sets with a ratio of 9:1, respectively. The validation set is 10% of the training set. All training, validation, and testing sets are disjoint in each fold. The stratified cross-validation ensures the same ratio of healthy to AD subjects in each fold. Table 3.8 presents the results for 11-fold cross validation

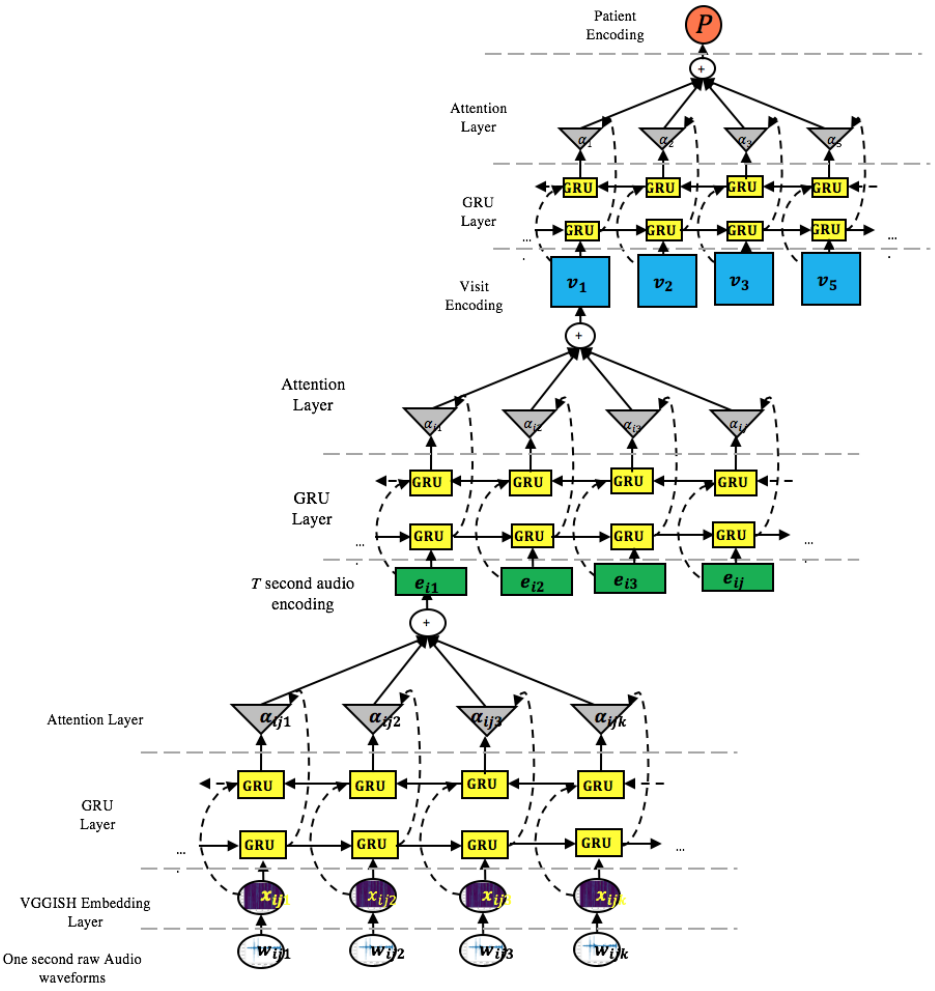


Figure 3.9: Hierarchical deep audio model

using our hierarchical deep audio model.

Our three-level hierarchical deep audio model achieves a mean accuracy of 90% ranging from 76% to 100% across 11-folds cross-validation to detect AD using patients' raw interview audio recordings, which is a new benchmark accuracy for audio classification task on the DementiaBank dataset. The mean precision, recall and F1 score across 11 folds are 93%, 92% and 92%, respectively.

To evaluate our hierarchical model on balanced input data, we augmented audio data by adding the random noise to the original audio recordings with the noise factor of 0.005. The audio augmentation technique was applied to both AD and healthy subjects to create a balanced dataset such that we could obtain 198 subjects for each AD and healthy category.

We applied 11-fold cross validation with the same experiment settings as the initial

Table 3.8: 11-fold cross validation results using hierarchical deep audio model

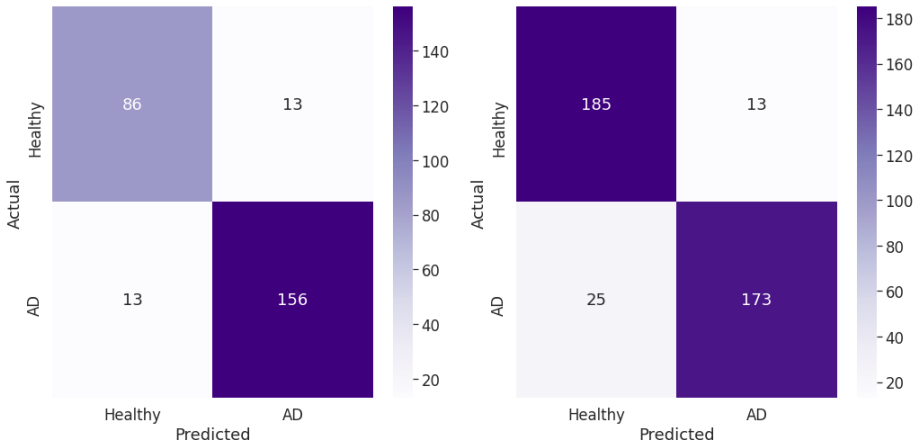
Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.76	0.78	0.88	0.82	0.80
2	0.80	0.87	0.81	0.84	0.89
3	0.87	1.00	0.81	0.90	0.99
4	0.92	0.89	1.00	0.94	0.92
5	1.00	1.00	1.00	1.00	1.00
6	0.91	0.88	1.00	0.94	0.90
7	0.91	0.88	1.00	0.94	0.98
8	1.00	1.00	1.00	1.00	1.00
9	0.95	1.00	0.93	0.97	0.98
10	0.91	0.93	0.93	0.93	0.93
11	0.88	1.00	0.80	0.89	0.96
Mean	**0.90**	**0.93**	**0.92**	**0.92**	**0.94**

unbalanced experiment (Section 3.5). Table 3.9 presents the results using augmented audio data. The mean accuracy, precision, recall and F1 score across 11 folds are 90%, 93%, 87% and 90%, respectively.

Table 3.9: 11-fold cross validation results using augmented balanced data

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	0.67	0.71	0.56	0.63	0.76
2	0.83	0.83	0.83	0.83	0.89
3	0.94	0.94	0.94	0.94	0.94
4	0.89	0.94	0.83	0.88	0.91
5	0.89	0.94	0.83	0.88	0.94
6	0.89	0.89	0.89	0.89	0.95
7	0.94	1.00	0.89	0.94	1.00
8	0.94	1.00	0.89	0.94	0.99
9	0.97	0.95	1.00	0.97	1.00
10	1.00	1.00	1.00	1.00	1.00
11	0.97	1.00	0.94	0.97	1.00
Mean	**0.90**	**0.93**	**0.87**	**0.90**	**0.94**

Figure 3.10 shows the confusion matrix for our hierarchical deep audio model for both unbalanced (a) and augmented balanced (b) input data. It is clear from Figure 3.10 (a) that only 13 probable AD patients are miss-classified as healthy (false-negative) and only 13 health subjects are miss-classified as probable AD (false-positive). Furthermore,



(a) Unbalanced data (b) Balanced data (Augmented)

Figure 3.10: Confusion matrix for classification of AD and Healthy subject using our hierarchical deep model

Figure 3.10 (b) shows that our model performs reasonably robust again when using balanced data. We have 25 probable AD patients that falsely predicted as healthy and 13 healthy samples that falsely predicted as AD.

We also evaluated the performance of our hierarchical deep audio model against three models, including non-hierarchical deep neural networks: *bidirectional GRU* and *attention based bidirectional GRU*; and non-deep learning method including Support Vector Machine (SVM). The first step of extracting feature embeddings using the VGGish transfer learning method is the same for all these three models. However, in order to evaluate the impact of hierarchical models, we fed these models with the full pool of patients' visits regardless of the visit time, which includes 309 audio recordings for AD subjects, and 243 audio recordings for healthy subjects. The mean accuracy for SVM was 60% and for bidirectional GRU and attention-based bidirectional GRU were 79% and 83%, respectively (Table 3.10). Not only is the prediction performance of non-hierarchical models lower than that of our novel hierarchical deep audio model, which was 90%, but also those models are black boxes and cannot provide any insights within the model.

Table 3.10: Testing accuracy results for different models

Model	1D-CNN-Raw Audio	SVM-VGGish	BGRU-VGGish	Attention-BGRU-VGGish	Hierarchcal Model
Accuracy	71%	60%	79%	83%	90%

To show the interpretability of our model, we presented five samples of the patient's visits (Figures 3.11, 3.12, 3.13,3.14, 3.15). In each sample, the whole audio waveform and its corresponding attention scores are depicted as plots. The plot of attention score shows how our model gives different attention scores to different parts of the interview audio recording. We listened to the audio recordings to extract those parts of dialogue with high attention scores. Each audio interview transcript is presented in each figure, and those parts with higher attention scores are highlighted with yellow and red colors. Red highlight signals those parts with the highest attention scores (peaks in the attention plot) and yellow highlights related to those interview parts with lower attention scores compared to the red highlight. We can see that all these highlighted parts are the signals of disfluencies and memory loss in patients' speech, and our model can automatically capture those parts very well. The implementation codes are available online at: `https://github.com/marynik66/AD-Audio`

3.7 Discussion

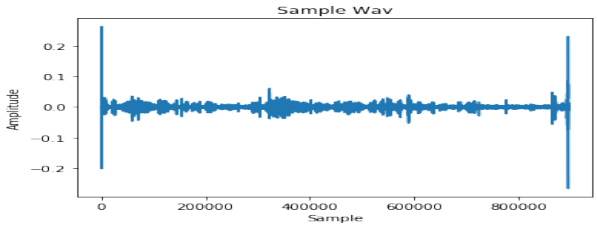
In this study, we developed an interpretable three-level hierarchical deep audio model to detect the onset of Alzheimer's using raw audio interviews of patients in their annual follow-up visits. Our hierarchical deep audio model achieved the mean accuracy of 90% across 11-fold cross-validation on Dementiabank audio dataset, which is a new benchmark performance compared to similar works ([114]).

We explored our model interpretability by providing interview samples where the parts of speech with high attention scores were highlighted. The highlighted speech of interview transcripts in figures 3.11, 3.12, 3.13, 3.14 and 3.15 show more of Alzheimer's disfluency codes (Table 2.1). For example, we can see that the red highlighted sentences in figure 3.14 is "**&plat <off the> [/] &uh off the &uh &sh &pl chair**" which has the most pauses including "uh","sh" as well as phrase revision "off the". This approves that our model can capture those disfluencies very accurately.

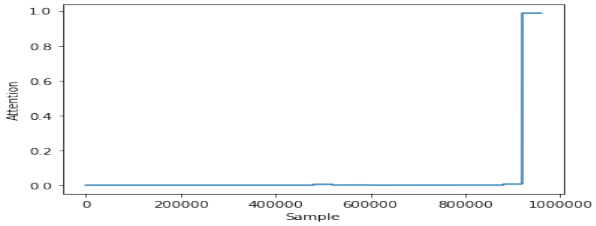
Our model also compensates the small size of the Dementiabank dataset by the VGGish transfer learning method, which is trained on two million 10-second YouTube-8M. Transfer learning makes our model more applicable in real healthcare problems where collecting datasets is a challenging step towards developing automatic learning models.

To conclude, we have developed an interpretable deep audio model to predict Alzheimer's disease's onset. The results of our experiments showed an outstanding

accuracy performance (90%). It can also tell clinicians the exact parts of patients' interview audio that are affected more by memory loss issues. Our model's transparency is crucial for clinicians to trust and apply automatic predictive tools in practice.



(a) Audio wave

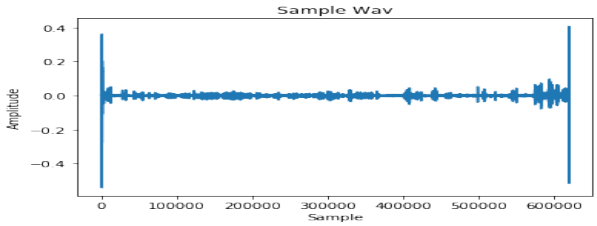


(b) Attention score

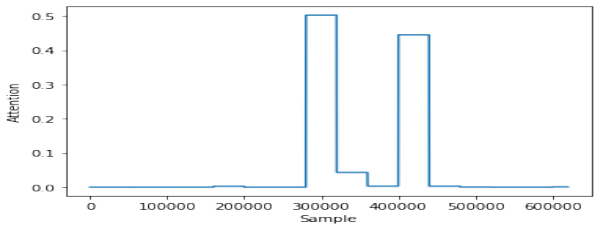
*PAR: &uh everything that's going to happen, huh ? [+ exc]
*INV: mhm .
*PAR: looking out the window <maybe it> [//] I don't know if it's &uh the breeze blowin(g) the curtains because <the curtains look a little wee bit> [//] &th <over the> [/] <over the window blind> [//] over the window they look a little wee bit puffy .
*PAR: so maybe there's a slight breeze coming <in the> [/] in the kitchen window .
*PAR: mama's drying the dishes and she forgettin(g) herself and the water is overflowing from the sink down onto the floor splashing into her shoes .
*PAR: alright there are two cups and a plate on the counter to the right of the sink .
PAR: and &uh <the boy> [//] well I did tell you the boy is gettin(g) ready to fall off of the chair [: stool] [s:r] .
*PAR: his sister is reaching up for a cookie, gettin(g) it from the cookie jar .
*PAR: so what else ? [+ exc]
*PAR: the [/] the door is open of_course to get into the cupboard .

(c) Interview transcript

Figure 3.11: First visit (first year)



(a) Audio wave



(b) Attention Score

*PAR: and the boy looks like he's gonna fall down and hurt himself or fall against his mother . ▶

*PAR: and the girl is whispering “don't make too much noise” to him . ▶

*PAR: she's [//] &let or else she's laughin(g) at him . ▶

*PAR: they got the cookies . ▶

*PAR: alright now though the window, let's see . ▶

*PAR: there's a nice look outside, real nice . ▶

*PAR: I told you the water was running over and splashing onto the floor . ▶

*INV: mhm . ▶

*PAR: +< and the mother doesn't seem too too affected by it . ▶

*PAR: she's dryin(g) a dish or wiping it . ▶

*PAR: let's see . [+ exc] ▶

*PAR: I guess the girl is laughing at her brother because he's going to fall . ▶

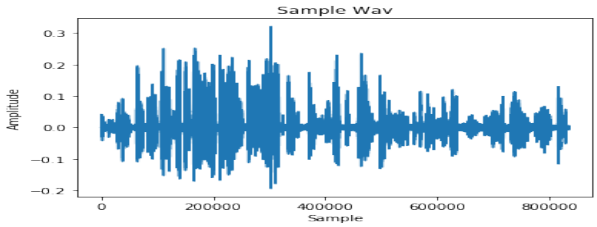
*INV: okay . ▶

*PAR: looks like a nice house . ▶

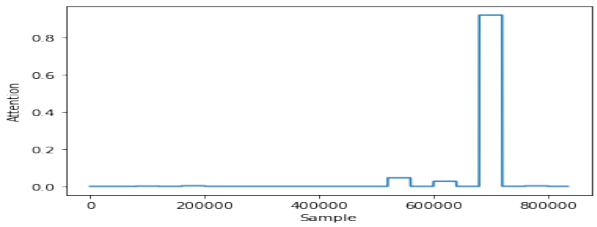
3

(c) Interview transcript

Figure 3.12: Second visit (second year)



(a) Audio wave

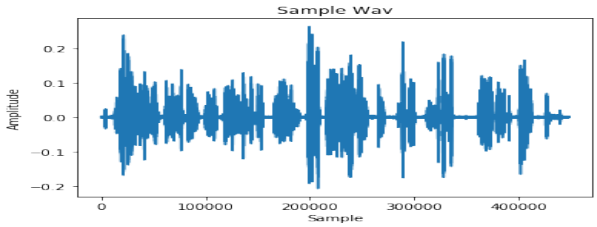


(b) Attention Score

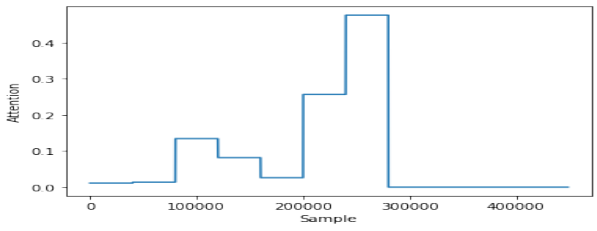
*PAR: she's lookin(g) out the window . ▶
*PAR: and in looking out the window why she's lettin(g) her sink run over and the water's splashin(g) on the floor all over her . ▶
*PAR: and &uh what else ? [+ exc] ▶
*PAR: &=noises:thinking the outside looks alright . ▶
*PAR: they're doin(g) more things on the outside . ▶
*PAR: there's some more there . [+ es] ▶
*PAR: and then <she's dryin(g)> [/] she's drying a dish . ▶
*PAR: and she has two cups and a plate on the table . ▶
*PAR: they've already eaten xxx . ▶
*PAR: <she could> [//] well I did say she was lettin(g) the &uh water run over the sink down onto the floor splashing onto her feet . ▶
*PAR: oh boy . [+ exc] ▶
*PAR: then there's an angle here that is incomplete <of the> [/] &st of the corner where the wall comes together . ▶
*PAR: &hm let's see what else should there be . [+ exc] ▶
*PAR: oh: let's see . [+ exc] ▶
*PAR: the [//] &ta there's a plate two [/] &sau two cups . ▶
*PAR: oh &sh I did say she's lettin(g) the water run over the sink didn't

(c) Interview transcript

Figure 3.13: Third visit (third year)



(a) Audio wave



(b) Attention Score

*PAR: <well the kid> [//] the girl's laughin(g) at her brother because he went into the cookie jar and he's fallin(g) over the cookie jar .

*PAR: and mother's [//] &d the mother was at the sink .

PAR: and the sink's [: water's] [s:r] splashin(g) [/] &s splashin(g) over the sink .

*PAR: and she sort_o(f) a little bit bitchy .

*PAR: and the water's goin(g) floodin(g) over the sink .

*PAR: and there's a saucer, there's a plate, there's a couple of dishes .

*PAR: well she's [//] and the mother's lookin(g) out the window .

*PAR: she don't know what the hell to think of it .

PAR: a girl laughin(g) at her brother who is taking cookies out_of the cookie jar and he's ready to fall <off the damn> [//] &plat <off the> [/] &uh off the &uh &sh &pl chair [: stool] [s:r] he's on .

PAR: the chair's [: stool's] [s:r] crooked .

*PAR: what the hell else ? [+ exc]

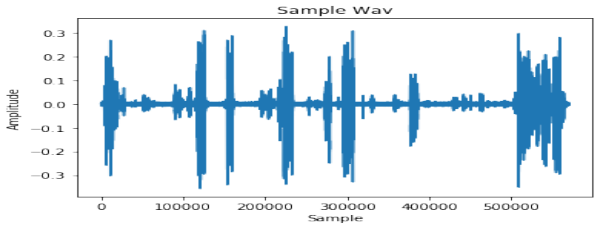
*PAR: and then there's a plate, saucer and two cups .

*PAR: she's looking out the window .

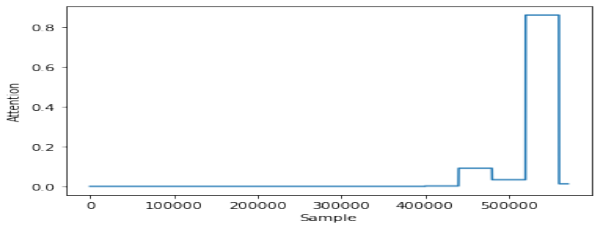
*PAR: the window's open .

(c) Interview transcript

Figure 3.14: Fourth visit (fourth year)



(a) Audio wave



(b) Attention Score

*INV: can you tell me what's going on in the picture ?
*PAR: kids are trying to get a &s [x 6] +...
*INV: can you tell me them then ?
*INV: tell me the mistakes .
*PAR: it's full o(f) mistakes .
*PAR: it's full of mistakes .
*INV: (o)kay .
*INV: what's going on right here ?
PAR: he's changin(g) [: taking] [s:ur-ret] [//] taking cookie jar . [+ gram]
*INV: uhhuh anything else ?
*PAR: that's all . [+ exc]
*INV: how_(a)bout what's going on over here ?
*PAR: the mother's just drying the dishes .
*INV: anything else going on ?
*PAR: &n [x 6] &s &s &n &s: xxx from the from xxx .
*PAR: this is &uh +...
*PAR: &=laughs .
*INV: okay .

(c) Interview transcript

Figure 3.15: Fifth visit (fifth year)

Chapter 4
Detection of COVID-19 from Chest Radiography Images

4.1 Introduction

The outbreak of the novel coronavirus known as Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) started in Wuhan, China, in December 2019 and rapidly spread worldwide. As the number of people with Coronavirus Disease 2019 (COVID-19) escalates in the United States and around the world, reducing the number of transmissions from infected individuals to the general population and healthcare workers becomes increasingly important and challenging.

Although about eight in ten people who contract the SARS-CoV-2 virus remain asymptomatic or develop only mild to moderate symptoms [115–118], others may develop life-threatening conditions, such as dyspnea, pneumonia, or severe acute respiratory syndrome, which require hospital or ICU care with supplemental oxygen or mechanical ventilation [119, 120]. The rapid spread of COVID-19 is, in part, due to the lack of sufficient testing and isolation of positive cases, which subsequently leads to community transmission from undiagnosed cases. Even where testing is available, the long wait for the test results, which is about 12-13 days in some cases due to the short supplies of laboratories for RT-PCR testing and the surge in the number of cases, makes the tests' utility marginal [121]. This rapid spread may result in overwhelming and collapsing healthcare systems, even in developed countries, due to the surge in demand for hospital and ICU care [122].

To control the transmissions and flatten the curve, we must use a widely-available, fast, and accurate COVID-19 detection method, and immediately isolate diagnosed cases until they are no longer infectious. The current gold standard screening method for COVID-19

is the direct detection of SARS-CoV-2 RNA by reverse transcription-polymerase chain reaction (RT-PCR) test [123]. Several RT-PCR assays are used in the U.S. and around the world. Each has different performance characteristics and turnaround times (ranging from minutes to several hours) and requires different specimen types [124]. The sensitivity of RT-PCR testing is widely variable. Depending on the assay, the type and quality of the specimen obtained, the stage of the disease, and the duration of infection, the sensitivity can vary between 32% to 73% [125–127]. Therefore, there is an immediate need for accessible, rapid, and accurate testing tools to help combat the spread of the SARS-CoV-2 virus.

Medical imaging modalities such as Chest X-ray (CXR) and Computed Tomography (CT) can be used as an alternative to RT-PCR testing to detect characteristic symptoms of COVID-19 in patients' chest images [128, 129]. Detecting COVID-19 from chest radiography images has shown promising results and higher sensitivity compared to RT-PCR testing [130]. Moreover, CT images of patients with COVID-19 may show abnormalities before the patient develops symptoms and before the detection of viral RNA from upper respiratory specimens [131, 132]. Even in asymptomatic cases with negative nucleic acid testing, CT images have proven definite features in screening and detecting COVID-19 patients [133].

Although radiologists can successfully distinguish COVID-19 from viral pneumonia using chest CT images with moderate to high accuracy [134], working with a high workload of images in emergencies like the COVID-19 pandemic leads to more fatigue and can affect human diagnostic performance [135].

Artificial intelligence (AI) methods have recently proved a huge impact in medical data analysis by automatically extracting rich features from multi-modal medical data [136]. In this context, artificial intelligence (AI) methods can be leveraged to automatically analyze medical images for subtle signs of SARS-CoV-2 infection and detect COVID-19 rapidly and accurately. However, the success of AI models highly relies on the availability of input data. Although many patients are infected by COVID-19, publicly available COVID-19 datasets with chest images are difficult to obtain due to patients' privacy issues. Therefore, we first describe the publicly available COVID-19 image datasets. Then, we review on the deep learning-based AI models developed to detect COVID-19 infection using those available datasets.

4.2 Dataset Description

To detect COVID-19 using AI models, we used two modalities of chest radiography images, including X-ray images and CT slices. These images are mainly extracted from five major data sources as follows:

- **Novel Corona Virus 2019 Dataset (S1)**: Joseph Paul Cohen and Paul Morrison and Lan Dao [137] have built an open public dataset which comprised of chest X-ray and CT images of patients diagnosed with COVID-19, Middle East respiratory syndrome (MERS), Severe acute respiratory syndrome (SARS) and Acute respiratory distress syndrome (ARDS). This dataset is collected indirectly from publications, hospitals and clinicians as well as directly from other public sub-datasets including:

  - https://radiopaedia.org/ (license CC BY-NC-SA)
  - https://www.sirm.org/category/senza-categoria/COVID-19/
  - https://www.eurorad.org/ (license CC BY-NC-SA)
  - https://coronacases.org/ (preferred for CT scans, license Apache 2.0)

  The details of dataset annotation are described in their Github repository: `https://github.com/ieee8023/COVID-chestxray-dataset` This dataset is This dataset is continuously updated and as of 1st August 2020, there were 475 chest X-rays and 59 CT images of confirmed COVID-19 cases. In this dataset, metadata is not complete for all patients. The average age is 56 ± 2, and the number of male patients was approximately twice more than female patients (326 vs. 174) for those images with demographic information.

- **Chest X-ray Images (pneumonia) Dataset (S2)**: Chest X-ray images from other infections (viral/bacterial) have been collected in Kaggle [138]. This dataset contains 3949 images from patients with other viral/bacterial infections as well as 1583 images from individuals with no pneumonia (i.e., normal). This dataset is collected from pediatric patients ages one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou.

- **COVID-19 Radiography Dataset (S3)**: Chowdhury et al. [139] developed the dataset of COVID-19 X-ray images from different sources including the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 dataset [140],

Novel Corona Virus 2019 Dataset [137] and images from 43 different publications. The distributions of images for each category are as follows: 219 COVID-19 positive images, 1341 normal images, and 1345 viral pneumonia images.

- **SARS-CoV-2 CT-scan Dataset (S4)**: Soares et al. [141] collected a public SARS-CoV-2 CT scan dataset from real patients in hospitals from Sao Paulo, Brazil. This dataset contains 1252 CT scans for COVID-19 and 1230 CT scans for patients not infected by COVID-19 (non-COVID19).

- **COVID-CT Dataset (S5)**: Zhao et al. [142] developed a COVID-19 dataset based on the chest CT images of patients. They collected CT images from COVID19-related papers (i.e, medRxiv, JAMA, Lancet, etc) from January 19th to March 25th. This dataset comprised of 349 COVID-19 patients and 397 not infected patients (non-COVID19).

4.3 Related works

Motivated by the need for a fast testing tool to detect COVID-19, several automatic detection models using deep learning models with radiography images as their input have been developed recently. Narin et al. [143] proposed three different convolutional neural network models to detect COVID-19 patients from healthy individuals (binary classification). They used chest X-ray images of 50 COVID-19 patients from S1 [137] and 50 normal chest X-ray images from S2 [144]. They obtained the highest accuracy of 98% compared to their other two models.

Ozturk et al. [145] developed a model called DarkCOVIDNet for COVID-19 detection in a binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. Pneumonia). They obtained X-ray images of COVID-19 patients from S1 [137] and images of Normal pneumonia patients from [146]. They achieved an accuracy performance of 98.08% for binary classification and 87.02% for multi-class classification.

Wang et al. [147] proposed a model called COVID-Net, which is tailored to detect COVID-19 using S1 [137] which is publicly available. They obtained an overall accuracy of 92.6% in a multi-class classification setting with 87.1 %, 90.0 %, and 97.0% sensitivity for COVID-19, NonCOVID-19, and normal subjects, respectively.

Yang et al. [148] built a new CT dataset, "COVID-CT-Dataset" (S5), which is publicly available. Using this dataset, they developed an AI model using multi-task learning and

contrastive self-supervised learning to detect COVID-19 and achieved the best accuracy performance of 89.1%.

Chowdhury et al. [139] collected a new X-ray dataset that is mainly created from public open-sourced datasets [139]. This dataset contains 423 COVID-19, 1485 viral pneumonia, and 1579 normal chest X-ray images. They also implemented different transfer learning methods and obtained the best classification accuracy of 97.94% using DenseNet201.

Soares et al. [141] built a large CT dataset called "SARS-CoV-2 CT-scan dataset" collected from different hospitals in Sao Paulo, Brazil. This dataset is made of 1252 CT scans belongs to COVID-19 patients, and 1230 CT scans belong to patients not identified with COVID-19 infection. Using this dataset, they developed the classification model of xDNN, which achieved an accuracy of 97.38% in the binary classification of COVID-19 vs. Non-COVID19 groups.

Tabik et al. [149], built a new database called COVIDGR-1.0, which contains all images from patients with all levels of severity, from Normal with positive RT-PCR, Mild, Moderate to Severe. They propose a COVID Smart Data based Network (COVID-SDNet) approach, which achieves an accuracy of 97.37% ± 1.86%, 88.14%±2.02%, 66.5%±8.04% in severe, moderate and mild COVID severity levels. This dataset can help to detect different severity levels of COVID-19 using Chest X-ray (CXR) images.

Karim et al. [150] proposed an explainable deep learning-based model called Deep-COVIDExplainer using CXR images. They obtained chest X-ray images from S1 [137], Ozturk et al. [145] and Wang et al. [147]. They obtained promising results in detecting COVID-19 with positive predictive value (PPV) of 91.6%, 92.45%, 96.12%, precision, recall, and F1 score of 94.6%, 94.3%, and 94.6% for normal, pneumonia, and COVID-19 patients, respectively. More recently, numerous studies has focused on developing deep learning-based models to detect COVID-19 using chest X-ray images [128, 151–154]. Li et al. [154] proposed a deep learning-based model called COVNet to detect COVID-19 using chest CT. COVNet can successfully detect COVID-19 from community-acquired pneumonia and Non-pneumonia classes. They used private CT images dataset, which was collected from six hospitals between August 2016 and February 2020. The model achieved a sensitivity of 89.8% and a specificity of 95.8% in detecting COVID-19.

Butt et al. [155] employed two CNN three-dimensional classification models on a private dataset. They obtained an accuracy of 86.7% to detect COVID-19 from normal patients. Apostolopoulos et al. [156] applied different transfer learning methods using public available X-ray images including S1 [137]. They achieved the best accuracy,

sensitivity, and specificity of 96.78%, 98.66%, and 96.46%, respectively.

Zhang et al. [157] propose a model called the confidence-aware anomaly detection (CAAD), which detects anomalies that exist on chest images. They obtained an AUC of 83.61% and a sensitivity of 71.70% using the X-COVID dataset, which comprised 106 confirmed COVID-19 cases and 107 normal controls. Wang et al. [158] applied AI models to detect COVID-19 using 453 CT images of COVID-19 patients from a private dataset. They achieved an accuracy of 73.1% with a specificity of 67% and a sensitivity of 74%.

Zheng et al. [159] proposed a weakly-supervised deep learning algorithm for lung segmentation using CT chest images collected from a single hospital of 313 confirmed COVID-19 patients and 229 patients without COVID-19. They proposed an end-to-end deep learning model which is called DeCoVNet. This model obtained the best accuracy of 90.8% without annotation of of pulmonary lesions in CT volumes.

Song et al. [160] proposed a deep learning-based model that can identify the COVID-19 infected patients and bacteria pneumonia-infected patients with a sensitivity of 0.96 using chest CT images collected from hospitals in China. Their model is also able to localize the main lesion features of CT images. Ardakani et al. [161] used ten well-known convolutional neural networks to detect COVID-19 from the non-COVID19 group. The ResNet-101 model obtained a sensitivity of 100%, a specificity of 99.02%, and an accuracy of 99.51%. They retrospectively collected chest CT images of patients from September 2019 to December 2019.

Minaee et al. [162] prepared a dataset using publicly available X-ray datasets which contains 184 COVID-19 samples and 5000 samples without COVID-19. They trained several convolutional neural networks, including ResNet18, ResNet50, SqueezeNet, and DenseNet-121 and achieved a sensitivity rate of 98% and a specificity rate of around 90%. Kumar et al. [163] proposed a prediction method that extracts deep features from X-ray images using ResNet152. Their dataset includes 62 images for COVID-19 patients, 1341 images for Normal patients and 1345 images for other Pneumonia patients. They used Random Forest and XGBoost for final classification of COVID-19, Normal and other Pneumonia using extracted deep features and achieved an accuracy of 97%.

Bullock et al. [164] provide a comprehensive review of recent studies based on Artificial Intelligence methods on different aspects of the COVID-19 crisis, including molecular, clinical, and societal applications. They also present a review on datasets, tools, and resources required to develop AI models to detect COVID-19.

In this study, we develop a new deep learning model to automatically detect COVID-19 using chest X-ray and CT scan images. We refer to this model as Artificial Intelligence

for Detection of COVID-19 (AIDCOV). AIDCOV employs a novel hierarchical attention structure, which can tell clinicians the specific locations of the lung affected by the SARS-CoV-2 infection rapidly and with high sensitivity and specificity.

4.4 Methods

AIDCOV includes a novel two-level hierarchical attention structure for classification of chest radiography images into one of the three classes: COVID-19 viral infection, other viral/bacterial infection (i.e., non-COVID19 infection), or normal (i.e., no infection). This hierarchical structure enables the model to capture the dependency of features extracted from chest images via a pre-trained network (e.g., VGG-16) in both horizontal and vertical directions and helps improve model performance. The attention mechanism makes the black-box deep neural network model interpretable such that the model can designate the specific locations of patients' lungs that manifest subtle signs of infection. AIDCOV is an end-to-end deep neural network model, which does not require any feature engineering.

4.4.1 Transfer Learning

Deep neural network models often include hundreds of thousands of hyperparameters, and thus they need to be trained on very large datasets. Transfer learning is a technique that allows training deep neural network models on small datasets by taking a pre-trained deep neural network model and repurposing it for a different task. We leverage the powerful idea of transfer learning by employing VGG-16 [100], a pre-trained convolutional neural network that is trained on a dataset of more than 15 million images [5]. VGG-16 has shown promising performance for medical image analysis [165–167]. VGG-16 has 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. We removed the 3 fully connected layers and replaced them with our novel hierarchical attention structure. While the early layers of VGG-16 learn low-level features of the image, our hierarchical attention model learns subtle signs of COVID-19 and other viral/bacterial infections and determines the final classification. Section 4.4.2 illustrates this process in details.

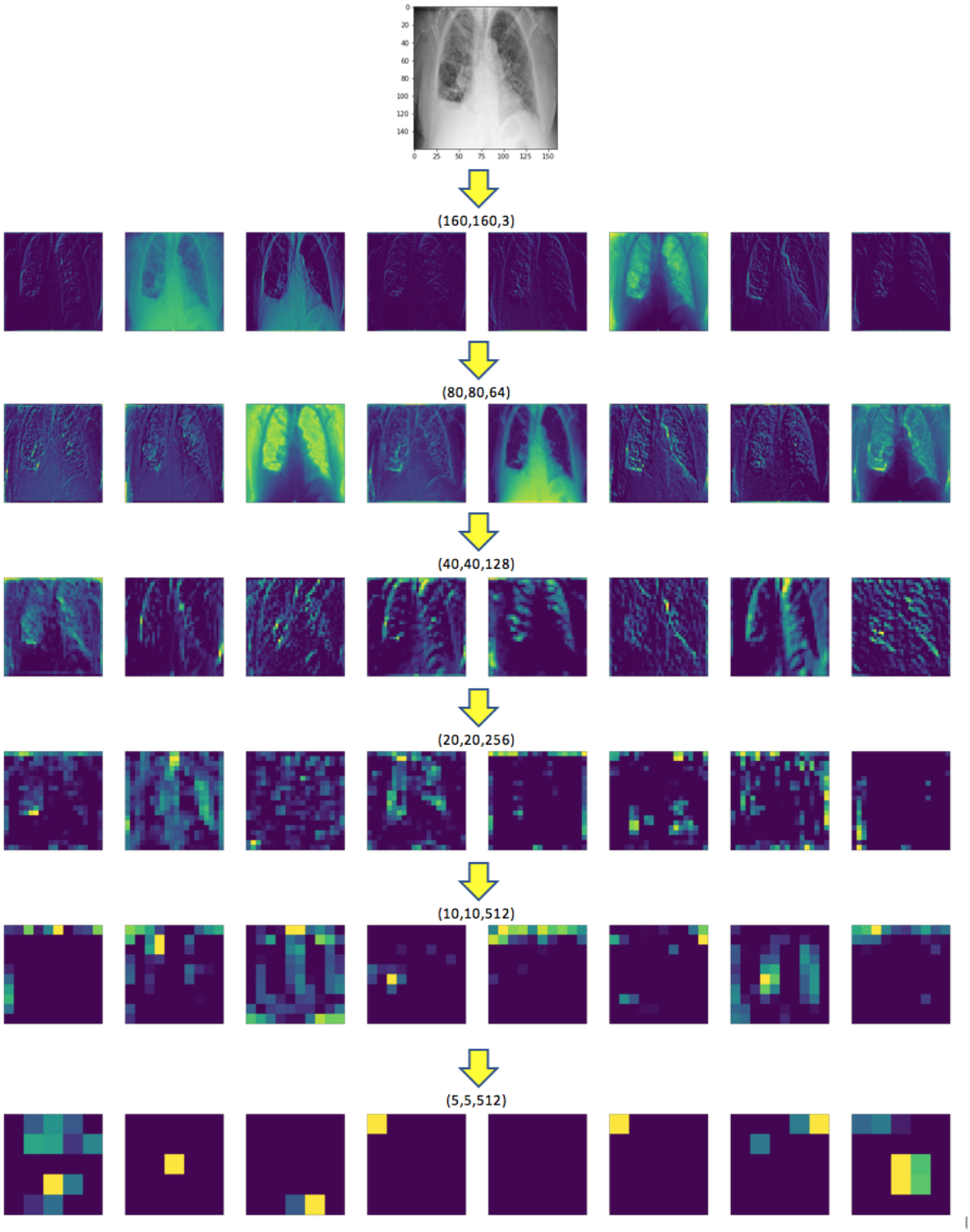


Figure 4.1: Low-dimensional feature extraction using the pre-trained VGG-16 model.

4.4.2 Two-level Hierarchical Deep Neural Network Model for Image Classification

In this section, we describe our novel hierarchical attention structure for image classification, which considers the dependencies of feature components in both horizontal (width) and vertical (height) directions.

Step 1: Resizing the image. First, we need to resize input chest radiography images to the format that is compatible with the pre-trained VGG-16 model. We resized all the input images to size $160 \times 160 \times 3$.

Step 2: Feature extraction. We use VGG-16 to obtain a low-dimensional feature representation vector for each resized image (Figure 4.1).

In general, if the input image is of size (A, B, 3), the output of the VGG-16 model will be a tensor of size (A/32,B/32, 512). In our case, since the input image is of size (160,160,3), the output of the model is of size $(5, 5, 512)$. We show this output by X, which is obtained as follows:

$$X = W_{VGG}C + b_{VGG}, \tag{4.1}$$

where C is the resized radiography image, and W_{VGG} and b_{VGG} are the trained weight and bias matrices obtained from the pre-trained VGG-16 model.

As mentioned above, X is the output of size $(5, 5, 512)$. We refer to each $(1, 1, 512)$ block of X as x_{ij}, where $i \in [1, 5]$ and $j \in [1, 5]$ (Figure 4.2).

Step 3: Horizontal feature encoding. Next, for each level of $i \in [1, 5]$, we encode the output block along the x axis (i.e., horizontally). To do so, we apply a GRU-BRNN on $x_{i1}, ..., x_{i5}$ for each level of i to incorporate horizontal dependencies (in both forward and backward directions) within each row of an image. Thus, we have:

$$h_{ij} = [\overrightarrow{\text{GRU}}(x_{ij}); \overleftarrow{\text{GRU}}(x_{ij'})], j \in [1, 5], j' \in [5, 1], \tag{4.2}$$

$$u_{ij} = \tanh(W_i h_{ij} + b_i), j \in [1, 5], \tag{4.3}$$

where h_{ij} is the output of the GRU-BRNN and u_{ij} is its hidden representation. W_i and

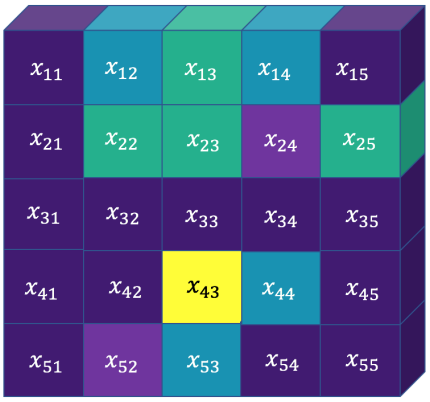


Figure 4.2: The output block of size $5 \times 5 \times 512$.

b_i are the weight matix and bias vector for each row of the x_{ij} block learned through training. Since we aim to determine the contribution of each x_{ij} block within each row (horizontal level) to the overall prediction, we applied an attention layer on top of the hidden representations u_{ij} to obtain the attention scores α_{ij} by learning the row context vector u_i,

$$\alpha_{ij} = \text{Softmax}({u_{ij}}^T u_i), j \in [1, 5]. \tag{4.4}$$

$$\tag{4.5}$$

Finally, we encode each $\hat{y}_i$ block as a weighted sum of h_{ij} and the attention scores α_{ij} (Figure 4.3),

$$\hat{y}_i = \Sigma_j \alpha_{ij} h_{ij}. \tag{4.6}$$

$$\tag{4.7}$$

Step 4: Vertical feature encoding. In this step, we encode the representations $\hat{y}_i$ computed from the previous step along the y axis (i.e., vertically). We further determine the attention scores α_i for each $\hat{y}_i$ block as follows.

$$h_i = [\overrightarrow{\text{GRU}}(\hat{y}_i); \overleftarrow{\text{GRU}}(\hat{y}_{i'})], i \in [1, 5], i' \in [5, 1], \tag{4.8}$$

$$u_i = \tanh(W h_i + b), \tag{4.9}$$

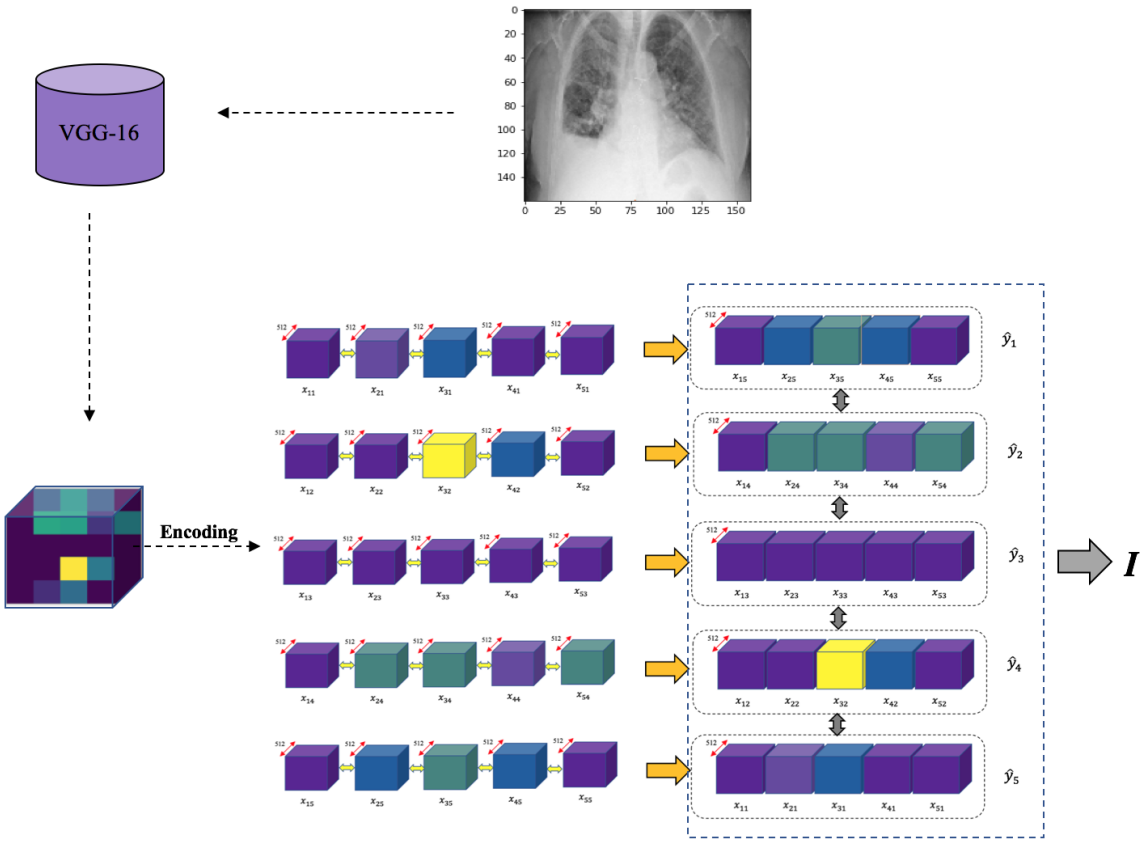


Figure 4.3: Encoding feature outputs in both x and y axis

$$\alpha_i = \text{Softmax}({u_i}^T u), \tag{4.10}$$

$$I = \Sigma_i \alpha_i h_i, \tag{4.11}$$

where h_i capture the dependencies of $\hat{y}_i$ blocks using GRU-BRNN and u_i is its hidden representation obtained through training of W and b parameters. The attention weights α_i for $\hat{y}_i$ are computed using u_i and the trained context vector of u. The image encoding is the weighted sum of h_i encodings and attention scores α_i. Finally, we use the image encoding I to build a multi-class classifier as follows:

$$p = \sigma(W_I I + b_I) \tag{4.12}$$

Figure 4.4 shows the hierarchical encoding network of Steps 3 and 4 described above.

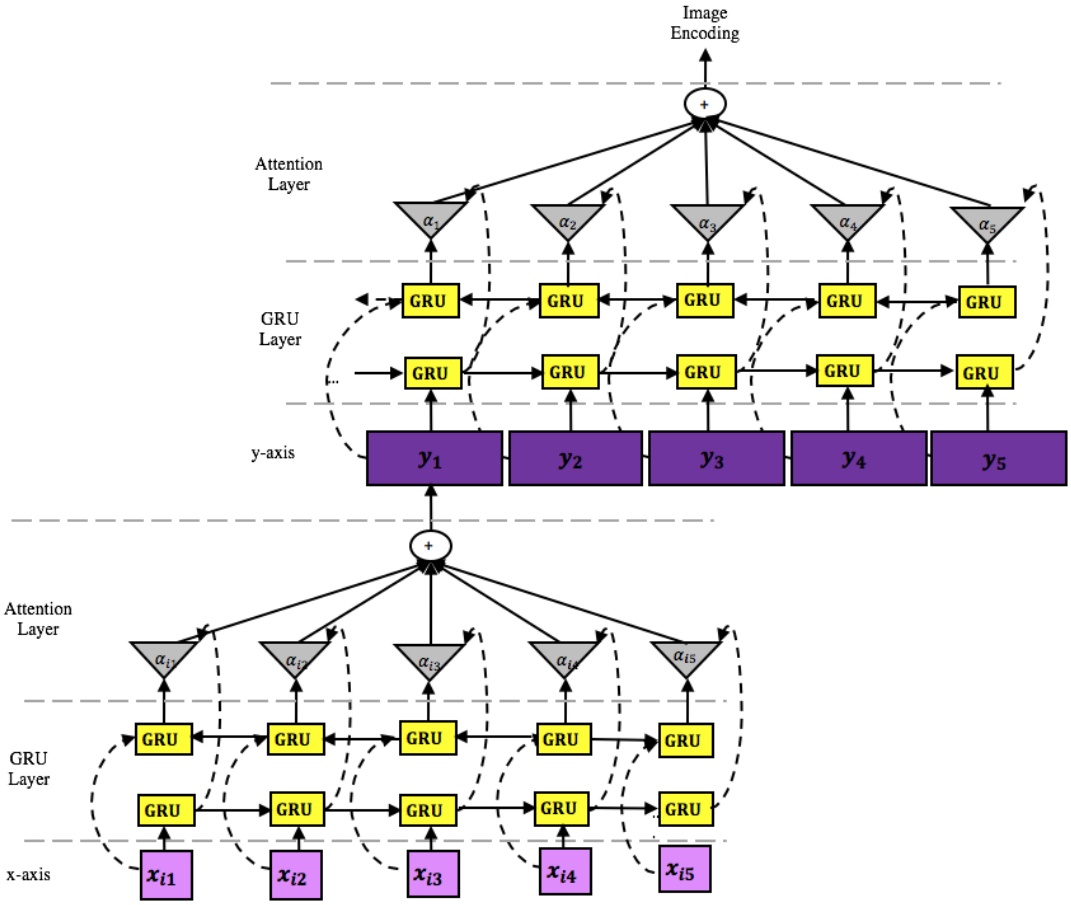


Figure 4.4: The hierarchical attention structure for image encoding.

4.4.3 Datasets

In our study, we leverage datasets from these five major data sources to build five different datasets (Table 4.1). Dataset-1 is the public dataset built by combining Novel Corona Virus 2019 Dataset [137] and Chest X-ray Images (pneumonia) dataset [138]. This dataset contains chest X-ray images of 475 COVID-19 patients as reported by Cohen et al. [137]. The other two categories of this dataset, including 3949 other infections (non-COVID19) and 1583 normal images, comes from Moony et al. [138]. Dataset-2 developed by Chowdhury et al. [139] and includes chest X-ray images from 219 COVID-19, 1341 normal and 1345 non-COVID19 images. Dataset-2 comes from SARS-CoV-2 CT scan dataset [141] containing 1252 COVID-19 CT images and 1229 non-COVID19 CT images.

Dataset-3 is based on CT images collected by Soares et al. [141]. This dataset contains

1252 and 1229 CT images for COVID-19 and non-COVID19 patients, respectively.

Dataset-4 originates from the COVID-CT dataset [142], which includes 349 CT images of COVID-19 patients and 397 CT images of not infected patients. Dataset-5 is built using 59 CT images from Novel Corona Virus 2019 Dataset [137] and 1252 CT images from SARS-CoV-2 CT scan dataset [141]. Table 4.1 illustrates the components of our five datasets. This study was exempt from institutional review board (IRB) review since it used publicly available, de-identified data.

Table 4.1: Publicly available datasets

Dataset	Sources	Type	Task Type	Categories	No.Images
Dataset-1	[137], [138]	X-ray	3-class	COVID-19	475
				Normal	1583
				Non-COVID19	3949
Dataset-2	[139]	X-ray	3-class	COVID-19	219
				Normal	1341
				Non-COVID19	1345
Dataset-3	[141]	CT	2-class	COVID-19	1252
				Non-COVID-19	1229
Dataset-4	[148]	CT	2-class	COVID-19	349
				Non-COVID-19	397
Dataset-5	[137], [141]	CT	2-class	COVID-19	1311
				Non-COVID-19	1229

4.4.4 Data Augmentation

Due to the limited number of COVID-19 images in our datasets, different image augmentation methods have been applied to increase COVID-19 images. Shorten et al. [168] provide a comprehensive review of image augmentation techniques. In this work, we applied a random rotation operation with the rotation range of 50 degrees. Also, we randomly flipped both vertically and horizontally. Horizontal and vertical shift augmentation was conducted with a range of 0.2. Random zoom augmentation has also been applied with a range value of 0.1. Wang and Perez [169] proved the effectiveness of data augmentation methods using a small subset of the ImageNet [101] dataset.

4.4.5 Implementation Details

We set the GRU dimension to 50 for both horizontal (Step 3) and vertical (Step 4) encoding levels. So, the bidirectional GRU has 100 dimensions. We used a mini-batch size of 20 images and trained on 20 epochs.

We used 'categorical cross-entropy' as our loss function to classify chest radiography images into one of the three classes (COVID-19, other infection, or normal). We employed the Adaptive Subgradient (Adagrad) as the optimizer.

To evaluate the model's performance in each experiment, we conducted 10-fold stratified cross-validation such that in each fold, the dataset is randomly divided into a training set and a testing set with the ratio of 9:1, respectively. The validation dataset is set as 10% of the training dataset. In each fold, the training, validation, and testing datasets are completely disjoint sets. The model is trained on the training set, validated on the validation set, and finally tested on an independent test set.

4.4.6 Experiments

Two different modalities of the dataset, including X-ray and CT images, were studied in this work. Thus, we split the experiments into two main categories. The first set of experiments was conducted using datasets that mainly included X-ray images. These experiments are multi-class such that the goal is to distinguish between three classes of patients including COVID-19, non-COVID19, and Normal. The second set of experiments was conducted on CT datasets and employ binary classifiers that try to distinguish COVID-19 CT images from non-COVID-19 CT images. Tables 4.2 and 4.3 presents both set of the experiments designed in this study. The details of these experiments are as follows:

- **Experiment1-1**: This experiment is designed on X-ray images using Dataset-1, which contains 475 COVID-19, 1583 Normal, and 3949 non-COVID-19 images. Since this dataset is unbalanced, we conducted 10-fold stratified cross-validation, which ensures the same ratio of subjects of each class in each fold.

- **Experiment1-2**: To design a more reliable experiment, we made Dataset-1 balanced using data augmentation for COVID-19 X-ray images and increased the number of images from 475 to 1583 (over-sampling). We used all the data from Normal class (1583), and we picked 1583 images out of 3949 images randomly from

the non-COVID-19 category. Thus, the resulted balanced dataset contains 1583 X-ray images for each category.

- **Experiment1-3**: In this experiment, we built a balanced dataset using under-sampling. We kept all the images in our minority class which is COVID-19 with 475 X-ray images. We then chose the first 475 images of other two classes.

- **Experiment1-4**: This experiment was designed for implementation on Dataset-2 specifically. This unbalanced dataset contains 219 X-ray images from COVID-19 class, 1351 Normal images, and 1345 non-COVID-19 images. To validate our results, we employed 10-fold stratified cross-validation.

- **Experiment1-5**: We designed this experiment to implement our model on a balanced dataset that originated from Dataset-2. To build a balanced dataset, we used under-sampling with the first 219 images for each category.

- **Experiment1-6**: This experiment is designed to implement our model on a larger balanced dataset. We leveraged data augmentation to increase the number of COVID-19 images from 219 to 1345 images. We picked the first 1345 images for non-COVID-19 category to have a balanced dataset which contains 1345 images for each class.

- **Experiment2-1**: This experiment was designed to evaluate the performance of our model on unbalanced CT images using Dataset-5. We conducted 10-fold stratified cross-validation to validate the results.

- **Experiment2-2**: This experiment is designed to implement our model on a different unbalanced CT dataset (Dataset-3). We used 10-fold stratified cross-validation for final evaluation.

- **Experiment2-3**: In this experiment, we made Dataset-3 balanced by reducing the number of COVID-19 images to the same number of images for non-COVID-19 (1229 images for each class).

- **Experiment2-4**: This experiment is designed to evaluate the model performance on a different small dataset, which includes 349 COVID-19 images and 397 non-COVID-19 images. The results are validated using a 10-fold stratified cross-validation.

Table 4.2: First set of experiments using X-ray images

Exp.	Task type	Modality	Design	Number of Samples	Dataset
1-1	3-class	X-ray	Unbalanced	COVID19: 475 Normal: 1583 Non-COVID19: 3949	Dataset-1
1-2	3-class	X-ray	Balanced (Augmented)	COVID19: 1583 Normal: 1583 Non-COVID19: 1583	Dataset-1
1-3	3-class	X-ray	Balanced	COVID19: 475 Normal: 475 Non-COVID19: 475	Dataset-1
1-4	3-class	X-ray	Unbalanced	COVID19: 219 Normal: 1351 Non-COVID19: 1345	Dataset-2
1-5	3-class	X-ray	Balanced	COVID19: 219 Normal: 219 Non-COVID19: 219	Dataset-2
1-6	3-class	X-ray	Balanced (Augmented)	COVID19: 1345 Normal: 1345 Non-COVID19: 1345	Dataset-2

4.4.7 Comparison with Simpler Structures

To better evaluate the value of our novel hierarchical attention structure, we developed two related but simpler deep learning models. In one model, we removed the attention mechanism from our base model, i.e., we fed the feature representations obtained from VGG-16 to the hierarchical structure of Figure 4.4 without the attention layers. In another model, we replaced the whole hierarchical attention network with a fully connected network.

Table 4.3: Second set of experiments using CT scans

Exp.	Task type	Modality	Design	Number of Samples	Dataset
2-1	2-class	CT	Unbalanced	COVID19: 1318 Non-COVID19: 1229	Dataset-5
2-2	2-class	CT	Unbalanced	COVID19: 1259 Non-COVID19: 1229	Dataset-3
2-3	2-class	CT	Balanced	COVID19: 1229 Non-COVID19: 1229	Dataset-3
2-4	2-class	CT	Unbalanced	COVID19: 349 Non-COVID19: 397	Dataset-4

4.5 Results

4.5.1 Model Performance

We conduct ten experiments using two different input data modalities, including CT and X-ray images (Tables 4.2 and 4.3 and). The model demonstrated outstanding performance in detecting COVID-19 for both modalities of data. The first six experiments are focused on datasets with X-ray images as their input. We first start with a large unbalanced dataset (Dataset-1) and evaluated our model using stratified cross-validation. In experiment 1-1, AIDCOV achieved a mean cross-validation accuracy of 98.4% across the 10 folds (Table 4.5). The hierarchical attention model had a sensitivity (true positive rate) of 99.8%, a specificity (true negative rate) of 100%, and an F1-Score of 99.8% for detecting COVID-19 from chest radiography images (Table 4.5). AIDCOV also demonstrated promising performance for correctly classifying non-COVID19 chest images. The model had a sensitivity (specificity) of 98.6% (98.1%) for detecting other viral/bacterial infections and a sensitivity (specificity) of 97.5% (98.7%) for normal chest radiography images. The F1-Score for other infections and normal images were 98.8% and 96.9%, respectively.

We also got help from data augmentation techniques in image processing and created balanced input data in experiment 1-2 (over-sampling). The results in Table 4.5 demonstrates that our model detects COVID-19 patients with a sensitivity of 100%. The mean overall accuracy across ten folds is 98.9%. In our third experiment, we created a balanced dataset by reducing the number of samples of our two larger classes (under-sampling). The results show that our model can successfully detect COVID-19 patients again with a

sensitivity of 100%. The mean accuracy of 10 folds is 99.6%.

Experiments 1-4 to 1-6 are implemented on our second dataset (Dataset-2). Experiment 4 conducted on the original unbalances dataset. We have balanced the input data in experiments 1-5 and 1-6 by reducing and augmenting each class's images. Our results (Table 4.6) show that AIDCOV achieves the mean accuracy (sensitivity) of 98.6% (99.2%), 98.6% (98.8%) and 99.2% (100%) in experiments 1-4, 1-5 and 1-6, respectively.

The second set of experiments includes binary classifications that have been conducted on CT images. Table 4.4 presents the results of four experiments in terms of accuracy, sensitivity, specificity, and F1-Score for each class. Experiments 2-1 and 2-2 are implemented on large unbalanced CT datasets (Dataset-5 and Dataset-3). AIDCOV model achieved the overall mean accuracy of 98.7% and 98.8% on experiments 2-1 and 2-2, respectively. In addition, the model detects COVID-19 patients with a sensitivity of 99% and 99.4% for those experiments.

We balanced the large CT dataset using an under-sampling approach to examine our model's performance on a balanced dataset. The results of experiment 2-3 demonstrate that the AIDCOV model can successfully detect COVID-19 with an accuracy and sensitivity of 99.3% and 99.4% (Table4.4). Finally, we conducted our model on a smaller CT dataset (Dataset-4).

Table 4.4: CT Results

Exp.	Fold	Acc.	**COVID-19**			**Non-COVID19**		
			Sens.	Spec.	F1.	Sens.	Spec.	F1.
2-1	1	92.9	94.3	91.6	92.8	91.6	94.3	93.0
	2	98.4	99.2	97.6	98.5	97.6	99.2	98.4
	3	98.4	98.6	98.3	98.6	98.3	98.6	98.3
	4	100	100	100	100	100	100	100
	5	98.4	98.6	98.2	98.6	98.2	98.6	98.2
	6	99.6	100	99.2	99.6	99.2	100	99.6
	7	99.6	99.3	100	99.6	100	99.3	99.6
	8	99.6	100	99.2	99.6	99.2	100	99.6
	9	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100
	Mean	98.7	99.0	98.4	98.7	98.4	99.0	98.7
2-2	1	94.4	98.4	90.2	94.7	90.2	98.4	94.1
	2	96.8	98.4	95.2	96.8	95.2	98.4	96.7

Table 4.4: CT Results

Exp.	Fold	Acc.	**COVID-19**			**Non-COVID19**		
			Sens.	Spec.	F1.	Sens.	Spec.	F1.
	3	98.8	100	97.5	98.9	97.5	100	98.7
	4	99.6	100	99.2	99.6	99.2	100	99.6
	5	99.6	99.3	100	99.6	100	99.3	99.6
	6	100	100	100	100	100	100	100
	7	99.2	98.3	100	99.2	100	98.3	99.2
	8	100	100	100	100	100	100	100
	9	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100
	Mean	98.8	99.4	98.2	98.9	98.2	99.4	98.8
2-3	1	95.9	95.3	96.6	96.0	96.6	95.3	95.8
	2	98.4	100	96.9	98.3	96.9	100	98.4
	3	99.6	99.2	100	99.6	100	99.2	99.6
	4	99.2	100	98.5	99.1	98.5	100	99.2
	5	99.6	99.3	100	99.6	100	99.3	99.5
	6	100	100	100	100	100	100	100
	7	100	100	100	100	100	100	100
	8	100	100	100	100	100	100	100
	9	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100
	Mean	99.3	99.4	99.2	99.3	99.2	99.4	99.3
2-4	1	78.7	77.4	79.5	75.0	79.5	77.4	81.4
	2	94.7	93.9	95.2	93.9	95.2	93.9	95.2
	3	96.0	97.3	94.7	96.0	94.7	97.3	96.0
	4	97.3	97.4	97.3	97.4	97.3	97.4	97.3
	5	98.7	97.0	100	98.5	100	97.0	98.8
	6	97.3	94.3	100	97.1	100	94.3	97.6
	7	100	100	100	100	100	100	100
	8	100	100	100	100	100	100	100
	9	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100
	Mean	96.3	95.7	96.7	95.8	96.7	95.7	96.6

The results in Table 4.4 presents the overall accuracy and sensitivity of 96.3% and 95.7%, respectively. These results suggest that AIDCOV performs well in detecting COVID-19, other viral/bacterial infections, and normal cases based on the chest radiography images.

4.5.2 The Value of Hierarchical Attention Structure

Using the first dataset (Dataset-1), the hierarchical model (without the attention layers) and the fully connected structure had lower accuracy levels than the hierarchical attention model. The hierarchical (no attention) model had an overall cross-validation accuracy of 97.5%, slightly lower than the hierarchical attention model. The sensitivity of this model to detect COVID-19 from chest radiography images was 99.3%, similar to the hierarchical attention model. The model with a fully connected structure had the poorest performance among the three models and resulted in an overall cross-validation accuracy of 96.0% when tested on our dataset. This model had a sensitivity of 93.3% in detecting COVID-19.

These results highlight the value of our novel hierarchical structure, which can capture the dependency of all feature representation blocks (obtained from the VGG-16) in both horizontal and vertical directions and improve model performance. The attention mechanism in our hierarchical attention model also helps with the interpretability and transparency of the model predictions.

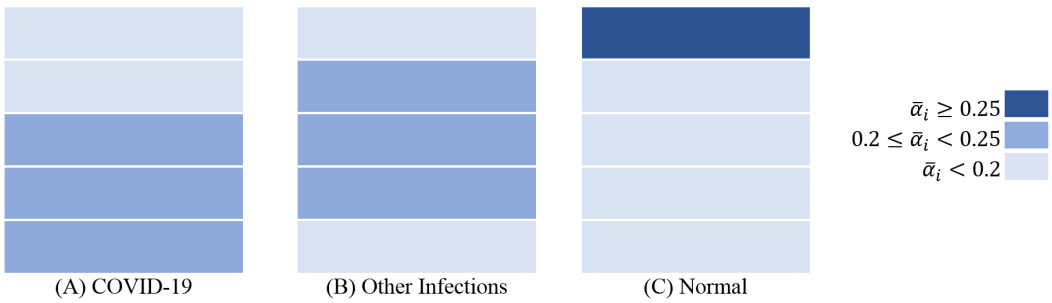


Figure 4.5: Average attention score for different zones of the lung for each image class

4.5.3 Interpretability

The strengths of AIDCOV are not limited to its superior sensitivity, specificity, and PPV in detecting COVID-19 and other infections. To gain deeper insights into how the model makes its predictions and identify the areas of the lung affected by the infection, we

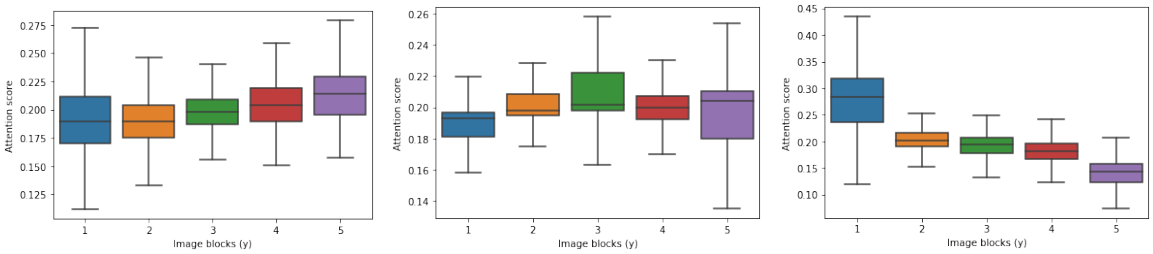
Table 4.5: X-ray Results

Exp.	Fold	Acc.	**COVID-19**			**Normal**			**Non-COVID19**		
			Sens.	Spec.	F1.	Sens.	Spec.	F1.	Sens.	Spec.	F1.
1-1	1	96.0	100	99.8	98.6	90.9	97.6	91.6	97.4	93.3	97.3
	2	97.2	100	100	100	94.9	98.1	95.2	97.9	95.9	97.8
	3	98.2	100	100	100	96.8	98.7	96.5	98.5	97.5	98.6
	4	98.3	98.1	100	99.0	96.1	99.1	96.8	99.2	97.1	98.9
	5	98.0	100	100	100	99.3	97.6	96.1	97.3	99.5	98.5
	6	98.7	100	100	100	98.6	98.7	97.3	98.5	99.0	99.0
	7	99.5	100	100	100	100	99.3	99.1	99.2	100	99.6
	8	99.2	100	100	100	99.4	99.1	98.4	99.0	99.5	99.4
	9	99.0	100	100	100	98.8	99.1	98.3	98.9	99.1	99.2
	10	99.8	100	100	100	100	99.8	99.7	99.7	100	99.9
	Mean	98.4	99.8	100	99.8	97.5	98.7	96.9	98.6	98.1	98.8
1-2	1	97.1	100	100	100	96.1	97.5	95.5	95.0	98.1	95.6
	2	97.9	100	100	100	95.3	99.1	96.6	98.2	97.7	97.0
	3	97.9	100	100	100	97.6	98.1	97.0	95.9	98.8	96.5
	4	99.2	100	100	100	97.2	100	98.6	100	98.7	98.7
	5	99.4	100	100	100	98.8	99.7	99.1	99.3	99.4	98.9
	6	99.8	100	100	100	99.3	100	99.7	100	99.7	99.7
	7	99.8	100	100	100	99.4	100	99.7	100	99.7	99.7
	8	99.4	100	100	100	98.2	100	99.1	100	99.0	99.1
	9	99.4	100	100	100	99.4	99.4	99.0	98.9	99.7	99.2
	10	99.8	100	100	100	100	99.7	99.7	99.3	100	99.7
	Mean	98.9	100	100	100	98.1	99.3	98.4	98.7	99.1	98.4
1-3	1	97.2	100	97.9	98.0	96.1	97.8	96.1	95.5	100	97.7
	2	99.3	100	100	100	97.5	100	98.7	100	99.0	99.0
	3	100	100	100	100	100	100	100	100	100	100
	4	99.3	100	100	100	100	99.1	98.7	98.4	100	99.2
	5	100	100	100	100	100	100	100	100	100	100
	6	100	100	100	100	100	100	100	100	100	100
	7	100	100	100	100	100	100	100	100	100	100
	8	100	100	100	100	100	100	100	100	100	100
	9	100	100	100	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100	100	100	100
	Mean	99.6	100	99.8	99.8	99.4	99.7	99.4	99.4	99.9	99.6

Table 4.6: X-ray Results

Exp.	Fold	Acc.	**COVID-19**			**Normal**			**Non-COVID19**		
			Sens.	Spec.	F1.	Sens.	Spec.	F1.	Sens.	Spec.	F1.
1-4	1	95.9	100	100	100	97.8	94.1	95.8	93.4	98.1	95.5
	2	98.6	95.5	100	97.7	98.5	98.7	98.5	99.3	98.7	98.9
	3	99.7	100	100	100	100	99.3	99.7	99.2	100	99.6
	4	95.2	100	99.6	97.4	99.2	92.6	95.2	90.9	99.3	94.9
	5	99.3	96.9	100	98.4	99.3	99.4	99.3	100	99.4	99.6
	6	98.3	100	100	100	100	97.1	97.9	96.7	100	98.3
	7	100	100	100	100	100	100	100	100	100	100
	8	99.3	100	100	100	100	98.7	99.3	98.4	100	99.2
	9	100	100	100	100	100	100	100	100	100	100
	10	99.7	100	100	100	99.3	100	99.6	100	99.4	99.6
	Mean	98.6	99.2	100	99.4	99.4	98.0	98.5	97.8	99.5	98.6
1-5	1	92.4	95.0	100	97.4	100	87.8	90.9	81.0	100	89.5
	2	98.5	100	100	100	95.2	100	97.6	100	97.9	97.4
	3	97.0	100	100	100	91.7	100	95.7	100	95.0	96.3
	4	100	100	100	100	100	100	100	100	100	100
	5	100	100	100	100	100	100	100	100	100	100
	6	100	100	100	100	100	100	100	100	100	100
	7	98.5	93.3	100	96.6	100	97.6	98.0	100	100	100
	8	100	100	100	100	100	100	100	100	100	100
	9	100	100	100	100	100	100	100	100	100	100
	10	100	100	100	100	100	100	100	100	100	100
	Mean	98.6	98.8	100	99.4	98.7	98.5	98.2	98.1	99.3	98.3
1-6	1	96.0	100	99.6	99.7	96.3	96.3	94.5	91.3	98.2	93.6
	2	99.0	100	100	100	99.2	99.0	98.3	97.8	99.6	98.5
	3	99.8	100	100	100	99.2	100	99.6	100	99.6	99.7
	4	99.5	100	100	100	99.2	99.6	99.2	99.2	99.6	99.2
	5	99.5	100	100	100	100	99.3	99.3	98.4	100	99.2
	6	99.8	100	100	100	99.3	100	99.6	100	99.6	99.7
	7	99.8	100	100	100	99.3	100	99.6	100	99.6	99.6
	8	99.5	100	100	100	100	99.2	99.4	98.5	100	99.3
	9	99.8	100	100	100	100	99.7	99.6	99.3	100	99.6
	10	99.8	100	100	100	99.3	100	99.7	100	99.6	99.6
	Mean	99.2	100	100	100	99.2	99.3	98.9	98.5	99.6	98.8

extracted attention scores for each image. Figure 4.5 illustrates the areas of the chest radiography images for each type (COVID-19, other infections, and normal) that the model paid more attention to based on the final attention scores averaged over all images of that type. We also provide the box plot of horizontal attention scores in Figure 4.6. The numbers 1 to 5 in the horizontal axis of all three boxplots represent the y blocks. The number "1" presents the upper, and the number "5" represents the lowest parts of the image, respectively. We see that in both Figures 4.5 and 4.6 the model identified signs of SARS-CoV-2 infection mostly in the lower zone and other infections around the middle zone of the lung. Since the model determines the attention scores relatively, the normal radiography images had the highest attention score on the very top level corresponding to the upper zone of the lung. In other words, since the lower and middle zones of the lung contain signs of COVID-19 and other infections and, therefore, dominate the attention scores for these zones, the normal images receive lower attention scores for the middle and lower zones and higher attention scores for the upper zone of the lung.



(a) COVID-19 CXRs (b) Non-COVID19 CXRs (c) Normal CXRs

Figure 4.6: Horizontal attention scores boxplots

Moreover, AIDCOV can identify the specific blocks within each individual's chest image that may include subtle signs of infection via the attention scores of both encoding levels (i.e., horizontal and vertical). To better illustrate the model's interpretability, we included the chest images of 6 patients with COVID-19 in Figures 4.7 and 4.8. We indicated the attention scores for each zone of the lung and each block of the image. Given that there are five zones in each image, we highlighted the zones that received an attention score of 0.2 or higher. Then, within those zones, we highlighted the blocks that received an attention score greater than or equal to 0.2.

In Figure 4.7, the first patient (Panels A and B) was admitted to the hospital with fever, shortness of breath (dyspnea), and low oxygen saturation. Radiological worsening with changes within the lower lobes was noted on her chest Xray. Our model correctly gave a higher attention score to the lower zone of the lung and identified the specific

areas of the lower zone that demonstrated signs of SARS-CoV-2 infection.

The second patient (Panels C and D) in Figure 4.7 came to the hospital with suspected pneumonia. The radiographic investigation indicated abnormalities in the middle part of the right lung. As seen in Figure 4.7.C and 4.7.D, our model correctly identified these abnormalities in the middle zone of the right lung. (Note that the right lung appears on the left side of the radiography image.)

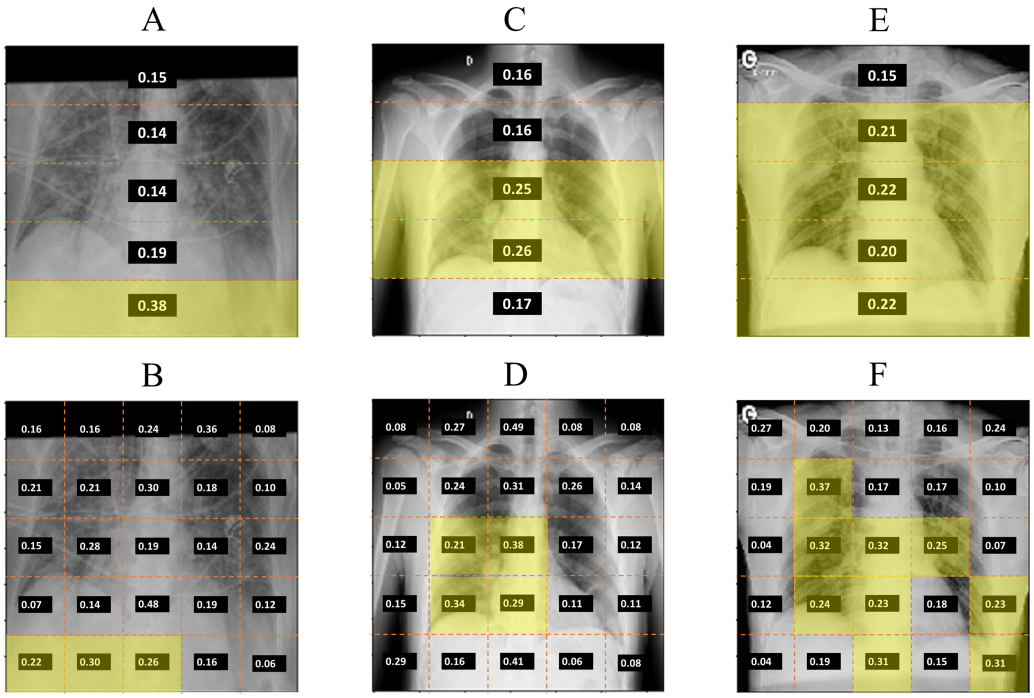


Figure 4.7: Attention scores for different zones of the lung (horizontal level) and different blocks of the image for 3 patients with COVID-19. Signs of COVID-19 were detected in the lower zone for Patient 1 (A-B), middle zone for Patient 2 (C-D), and lower and middle zones for Patient 3 (E-F).

The chest X-ray image of the third patient (Panels E and F) in Figure 4.7 indicated small consolidation in the right upper lobe and ground-glass opacities in both lower lobes. Our deep learning model correctly identified the middle and the lower zones as the areas with abnormalities (Figure 4.7.E). It further pointed to the right lung (that appears on the left side of Figure 4.7.F) and the lower zone of both lungs.

We have provided five more instances of COVID-19 chest X-ray images with their horizontal and vertical attention scores (Figures 4.10 and 4.11). The clinical notes of these images are presented in the caption. We can see that all the result of the interpretability of our model confirms their corresponding clinical notes.

We present chest CT images of three more patients in Figure 4.8. In CT images, ground-glass opacities (GGOs) are one of the main signs of the COVID-19 infections in patients' chest images [170–172]. Basically, GGOs are the transparent lighter (grey color) pattern within chest CT images where the lung's underlying structures are still visible. In Figure 4.8, panels H and I belong to the first patient; Panels J and K belong to the second patient, and Panels L and M are related to the third patient. We can see that the AIDCOV model can automatically localize the GGOs in all of these CT images (Figure 4.8). AIDCOV allocates higher attention scores to those infected regions in both horizontal and vertical levels. These scores can help radiologists to accelerate COVID-19 diagnosis. We made our code available at: `https://github.com/marynik66/COVID-19`

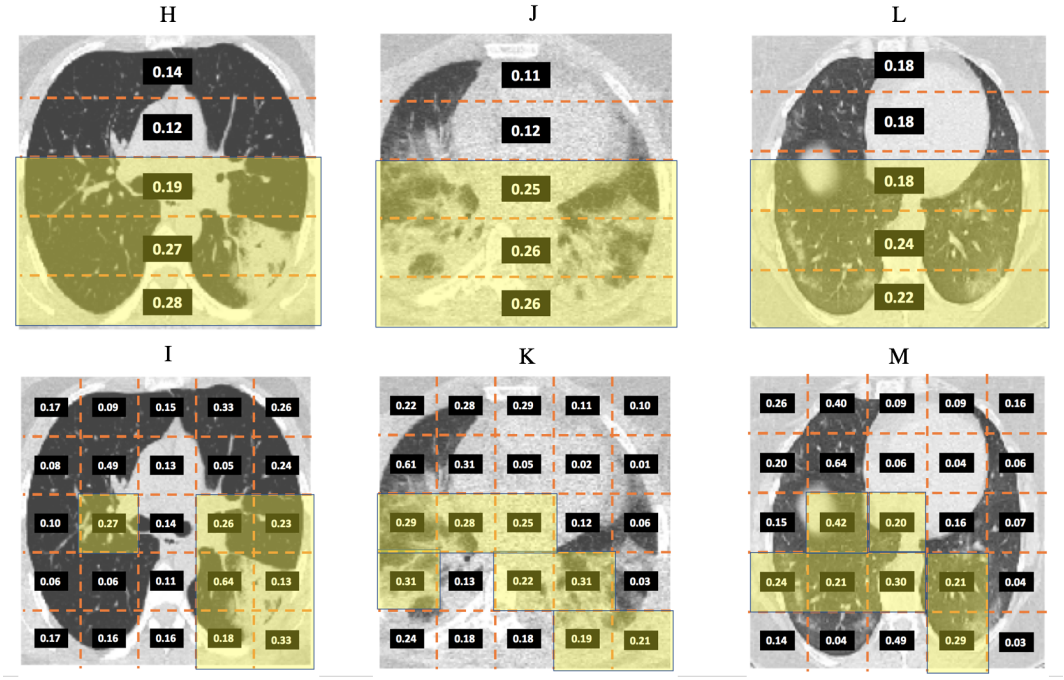


Figure 4.8: Attention scores for abnormalities in chest CT images (horizontal level) and different blocks of the image for 3 patients with COVID-19.

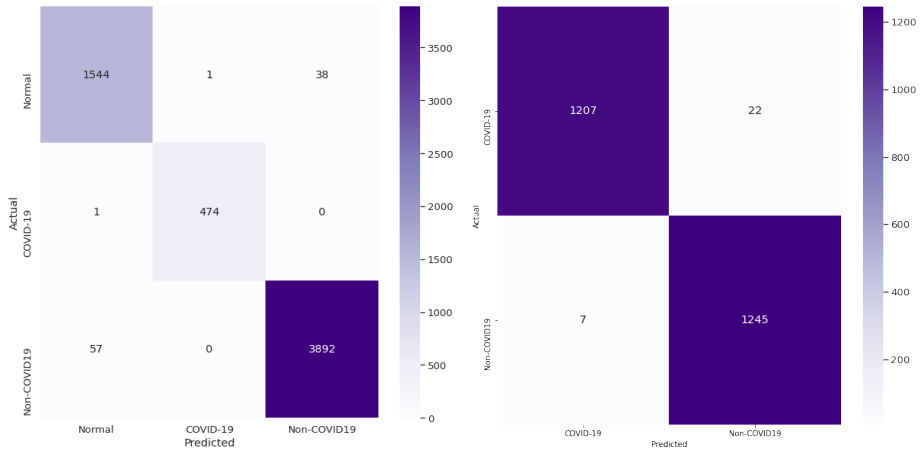
4.6 Discussion

In this study, we introduced AIDCOV, an artificial intelligence model for the detection of COVID-19 from chest radiography images. The model performs multi-class prediction, i.e., it labels each image as belonging to a person with either COVID-19, other infections, or no pneumonia (i.e., normal). AIDCOV leverages VGG-16 to obtain a low-dimensional

feature representation for each image. It then encodes the features along the horizontal (width) and vertical (height) directions using a novel two-level hierarchical attention structure. This allows the model to capture the horizontal and vertical dependencies of the features, which is ignored in a fully connected network. The attention mechanism further helps make the model interpretable and gives transparency to model predictions.

We trained and tested AIDCOV on publicly available datasets (Table 4.1). We designed six different experiments on both X-ray and CT modalities (Tables 4.2 and 4.3). Using the first unbalanced dataset (Dataset-1), We demonstrated that the model has an overall accuracy of 98.4% across the ten folds of cross-validation. AIDCOV showed excellent sensitivity (99.8%), specificity (100%) in detecting COVID-19 from chest X-ray images. Our model proved a better performance when implemented on a balanced dataset. The results of experiment 1-3 demonstrated outstanding accuracy (sensitivity) of 99.6% (100%) (Table 4.5).

In addition, the AIDCOV model kept its superior performance when it was run on CT images as the input. Our model obtained an accuracy (sensitivity) of 99.3% (99.4%) using Dataset-3 which is a large balanced CT dataset (Experiment 2-2) (Table 4.4). We observed that our model obtained an accuracy (sensitivity) of 96.3% (95.7%) when implemented on Dataset-4 (Experiment 2-4). The model does not perform as accurate in the other three experiments (Experiment 2-1, 2-2, and 2-3). This can be due to the lower quality of CT images of input data. Since Dataset-4 is extracted from CT images presented in preprint manuscripts, the quality of CT images may be diminished.



(a) Experiment 1-1 (X-ray images) (b) Experiment 2-2 (CT scans)

Figure 4.9: Confusion matrix

Figure 4.9 represents the confusion matrix for experiments 1-1 and 2-2. In the

three-class problem using X-ray images (Figure 4.9(a)), it is clear that our model miss-classified only one COVID-19 image out of 475 images as normal. Using CT images in a binary classification, our model miss-classified 22 COVID-19 CT scans as non-COVID19 class. The high sensitivity and specificity of our model are critical in practice since a false negative result can lead to not isolating an individual with COVID-19, which can subsequently result in many transmissions from that person to others, including to the healthcare workers. Transmission from patients to healthcare workers results in undermining the healthcare capacity. In China, about 5%, and in Italy, about 10% of infections were among the healthcare workers [173]. Thus, leveraging chest radiographs in hospitals to identify patients with COVID-19 and using appropriate personal protective equipment (PPE) when providing care to these patients can be beneficial.

To better assess the value of our hierarchical attention structure, we developed two simpler models. In one model, we kept the hierarchical structure but removed the attention mechanism from it. In another model, we replaced the whole hierarchical attention structure with a fully connected network. The model with a fully connected network had the poorest performance among the three models and achieved an accuracy of 96.0%. The hierarchical model without attention had an accuracy of 97.5%, which is only slightly below our hierarchical attention model's accuracy. However, the main advantage of including an attention mechanism is making the model interpretable and providing transparency to model predictions.

Using an analysis of the attention scores that AIDCOV gave to COVID-19, other infections, and normal samples, we demonstrated that the model identified signs of SARS-CoV-2 infection, on average, in the lower and middle zones of the lung more frequently. This is consistent with findings from radiology reports, which indicate that abnormalities due to SARS-CoV-2 infection are more commonly found in inferior and middle lobes of the lung, corresponding to the lower and middle zones of the chest radiography images [174–178]. Furthermore, our hierarchical attention model divides each image into 25 blocks (5×5) and provides the attention score for each block. This can shed light on particular regions within each radiography image that may contain subtle signs of infection and makes the model more transparent and trustable.

Currently, the primary screening tool to detect COVID-19 is reverse transcription-polymerase chain reaction (RT-PCR), which is a laboratory test to detect viral nucleic acid [123]. However, not only the capacity for RT-PCR testing is limited (both in the U.S. and in many other countries), but also the result return time can range between several minutes to hours [124]. More importantly, the RT-PCT test has a sensitivity of around

70%, which can be even lower depending on the assay, type and quality of the specimen, and the disease stage [125–127, 179, 180]. That is, about three in ten individuals with COVID-19 receive a false-negative test result. In this context, AIDCOV provides an accurate and fast alternative to RT-PCR testing that can quickly detect COVID-19 from chest radiography images or can be additionally used to augment the RT-PCR tests. Moreover, given that RT-PCR testing kits are in short supply in many resource-limited settings, using chest Xray and CT scan images, which are generally more available around the world, as a screening tool for COVID-19 may be worth considering. We demonstrated that our model is a viable alternative to RT-PCR testing when symptoms are visible in the lungs, and it can be used to detect and quarantine infected individuals and stop the spread of the SARS-CoV-2 virus.

A number of other artificial intelligence models to detect COVID-19 from chest radiography images have also been developed recently (Table 4.7). Li et al. developed COVNet, a neural network model to detect COVID-19 and community-acquired pneumonia from CT images [154]. COVNet demonstrated a sensitivity of 90.0% and a specificity of 96.0% in detecting COVID-19. It also achieved a sensitivity and specificity of 86.9% and 92.3%, respectively, in detecting community-acquired pneumonia from CT images. Wang et al. developed COVID-Net using a deep convolutional neural network structure designed for detecting COVID-19 and other infections from chest X-ray images [147]. COVID-Net reports a sensitivity of 91.0% and a positive predictive value (PPV) of 98.9% in detecting COVID-19. The sensitivity and PPV of COVID-Net in detecting other infections were 94.0% and 91.3%, respectively. Zhang et al. developed a deep learning model to detect COVID-19 from chest X-ray images [157]. Their model had a sensitivity (specificity) of 71.7% (73.8%). Table 4.7 presents the performance of other state-of-the-art studies with the characteristics of their datasets to detect COVID-19 using AI-based models. The comparison of results in Table 4.7 confirms that AIDCOV outperforms other models in terms of overall accuracy, sensitivity, and specificity using similar datasets.

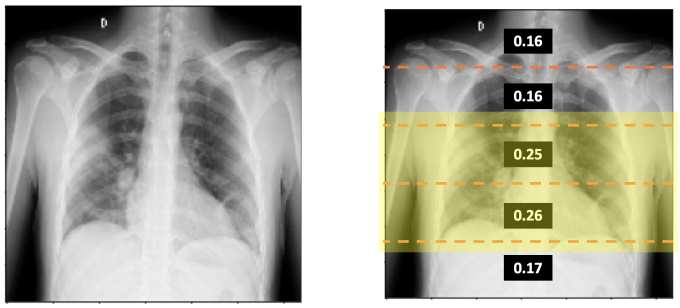
Our study has a number of limitations and, therefore, our results should be interpreted with caution. First, as was the case with the other related studies, our dataset was limited in size and had only 475 X-ray and 1311 CT images of individuals with COVID-19. Further validation on datasets with a larger number of chest radiography images from patients with COVID-19 would be valuable. Second, chest X-ray images may not show signs of SARS-CoV-2 infections in the early stages of illness. Abnormalities are more likely to develop over the course of the disease [183, 184]. However, some preliminary data suggest that abnormalities may show in CT images in the presymptomatic stage

Table 4.7: Results comparison with state-of-the-art studies

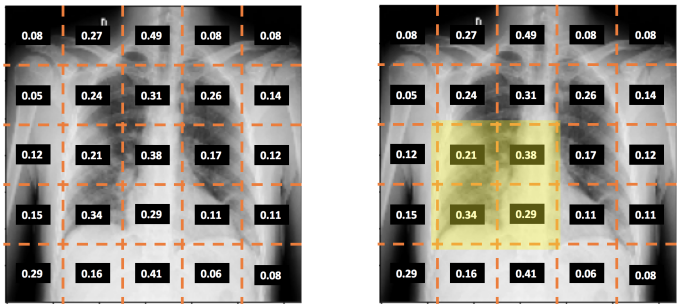
Study	Modality	COVID19	Normal	Other	Acc.%	Sens.%	Spec.%
Wang et al. [147]	X-ray	358	8066	5538	93.3	91.0	-
Chowdhury et al. [139]	X-ray	219	1341	1345	97.9	97.9	98.8
Apostolopoulos et al. [156]	X-ray	224	700	504	93.5	92.8	98.7
Sethy et al. [181]	X-ray	25	25	-	95.4	97.3	93.5
Zhang et al. [157]	X-ray	106	107	-	72.7	71.7	73.8
Hemdan et al. [182]	X-ray	25	25	-	90.0	100	-
Narin et al. [143]	X-ray	50	50	-	98.0	96.0	100
Ozturk et al. [145]	X-ray	125	500	500	87.0	85.3	92.1
Kumar et al. [163]	X-ray	62	1341	1345	97.7%	97.7%	98.8%
Yang et al. [148]	CT	349	-	397	89.1	-	-
Soares et al. [141]	CT	1252	-	1229	97.4	95.5	-
Song et al. [160]	CT	777	708	-	86.0	96.0	-
Zheng et al. [159]	CT	313	229	-	90.8	-	-
Wang et al. [158]	CT	195	258	-	82.9	81.0	84.0
Ardakani et al. [161]	CT	510	510	-	99.5	100	99.0
Li et al. [154]	CT	1296	1735	1325	-	90.0	96.0
Butt et al. [155]	CT	357	1353	-	86.7	-	-
Our model	X-ray	475	1583	3949	98.4	99.8	100
Our model	X-ray	475	475	475	99.6	100	99.8
Our model	CT	1259	-	1229	98.8	99.4	98.2
Our model	CT	1229	-	1229	99.3	99.4	99.2

and prior to the detection of viral RNA from upper respiratory specimens [131, 132]. Our dataset does not include information on time since symptom onset or the disease stage at the time the image was taken; thus, we could not assess our model's accuracy based on these factors. Using chest X-ray to detect COVID-19 is currently not being used as the first modality of testing perhaps due to its higher cost compared to RT-PCR testing. However, we showed that a higher sensitivity and specificity can be achieved a very if we use AIDCOV on chest images.

In conclusion, AIDCOV demonstrated high sensitivity, specificity, and positive predictive value in detecting COVID-19 from chest X-ray and CT images. Given that radiography is widely available in many countries around the world, AIDCOV can be used in conjunction with or instead of RT-PCR testing (e.g., where RT-PCR testing is unavailable) to find individuals infected with the SARS-CoV-2 virus, isolate them, and prevent the spread of COVID-19.

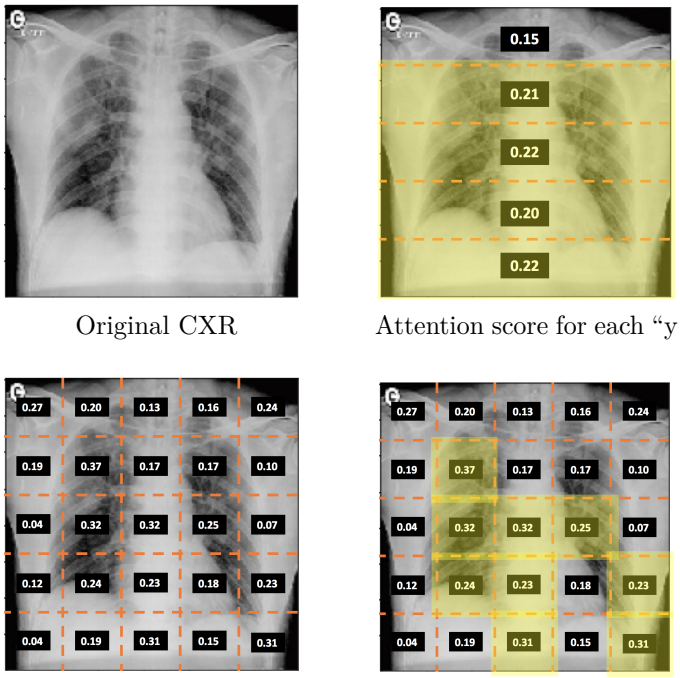


Original CXR Attention score for each “y"



Original CXR Attention score for each “x"

Figure 4.10: The radiographic investigation demonstrates the presence of an increase in the peribroncovascular interstitial plot with associated parenchymal thickenings especially in the **basal and lateral subpleural site at the level of the middle-upper field of the right lung.**



Original CXR Attention score for each “y"

Original CXR Attention score for each “x"

Figure 4.11: Small consolidation in **right upper lobe and ground-glass opacities in both lower lobes** were observed on high-resolution computed tomography scan

Chapter 5
Conclusion and Future works

In this dissertation, we developed interpretable deep learning methods to detect chronic and infectious diseases. We studied three specific problems in healthcare. In the first problem, we studied the detection of Alzheimer's disease using interview transcripts of patients. In this study, we developed an interpretable hierarchical deep learning model to detect the onset of Alzheimer's disease from interview transcripts of individuals who were asked to describe a picture. We demonstrated the interpretability of the model with the importance score of words, sentences, and transcripts extracted from our three-level hierarchical model. In the second problem, we studied the detection of Alzheimer's disease using raw audio interviews of patients. Since linguistic deficits are the first signs of cognitive decline in AD patients, speech analysis can help with the early detection of Alzheimer's, which will lead to better management of the disease. In this study, we developed a hierarchical deep audio model to detect the onset of Alzheimer's. Our deep hierarchical model is interpretable such that it signals the signs of memory loss in patients' speech very accurately. In the third problem, we studied the detection of COVID-19 using chest radiography images. As the Coronavirus Disease 2019 (COVID-19) pandemic continues to grow globally, testing to detect COVID-19 and isolating individuals who test positive remains to be the primary strategy for preventing community spread of the disease. The current gold standard method of testing for COVID-19 is the reverse transcription-polymerase chain reaction (RT-PCR) test. However, the RT-PCR test has an imperfect sensitivity (around 70%), is time-consuming and labor-intensive, and is in short supply, particularly in resource-limited countries. Therefore, automatic and accurate detection of COVID-19 using medical imaging modalities such as chest X-ray and Computed Tomography (CT), which are more widely available and accessible, can be beneficial. We develop a novel hierarchical attention neural network model to classify chest radiography images as belonging to a person with either COVID-19, other infections,

or no pneumonia (i.e., normal). We refer to this model as Artificial Intelligence for Detection of COVID-19 (AIDCOV). The hierarchical structure in AIDCOV captures the dependency of features and improves model performance while the attention mechanism makes the model interpretable and transparent. AIDCOV can be used in conjunction with or instead of RT-PCR testing (where RT-PCR testing is unavailable) to identify and isolate individuals with COVID-19 and prevent onward transmission to the general population and healthcare workers.

Three main modalities of medical data, including text, audio, and image, were considered in the deep learning models. We demonstrated that using attention mechanism combined with hierarchical architecture can be used across different modalities. Our models demonstrated excellent performance for both prediction and interpretability capability. Our interpretable deep learning models can be extended to other types of medical data. Our hierarchical deep audio model can be applied in signal processing, including to study electroencephalogram (EEG) and electrocardiogram (ECG) data. Our AIDCOV structures can be used for other types of medical images, including pathology CT scans and brain scans, to detect anomalies. Moreover, our interpretable text model can extract useful information from electronic health records (EHRs). In the future, we aim to combine different data modalities to investigate how it can improve the prediction results. For instance, we can combine text and audio interviews to detect the onset of Alzheimer's disease. We also aim to use the recent word embedding tools, including the pre-training of deep Bidirectional Transformers for Language Understanding (BERT) [60], to investigate how they can improve the model prediction performance. Last but not least, we hope to build a mobile app to detect AD's onset using audio recordings of patients in a real-time manner. Such an app can help improve the detection of AD such that susceptible individuals can see their primary doctors for further experiments and start AD treatment earlier before the cognitive decline worsens.

Bibliography

[1] Becker, J. T., F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle (1994) "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, **51**(6), pp. 585–594.

[2] Dai, W., C. Dai, S. Qu, J. Li, and S. Das (2017) "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 421–425.

[3] Abdoli, S., P. Cardinal, and A. L. Koerich (2019) "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Systems with Applications*, **136**, pp. 252–263.

[4] LeCun, Y., Y. Bengio, and G. Hinton (2015) "Deep learning," *nature*, **521**(7553), p. 436.

[5] Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012) "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105.

[6] Bahdanau, D., K. Cho, and Y. Bengio (2014) "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*.

[7] Karpathy, A. and L. Fei-Fei (2015) "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137.

[8] Hinton, G., L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al. (2012) "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, **29**.

[9] Arisoy, E., T. N. Sainath, B. Kingsbury, and B. Ramabhadran (2012) "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, Association for Computational Linguistics, pp. 20–28.

[10] Katzman, R. (1976) "The prevalence and malignancy of Alzheimer disease: a major killer," *Archives of neurology*, **33**(4), pp. 217–218.

[11] Williams, J. A., A. Weakley, D. J. Cook, and M. Schmitter-Edgecombe (2013) "Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia," in *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 71–76.

[12] Johnson, P., L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, L. S. Macaulay, K. A. Ellis, C. Szoeke, R. N. Martins, et al. (2014) "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease," *BMC bioinformatics*, **15**(16), p. S11.

[13] Mitolo, M., S. Gardini, P. Caffarra, L. Ronconi, A. Venneri, and F. Pazzaglia (2015) "Relationship between spatial ability, visuospatial working memory and self-assessed spatial orientation ability: a study in older adults," *Cognitive processing*, **16**(2), pp. 165–176.

[14] Goryawala, M., Q. Zhou, W. Barker, D. A. Loewenstein, R. Duara, and M. Adjouadi (2015) "Inclusion of neuropsychological scores in atrophy models improves diagnostic classification of Alzheimer's disease and Mild Cognitive Impairment," *Computational intelligence and neuroscience*, **2015**, p. 56.

[15] Association, A. et al. (2016) "2016 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, **12**(4), pp. 459–509.

[16] Querbes, O., F. Aubry, J. Pariente, J.-A. Lotterie, J.-F. Démonet, V. Duret, M. Puel, I. Berry, J.-C. Fort, P. Celsis, et al. (2009) "Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve," *Brain*, **132**(8), pp. 2036–2047.

[17] Ewers, M., R. A. Sperling, W. E. Klunk, M. W. Weiner, and H. Hampel (2011) "Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia," *Trends in neurosciences*, **34**(8), pp. 430–442.

[18] Scheubert, L., M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen (2012) "Tissue-based Alzheimer gene expression markers–comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC bioinformatics*, **13**(1), p. 266.

[19] Orimaye, S. O., J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri (2017) "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC bioinformatics*, **18**(1), p. 34.

[20] Sundermeyer, M., R. Schlüter, and H. Ney (2012) "LSTM neural networks for language modeling," in *Thirteenth Annual Conference of the International Speech Communication Association*.

[21] Ribeiro, M. T., S. Singh, and C. Guestrin (2016) "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, pp. 1135–1144.

[22] Terry, R. D. and P. Davies (1980) "Dementia of the Alzheimer type," *Annual review of neuroscience*, **3**(1), pp. 77–95.

[23] James, B. D., S. E. Leurgans, L. E. Hebert, P. A. Scherr, K. Yaffe, and D. A. Bennett (2014) "Contribution of Alzheimer disease to mortality in the United States," *Neurology*, **82**(12), pp. 1045–1050.

[24] Laske, C., H. R. Sohrabi, S. M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, et al. (2015) "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's & Dementia*, **11**(5), pp. 561–578.

[25] Mortamais, M., J. A. Ash, J. Harrison, J. Kaye, J. Kramer, C. Randolph, C. Pose, B. Albala, M. Ropacki, C. W. Ritchie, et al. (2017) "Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility," *Alzheimer's & Dementia*, **13**(4), pp. 468–492.

[26] Damian, A. M., S. A. Jacobson, J. G. Hentz, C. M. Belden, H. A. Shill, M. N. Sabbagh, J. N. Caviness, and C. H. Adler (2011) "The Montreal Cognitive Assessment and the Mini-Mental State Examination as screening instruments for cognitive impairment: item analyses and threshold scores," *Dementia and geriatric cognitive disorders*, **31**(2), pp. 126–131.

[27] Mitchell, A. J. (2009) "A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment," *Journal of psychiatric research*, **43**(4), pp. 411–431.

[28] Albert, M. S., S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, et al. (2011) "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association*, **7**(3), pp. 270–279.

[29] Alzheimer's Association (2018) "2018 Alzheimer's disease facts and figures," *Alzheimers Dement*, **14**(3), pp. 367–429.

[30] Aramaki, E., S. Shikata, M. Miyabe, and A. Kinoshita (2016) "Vocabulary size in speech may be an early indicator of cognitive impairment," *PloS one*, **11**(5), p. e0155195.

[31] Blair, M., C. A. Marczinski, N. Davis-Faroque, and A. Kertesz (2007) "A longitudinal study of language decline in Alzheimer's disease and frontotemporal dementia," *Journal of the International Neuropsychological Society*, **13**(2), pp. 237–245.

[32] Fraser, K. C., J. A. Meltzer, and F. Rudzicz (2016) "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, **49**(2), pp. 407–422.

[33] Bengio, Y. et al. (2009) "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, **2**(1), pp. 1–127.

[34] Orr, G. B. and K.-R. Müller (2003) *Neural networks: tricks of the trade*, Springer.

[35] Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014) "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*.

[36] Lopez-de Ipiña, K., J. B. Alonso, J. Solé-Casals, N. Barroso, M. Faundez-Zanuy, M. Ecay-Torres, C. M. Travieso, A. Ezeiza, A. Estanga, et al. (2012) "Alzheimer disease diagnosis based on automatic spontaneous speech analysis," .

[37] Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013) "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

[38] Sainath, T. N., A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran (2013) "Deep convolutional neural networks for LVCSR," in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, pp. 8614–8618.

[39] Bordes, A., S. Chopra, and J. Weston (2014) "Question answering with subgraph embeddings," *arXiv preprint arXiv:1406.3676*.

[40] Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011) "Natural language processing (almost) from scratch," *Journal of machine learning research*, **12**(Aug), pp. 2493–2537.

[41] Zimmerer, V. C., M. Wibrow, and R. A. Varley (2016) "Formulaic language in people with probable Alzheimer's disease: A frequency-based approach," *Journal of Alzheimer's Disease*, **53**(3), pp. 1145–1160.

[42] Wankerl, S., E. Nöth, and S. Evert (2017) "An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language." in *INTERSPEECH*, pp. 3162–3166.

[43] Orimaye, S. O., J. S.-M. Wong, and C. P. Wong (2018) "Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia," *PloS one*, **13**(11), p. e0205636.

[44] Karlekar, S., T. Niu, and M. Bansal (2018) "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models," *arXiv preprint arXiv:1804.06440*.

[45] Chen, J., J. Zhu, and J. Ye (2019) "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech." in *INTERSPEECH*, pp. 4085–4089.

[46] Fritsch, J., S. Wankerl, and E. Nöth (2019) "Automatic diagnosis of Alzheimer's disease using neural network language models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5841–5845.

[47] Pan, Y., B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen (2019) "Automatic Hierarchical Attention Neural Network for Detecting AD." in *Interspeech*, pp. 4105–4109.

[48] Chien, Y.-W., S.-Y. Hong, W.-T. Cheah, L.-H. Yao, Y.-L. Chang, and L.-C. Fu (2019) "An Automatic Assessment System for Alzheimer's Disease Based on Speech Using feature Sequence Generator and Recurrent neural network," *Scientific Reports*, **9**(1), pp. 1–10.

[49] Kong, W., H. Jang, G. Carenini, and T. Field (2019) "A Neural Model for Predicting Dementia from Language," in *Machine Learning for Healthcare Conference*, pp. 270–286.

[50] Kim, Y. (2014) "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*.

[51] Zhang, X. and M. Lapata (2014) "Chinese poetry generation with recurrent neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680.

[52] Irsoy, O. and C. Cardie (2014) "Opinion mining with deep recurrent neural networks," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 720–728.

[53] Bengio, Y., P. Simard, and P. Frasconi (1994) "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, **5**(2), pp. 157–166.

[54] Hochreiter, S. and J. Schmidhuber (1997) "Long short-term memory," *Neural computation*, **9**(8), pp. 1735–1780.

[55] Rong, X. (2014) "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*.

[56] Pennington, J., R. Socher, and C. Manning (2014) "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

[57] Koren, Y., R. Bell, and C. Volinsky (2009) "Matrix factorization techniques for recommender systems," *Computer*, **42**(8), pp. 30–37.

[58] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013) "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

[59] Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018) "Deep contextualized word representations," in *Proc. of NAACL*.

[60] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

[61] Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019) "Language models are unsupervised multitask learners," *OpenAI Blog*, **1**(8), p. 9.

[62] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017) "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008.

[63] MacWhinney, B. (2000) *The CHILDES project: The database*, vol. 2, Psychology Press.

[64] Pascanu, R., T. Mikolov, and Y. Bengio (2013) "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, pp. 1310–1318.

[65] Schuster, M. and K. K. Paliwal (1997) "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, **45**(11), pp. 2673–2681.

[66] Chorowski, J. K., D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio (2015) "Attention-based models for speech recognition," in *Advances in neural information processing systems*, pp. 577–585.

[67] Kingma, D. P. and J. Ba (2014) "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*.

[68] Salton, G. and M. J. McGill (1986) "Introduction to modern information retrieval," .

[69] Manning, C. D. and H. Schütze (1999) *Foundations of statistical natural language processing*, MIT press.

[70] Jurafsky, D. and J. H. Martin (2014) *Speech and language processing*, vol. 3, Pearson London:.

[71] Orimaye, S. O., J. S.-M. Wong, and J. S. G. Fernandez (2016) "Deep-Deep Neural Network Language Models for Predicting Mild Cognitive Impairment." in *BAI@ IJCAI*, pp. 14–20.

[72] Alvarez-Melis, D. and T. S. Jaakkola (2018) "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*.

[73] Choi, E., M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart (2016) "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, pp. 3504–3512.

[74] Ma, F., R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao (2017) "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, pp. 1903–1911.

[75] Sha, Y. and M. D. Wang (2017) "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, pp. 233–240.

[76] Ma, X. and E. Hovy (2016) "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354*.

[77] Chung, J., C. Gulcehre, K. Cho, and Y. Bengio (2014) "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*.

[78] Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014) "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*.

[79] Lin, Z., M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio (2017) "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*.

[80] Feng, C., F. Cai, H. Chen, and M. de Rijke (2018) "Attentive Encoder-based Extractive Text Summarization," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, pp. 1499–1502.

[81] Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy (2016) "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489.

[82] Wei, J. and K. Zou (2019) "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*.

[83] Yang, J. and Y. Zhang (2018) "NCRF++: An Open-source Neural Sequence Labeling Toolkit," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
URL `http://aclweb.org/anthology/P18-4013`

[84] Verma, M. and R. J. Howard (2012) "Semantic memory and language dysfunction in early Alzheimer's disease: a review," *International journal of geriatric psychiatry*, **27**(12), pp. 1209–1217.

[85] Szatloczki, G., I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski (2015) "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, **7**, p. 195.

[86] Meilan, J. J., F. Martinez-Sanchez, J. Carro, N. Carcavilla, and O. Ivanova (2018) "Voice markers of lexical access in mild cognitive impairment and Alzheimer's disease," *Current Alzheimer Research*, **15**(2), pp. 111–119.

[87] Meilán, J. J., F. Martínez-Sánchez, J. Carro, J. A. Sánchez, and E. Pérez (2012) "Acoustic markers associated with impairment in language processing in Alzheimer's disease," *The Spanish journal of psychology*, **15**(2), pp. 487–494.

[88] Meilán, J. J. G., F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana (2014) "Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, **37**(5-6), pp. 327–334.

[89] "Voice Problems and Alzheimer's Disease," `https://www.webmd.com/alzheimers/voice-speaking-problems-alzheimers#1`.

[90] Huang, Z., M. Dong, Q. Mao, and Y. Zhan (2014) "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 801–804.

[91] Keren, G. and B. Schuller (2016) "Convolutional RNN: an enhanced model for extracting features from sequential data," in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 3412–3419.

[92] Khodabakhsh, A., F. Yesil, E. Guner, and C. Demiroglu (2015) "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, **2015**(1), p. 9.

[93] Sadeghian, R., J. D. Schaffer, and S. A. Zahorian (2017) "Speech processing approach for diagnosing dementia in an early stage," .

[94] Weiner, J., C. Herff, and T. Schultz (2016) "Speech-Based Detection of Alzheimer's Disease in Conversational German." in *INTERSPEECH*, pp. 1938–1942.

[95] Roark, B., M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye (2011) "Spoken language derived measures for detecting mild cognitive impairment," *IEEE transactions on audio, speech, and language processing*, **19**(7), pp. 2081–2090.

[96] Chakraborty, R., M. Pandharipande, C. Bhat, and S. K. Kopparapu (2020) "Identification of Dementia Using Audio Biomarkers," *arXiv preprint arXiv:2002.12788*.

[97] Warnita, T., M. R. Makiuchi, N. Inoue, K. Shinoda, M. Yoshimura, M. Kitazawa, K. Funaki, Y. Eguchi, and T. Kishimoto (2020) "Speech Paralinguistic Approach for Detecting Dementia Using Gated Convolutional Neural Network," *arXiv preprint arXiv:2004.07992*.

[98] Liu, L., S. Zhao, H. Chen, and A. Wang (2020) "A new machine learning method for identifying Alzheimer's disease," *Simulation Modelling Practice and Theory*, **99**, p. 102023.

[99] He, K., X. Zhang, S. Ren, and J. Sun (2016) "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[100] Simonyan, K. and A. Zisserman (2014) "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*.

[101] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009) "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248–255.

[102] Salamon, J. and J. P. Bello (2017) "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, **24**(3), pp. 279–283.

[103] Pons, J. and X. Serra (2019) "Randomly weighted CNNs for (music) audio classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 336–340.

[104] Costa, Y. M., L. S. Oliveira, and C. N. Silla Jr (2017) "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, **52**, pp. 28–38.

[105] Dieleman, S. and B. Schrauwen (2014) "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6964–6968.

[106] Hershey, S., S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. (2017) "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, pp. 131–135.

[107] Piczak, K. J. (2015) "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018.

[108] Mesaros, A., T. Heittola, A. Eronen, and T. Virtanen (2010) "Acoustic event detection in real life recordings," in *2010 18th European Signal Processing Conference*, IEEE, pp. 1267–1271.

[109] Zhu, Z., J. H. Engel, and A. Hannun (2016) "Learning multiscale features directly from waveforms," *arXiv preprint arXiv:1603.09509*.

[110] Zeghidour, N., N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux (2018) "End-to-end speech recognition from the raw waveform," *arXiv preprint arXiv:1806.07098*.

[111] Pan, S. J. and Q. Yang (2009) "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, **22**(10), pp. 1345–1359.

[112] Salamon, J., C. Jacoby, and J. P. Bello (2014) "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044.

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

[114] Warnita, T., N. Inoue, and K. Shinoda (2018) "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data," *arXiv preprint arXiv:1803.11344*.

[115] Mizumoto, K., K. Kagaya, A. Zarebski, and G. Chowell (2020) "Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020," *Eurosurveillance*, **25**(10), p. 2000180.

[116] Arons, M. M., K. M. Hatfield, S. C. Reddy, A. Kimball, A. James, J. R. Jacobs, J. Taylor, K. Spicer, A. C. Bardossy, L. P. Oakley, et al. (2020) "Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility," *New England Journal of Medicine*.

[117] Wu, Z. and J. M. McGoogan (2020) "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention," *Jama*, **323**(13), pp. 1239–1242.

[118] Yang, P., Y. Ding, Z. Xu, R. Pu, P. Li, J. Yan, J. Liu, F. Meng, L. Huang, L. Shi, et al. (2020) "Epidemiological and clinical features of COVID-19 patients with and without pneumonia in Beijing, China," *Medrxiv*.

[119] Mahase, E. (2020), "Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate," .

[120] "WHO," `https://www.who.int/health-topics/coronavirus20.03.2020`.

[121] "The New york times," `https://www.nytimes.com/2020/08/04/us/virus-testing-delays.html`.

[122] Grasselli, G., A. Pesenti, and M. Cecconi (2020) "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response," *Jama*, **323**(16), pp. 1545–1546.

[123] Patel, A., D. B. Jernigan, et al. (2020) "Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak—United States, December 31, 2019–February 4, 2020," *Morbidity and Mortality Weekly Report*, **69**(5), p. 140.

[124] Organization, W. H. et al. (2020) "Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases, Interim guidance, 2 March 2020," .

[125] Wang, W., Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan (2020) "Detection of SARS-CoV-2 in different types of clinical specimens," *Jama*, **323**(18), pp. 1843–1844.

[126] Kucirka, L. M., S. A. Lauer, O. Laeyendecker, D. Boon, and J. Lessler (2020) "Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure," *Annals of Internal Medicine*.

[127] Guo, L., L. Ren, S. Yang, M. Xiao, D. Chang, F. Yang, C. S. Dela Cruz, Y. Wang, C. Wu, Y. Xiao, et al. (2020) "Profiling early humoral response to diagnose novel coronavirus disease (COVID-19)," *Clinical Infectious Diseases*.

[128] Xu, X., X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang, et al. (2020) "Deep learning system to screen coronavirus disease 2019 pneumonia," *arXiv preprint arXiv:2002.09334*.

[129] Ng, M.-Y., E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, et al. (2020) "Imaging profile of the COVID-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, **2**(1), p. e200034.

[130] Fang, Y., H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji (2020) "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology*, p. 200432.

[131] Sutton, D., K. Fuchs, M. D'alton, and D. Goffman (2020) "Universal screening for SARS-CoV-2 in women admitted for delivery," *New England Journal of Medicine*.

[132] Zhao, W., Z. Zhong, X. Xie, Q. Yu, and J. Liu (2020) "Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study," *American Journal of Roentgenology*, **214**(5), pp. 1072–1077.

[133] Meng, H., R. Xiong, R. He, W. Lin, B. Hao, L. Zhang, Z. Lu, X. Shen, T. Fan, W. Jiang, et al. (2020) "CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia at admission in Wuhan, China," *Journal of Infection*.

[134] Bai, H. X., B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, et al. (2020) "Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT," *Radiology*, p. 200823.

[135] Nihashi, T., T. Ishigaki, H. Satake, S. Ito, O. Kaii, Y. Mori, K. Shimamoto, H. Fukushima, K. Suzuki, H. Umakoshi, et al. (2019) "Monitoring of fatigue in radiologists during prolonged image interpretation using fNIRS," *Japanese journal of radiology*, **37**(6), pp. 437–448.

[136] Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez (2017) "A survey on deep learning in medical image analysis," *Medical image analysis*, **42**, pp. 60–88.

[137] Cohen, J. P., P. Morrison, and L. Dao (2020) "COVID-19 image data collection," *arXiv 2003.11597*.
URL `https://github.com/ieee8023/covid-chestxray-dataset`

[138] Mooney, P. (2018) "Chest x-ray images (pneumonia)," *Online], https://www. kaggle. com/paultimothymooney/chest-xray-pneumonia, tanggal akses*.

[139] Chowdhury, M. E., T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, et al. (2020) "Can AI help in screening viral and COVID-19 pneumonia?" *arXiv preprint arXiv:2003.13145*.

[140] o. M. a. I. Radiology, S. S. (2020), "COVID-19 Database," `https://www.sirm.org/category/senza-categoria/covid-19/`.

[141] Soares, E., P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe (2020) "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," *medRxiv*.

[142] Zhao, J., Y. Zhang, X. He, and P. Xie (2020) "COVID-CT-Dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:2003.13865*.

[143] Narin, A., C. Kaya, and Z. Pamuk (2020) "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*.

[144] "Chest X-Ray Images (Pneumonia)," `https://www.kaggle.com/paultimothymooney/chestxray-pneumonia`.

[145] Ozturk, T., M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya (2020) "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, p. 103792.

[146] Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers (2017) "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

[147] Wang, L. and A. Wong (2020) "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*.

[148] Yang, X., X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie (2020) "COVID-CT-Dataset: A CT Image Dataset about COVID-19," *arXiv preprint arXiv:2003.13865*.

[149] Tabik, S., A. Gómez-Ríos, J. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. Suárez, J. Luengo, M. Valero-González, et al. (2020) "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on Chest X-Ray images," *arXiv preprint arXiv:2006.01409*.

[150] Karim, M., T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, et al. (2020) "Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images," *arXiv preprint arXiv:2004.04582*.

[151] Majeed, T., R. Rashid, D. Ali, and A. Asaad (2020) "Covid-19 detection using CNN transfer learning from X-ray Images," *medRxiv*.

[152] Makris, A., I. Kontopoulos, and K. Tserpes (2020) "COVID-19 detection from chest X-Ray images using Deep Learning and Convolutional Neural Networks," *medRxiv*.

[153] Gozes, O., M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel (2020) "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv preprint arXiv:2003.05037*.

[154] Li, L., L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al. (2020) "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, p. 200905.

[155] Butt, C., J. Gill, D. Chun, and B. A. Babu (2020) "Deep learning system to screen coronavirus disease 2019 pneumonia," *Applied Intelligence*, p. 1.

[156] Apostolopoulos, I. D. and T. A. Mpesiana (2020) "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1.

[157] Zhang, J., Y. Xie, Y. Li, C. Shen, and Y. Xia (2020) "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv preprint arXiv:2003.12338*.

[158] Wang, S., B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al. (2020) "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *MedRxiv*.

[159] Zheng, C., X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang (2020) "Deep learning-based detection for COVID-19 from chest CT using weak label," *medRxiv*.

[160] Song, Y., S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, H. Zhao, Y. Jie, R. Wang, et al. (2020) "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *medRxiv*.

[161] Ardakani, A. A., A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi (2020) "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks," *Computers in Biology and Medicine*, p. 103795.

[162] Minaee, S., R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi (2020) "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *arXiv preprint arXiv:2004.09363*.

[163] Kumar, R., R. Arora, V. Bansal, V. J. Sahayasheela, H. Buckchash, J. Imran, N. Narayanan, G. N. Pandian, and B. Raman (2020) "Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers," *medRxiv*.

[164] Bullock, J., K. H. Pham, C. S. N. Lam, M. Luengo-Oroz, et al. (2020) "Mapping the landscape of artificial intelligence applications against COVID-19," *arXiv preprint arXiv:2003.11336*.

[165] Yadav, S. S. and S. M. Jadhav (2019) "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, **6**(1), p. 113.

[166] Shen, L., L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh (2019) "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, **9**(1), pp. 1–12.

[167] Guan, Q., Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang (2019) "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *Journal of Cancer*, **10**(20), p. 4876.

[168] Shorten, C. and T. M. Khoshgoftaar (2019) "A survey on image data augmentation for deep learning," *Journal of Big Data*, **6**(1), p. 60.

[169] Perez, L. and J. Wang (2017) "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*.

[170] Lei, J., J. Li, X. Li, and X. Qi (2020) "CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia," *Radiology*, **295**(1), pp. 18–18.

[171] Shi, H., X. Han, and C. Zheng (2020) "Evolution of CT manifestations in a patient recovered from 2019 novel coronavirus (2019-nCoV) pneumonia in Wuhan, China," *Radiology*, **295**(1), pp. 20–20.

[172] Song, F., N. Shi, F. Shan, Z. Zhang, J. Shen, H. Lu, Y. Ling, Y. Jiang, and Y. Shi (2020) "Emerging 2019 novel coronavirus (2019-nCoV) pneumonia," *Radiology*, **295**(1), pp. 210–217.

[173] Wang, J., M. Zhou, and F. Liu (2020) "Reasons for healthcare workers becoming infected with novel coronavirus disease 2019 (COVID-19) in China," *Journal of Hospital Infection*, **105**(1), pp. 100–101.

[174] Pan, F., T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang, et al. (2020) "Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia," *Radiology*, p. 200370.

[175] Zhou, S., Y. Wang, T. Zhu, and L. Xia (2020) "CT features of coronavirus disease 2019 (COVID-19) pneumonia in 62 patients in Wuhan, China," *American Journal of Roentgenology*, pp. 1–8.

[176] Bernheim, A., X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al. (2020) "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," *Radiology*, p. 200463.

[177] Zu, Z. Y., M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang (2020) "Coronavirus disease 2019 (COVID-19): a perspective from China," *Radiology*, p. 200490.

[178] Wong, H. Y. F., H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. Chung, et al. (2020) "Frequency and distribution of chest radiographic findings in COVID-19 positive patients," *Radiology*, p. 201160.

[179] Huang, P., T. Liu, L. Huang, H. Liu, M. Lei, W. Xu, X. Hu, J. Chen, and B. Liu (2020) "Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, **295**(1), pp. 22–23.

[180] Ai, T., Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia (2020) "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, p. 200642.

[181] Sethy, P. K. and S. K. Behera (2020) "Detection of coronavirus disease (covid-19) based on deep features," *Preprints*, **2020030300**, p. 2020.

[182] Hemdan, E. E.-D., M. A. Shouman, and M. E. Karar (2020) "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055*.

[183] Wang, Y., Y. Liu, L. Liu, X. Wang, N. Luo, and L. Li (2020) "Clinical Outcomes in 55 Patients With Severe Acute Respiratory Syndrome Coronavirus 2 Who Were Asymptomatic at Hospital Admission in Shenzhen, China," *The Journal of Infectious Diseases*, **221**(11), pp. 1770–1774.

[184] Simpson, S., F. U. Kay, S. Abbara, S. Bhalla, J. H. Chung, M. Chung, T. S. Henry, J. P. Kanne, S. Kligerman, J. P. Ko, et al. (2020) "Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the

American College of Radiology, and RSNA." *Radiology: Cardiothoracic Imaging*, **2**(2), p. e200152.

Vita

Maryam Zokaeinikoo

RESEARCH AREAS

- Machine Learning; Deep Learning; Natural Language Processing; Medical Image Analysis; Audio Processing; Data Analytics;
- Operations Research; Optimization; Stochastic Processes

EDUCATION

- **Ph.D.** in Industrial Engineering (2016-2020)
  Pennsylvania State University, State College, USA
  Advisor: Dr. Prasenjit Mitra and Dr. Soundar Kumara
  Focus Areas*:*
  Deep Learning; Machine Learning; Natural Language Processing; Big Data Analytics;
  Applied Operations Research; Artificial Intelligence in Healthcare
  Ph.D. Dissertation*:*
  "Interpretable Artificial Intelligence Methods to Detect Chronic and Infectious Diseases"

- **M.Sc.** in Industrial Engineering /Operations Research (2014-2016)
  University of Tennessee, Knoxville, USA

  Focus Areas*:*
  Machine Learning; Reinforcement Learning; Artificial Intelligence; Optimization Modeling;
  Stochastic Processes; Simulation Optimization; Healthcare Analytics

- **M.Sc.** in Industrial Engineering /Operations Research (2010-2012)
  Sharif University of Technology, Tehran, Iran
  Focus Areas*:*
  Scheduling Optimization; Mathematical Programming; Metaheuristic Algorithms; Graph Theory; Design of Experiments; Financial Engineering

- **B.Sc.** in Industrial and Systems Engineering / System Analysis (2005-2009)
  Khajeh Nasir Toosi University of Technology, Tehran, Iran