Classification and Characterization of Sleep Apnea using Machine Learning Methods on Sleep Studies

By

Linda Zhang

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

January 31, 2021

Nashville, Tennessee

Approved:

Daniel Fabbri, Ph.D.

David Kent, M.D.

Thomas Lasko, M.D., Ph.D.

Trent Rosenbloom M.D., M.P.H.

Colin Walsh, M.D., M.A.

ACKNOWLEDGEMENTS

There are a ton of people that I would like to thank for helping me along my dissertation journey, what an adventure it was! I would like to start off and give the greatest thanks to my mentor and committee chair Dr. Daniel Fabbri, who I've worked with and who has put up with me ever since I was a first year student in the program. He has taught me so much and been more lenient and understanding than I could have ever imagined. Following that, I would like to thank Dr. David Kent, the clinician on my committee who basically birthed the topic of my thesis, and who has worked closely with me for these last few years. I want to give a great big thanks to the rest of my committee: Dr. Tom Lasko, Dr. Trent Rosenbloom and Dr. Colin Walsh, who have encouraged me and given me many ideas throughout the progress of my dissertation.

I want to thank the entire Department of Biomedical Informatics at Vanderbilt – it's an absolutely fantastic community, and I've enjoyed my time here and learned so much. In particular, I want to give a great big thanks to everyone who has been a director of graduate studies in our department, Drs. Cindy Gadd, Gretchen Jackson, and Kim Unertl, who organize the program and have made it possible for me to be a part of it. I'd also like to thank the National Library of Medicine for funding the grant that enabled me to do so. I'd like to give a HUGE thank you to Rischelle Jenkins, the best program manager EVER. Any questions I've ever had, any problems I ran into, I could always turn to Rischelle and she would figure it out. And the biggest thank you to the graduate students of the DBMI program – in particular Alex Cheng, Lina Sulieman, Bryan Steitz and Matt Lenert, for being great friends and helping me with the last minute reorganization of my presentation. To all the students, it's been a blast learning, working and eating (☺) alongside all of you absolutely brilliant people.

I want to thank and say I love you to my family, who has always believed in me and supported me in all my endeavors. And lastly, I'd like to thank my boyfriend, Thomas Schlegel, who followed me headfirst into this adventure, and has been taking care of me all these years. His support has made it possible for me to pursue and finish this work. Here's to the start of another one!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Sleep disorders and their diagnosis

Obstructive Sleep Apnea (OSA) Syndrome is a sleep disorder in which breathing rapidly starts and stops during sleep. It has been independently linked to multiple health conditions including an increased risk of hypertension, diabetes, cardiovascular disease, stroke risk, and overall mortality.[1-4] It affects approximately 6% of women and 13% of men in the United States.[5] OSA is diagnosed using an overnight sleep study measuring multiple high-resolution physiologic signals called a polysomnogram (PSG).

Overnight polysomnography is central to the diagnosis and management of many sleep disorders. A patient will come in at nighttime, and a technologist will apply monitors to measure activity in the body related to sleep (Figure 1). Some of this activity includes:

- Wires with small cup electrodes attached to the scalp with a conductive paste to monitor brain activity.
- Wire electrodes taped to the face to show muscle activity.
- 2 elastic belts around the chest and stomach to measure breathing effort.
- A nasal cannula and small heat monitor to measure all breathing activity.
- A wire electrode on each leg to measure body movement/muscle activity.
- A monitor taped to a finger to detect oxygen levels during the study.
- 2-3 lead EKG monitors to show heart rate and rhythm.
- A small snore mic applied to the throat to detect snoring.

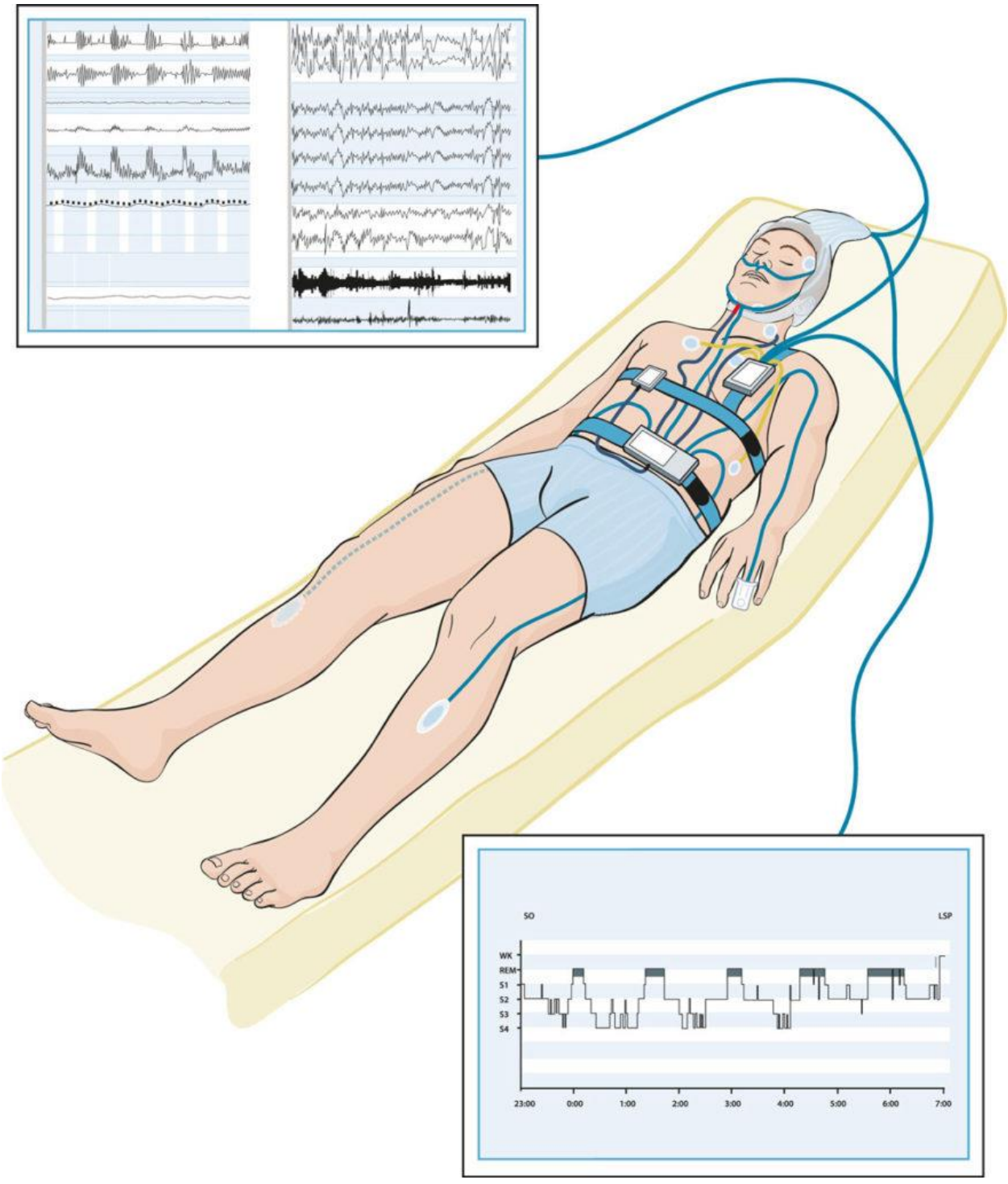

Figure 1. Sleep study setup[6]

After the study is completed, a sleep technologist will then score the study, labeling diagnostic events such as apneas and hypopneas, and sleep staging. Apneas and hypopneas are

the cessation and slowing of breathing. An apnea is technically defined by the American Acadamy of Sleep Medicine as the absence of airflow for > 10 seconds, while the latest hypopnea definition is a > 3% oxyhemoglobin desaturation or event-related arousal.[7] Scoring these events allows for the calculation of the Apnea-Hypopnea Index (AHI), used to diagnose sleep apnea.

The value of AHI is used to determine the severity categorization of sleep disordered breathing (SDB). These classifications are as follows: mild (AHI = 5-14), moderate (AHI = 15-30), severe (AHI > 30). AHI is defined by apnea and hypopnea occurrence and can vary widely with the use of different hypopnea scoring criteria. This metric can be difficult to automatically score not only because of nuances like the hypopnea scoring criteria, but because of variability in human event detection. Despite the variability with AHI, it has been found to correlate with cardiovascular risk and overall mortality. [1-4]

Sleep staging is the categorization of sleep into REM, non-REM and wake stages, and is essential for evaluating the quality of sleep and diagnosing its disorders. Sleep staging is evaluated using the brain activity (EEG), muscular activity (EMG) and eye activity (EOG) channels. The clinical standard for PSG sleep staging requires visual inspection of the data by trained sleep technicians and physicians. Sleep technicians look for specific waveforms in the EEG channel that indicate sleep stage, some of which include alpha activity, theta activity, vertex sharp waves, spindles, K complexes and slow waves. The EMG channel is used to distinguish between sleep and wake (muscular activity is lost in sleep), and the EOG channel is used to distinguish the presence of eye blinks (wake).[7]

The sleep staging problem

Staging historically followed the Rechtschaffen and Kales criteria until the American Academy of Sleep Medicine (AASM) published updated criteria in 2007.[8,9] The AASM rules divide sleep into five stages: Wake, Non-Rapid Eye Movement stages 1, 2, and 3 (N1, N2, and N3), and Rapid Eye Movement (REM). PSG scoring is a labor-intensive process that requires up to two hours for a sleep technologist to complete.[10] In addition, inter-rater and intra-rater

reliability of PSG staging and event scoring is also known to suffer from considerable variability.[11-20]

In the past, significant effort has been invested in developing computer-assistive or automated staging technologies, but they have struggled to achieve human-level performance.[10, 21-32] The problem is difficult because of multiple channels of timeseries data with specific rules and event triggers are used in order to define the stage of sleep. In order for a staging system to have clinical utility it should be at least as accurate and reliable as a trained human scorer. Therefore, a practical non-inferiority threshold for staging algorithms is an overall agreement of 82.0% (Cohen's kappa = 0.76), which is the overall inter-rater agreement between trained scorers at eight European centers using the 2007 AASM PSG scoring rules. We tackle the automatic sleep stage scoring problem in our first Aim.

The standard metric: Apnea-Hypopnea Index

The gold standard for the diagnosis of OSA is AHI, defined as the average number of airflow limitations (hypopneas) and cessations (apneas) observed per hour of sleep during a PSG.[7] The large volumes of data generated in these studies has been analyzed the same way for decades: a human interpreter looks for visual patterns in the data. This labor-intensive task is difficult and multiple studies have confirmed significant interrater variability in sleep and respiratory event scoring.[33-36] We explore the respiratory event scoring problem in our second Aim.

Furthermore, the definitions of an apnea and hypopnea have changed in recent years as the sleep medicine community continues to debate the fundamentals of clinically meaningful events during sleep. The cumulative effect of these subjective definitions and interpretations is reflected in recently published work that shows the overall predictive value of AHI for complications from OSA is low.[37-39] Furthermore, published research suggests that the definition of clinically significant events may need to change depending on the health outcome of interest. Currently, there is a critical need to personalize the field of sleep medicine by developing computational tools for PSGs that objectively discover clinically significant OSA phenotypes that may be too complex for human interpreters to reliably recognize.

Ambiguity in the hypopnea definition

The definition of a hypopnea has changed since its original definition in 1999. It has gone through three different definitions: 1) a > 4% oxyhemoglobin desaturation; 2) a > 3% oxyhemoglobin desaturation; 3) a > 3% oxyhemoglobin desaturation or event-related arousal. It has been found that as the definition of a hypopnea has changed, AHI has increased, resulting in a more severe SDB classification on average.[40] Because of the uncertainty and changing definition of hypopnea, there has been controversy surrounding the use of the AHI as the single disease scoring metric.

In addition to the various available definitions of hypopnea, there is a problem in measuring hypopnea severity because it is defined using a threshold. A problem with thresholds is that any events that exceed the threshold are represented as the equivalent, e.g. 4% desaturation is considered equivalent to a 10% desaturation. Because of this, we lose information about the severity of the hypopnea. Similarly, a 100-second long apnea or hypopnea contributes the same information as a 10-second long event to a patient's AHI. In general, the AHI is a useful metric that defines OSA, but does not measure its severity particularly well. Despite the problems with AHI, it has been found to correlate with cardiovascular risk and overall mortality. [1-4]

A growing body of evidence suggests that a common definition of AHI may be insufficient for predicting the many complications of OSA. This may be due, in part, to the fact that repetitive airway obstructions can variably precipitate other physiologic events such as tachycardia, blood oxygen desaturation, and neurologic arousal. While these related risks are likely important for quantifying associated disease risk, they are only coarsely captured by defining threshold criteria for apneas and hypopneas, if at all. In the Sleep Heart Health Study (SHHS), severe OSA has been shown to double the hazard ratio of cardiovascular disease risk in certain subpopulations, but in other groups the prognostic power of AHI appears more limited.[2,10,39]

AHI severity categories do not correlate well with symptom burden or comorbidity outcomes.[39] This has led to other signal patterns in PSG data being investigated for their

relationships with outcomes of interest. Two-percent oxygen desaturations have been found to predict insulin resistance, while 4% oxygen desaturations and REM-predominant OSA predict hypertension.[41-43] The time spent with an oxygen saturation below 90% predicts platelet aggregation, while an increased arousal index predicts memory impairment.[44-46] Definitions of disease burden may therefore need to change depending on a given patient's history, the outcome of interest, and the specific pathophysiologic mechanisms underpinning their OSA.[38] We tackle this problem of defining phenotypes in our third Aim.

Machine learning in sleep

Machine learning has started to become more widely used in the sleep medicine field in the past couple years. Problems such as apnea/hypopnea event detection or sleep staging are examples of some of the areas that machine learning has been used to explore and solve.[20-30] Polysomnogram data, the main diagnostic data available for sleep, is in the form of multichannel signal data. This type of data is both temporal and spatial in nature, and lends itself well to deep learning techniques, which have been widely used in other areas of signal processing (music, weather forecasting).[31,32] We apply a combination of traditional machine learning and deep learning techniques to polysomnogram data to try and solve the beforementioned problems.

Dissertation Aims

In this dissertation, we leverage a variety of machine learning methods in combination with clinical knowledge to build models that help to characterize, measure and phenotype sleep apnea. Deep learning can be used to take advantage of the structure of polysomnogram data and learn to recognize predefined patterns and definitions, as well as search for new patterns. We use these techniques to first solve supervised problems, automating sleep staging and apnea/hypopnea event detection, and then build models to predict comorbid outcomes such as cardiac arrest.

Aim 1: Predict sleep staging from polysomnograms using deep learning.

The first aim describes the development of an automated sleep staging model using polysomnogram data as input into deep neural network models. Chapter 3 presents the process and reasoning behind the model creation and presents results on the performance and model generalizability. We published this work in the journal of SLEEP medicine, under a title of "Automated Sleep Stage Scoring of the Sleep Heart Health Study Using Deep Neural Networks". Automating sleep stage scoring is an important step in the automation of polysomnogram scoring, as well as developing a model that can reduce variability in sleep stage scoring.

Aim 2: Predict AHI from polysomnograms using deep learning.

The second aim describes the development of an automated apnea/hypopnea event detection model and its use to predict AHI. Chapter 4 details the creation of this model as well as the effects of differing representations of the temporal data. This work explores deeper into the different ways to represent snippets of signal from a long overnight sleep study and how those representations affect prediction of events. Automating respiratory event detection is another important step in the automation of polysomnogram scoring and the methods used in this problem illustrate different ways that timeseries data can be represented.

Aim 3: Engineer PSG-derived features that predict sleep apnea-associated cardiac outcomes.

The third aim describes the development of a series of models that predict cardiac outcomes associated with sleep apnea. The base model is built to predict cardiovascular outcomes with multiple clinician-defined features (e.g. AHI, demographics, clinical data, etc.). We then developed engineered features derived from polysomnograms and literature in sleep apnea. These features significantly improve prediction of associated cardiac outcomes and show that there are sleep-derived patterns and features within polysomnograms that can be used to describe those outcomes. This work is important in starting to develop and categorize subtypes of OSA, allowing for more personalized treatment.

The work in this thesis is important in the advancement of the field of sleep medicine. There is a wealth of data from polysomnography that is perfect for modeling and pattern discovery. With the development of data-driven models, the time-consuming and variable tasks such as polysomnogram scoring can be automated. With the development of new data-driven phenotypes, we can develop better descriptors of sleep apnea, resulting in more personalized patient treatment plans.

CHAPTER 2

BACKGROUND

This dissertation explores sleep apnea and how we diagnose it using polysomnography. To do this, we must first understand what the signal data looks like and how we use machine learning to process and learn from it. This chapter describes the data and methods used, followed by previous research in this domain.

Polysomnography and signal data

Polysomnograms are an overnight record of a number of physiological signals of the sleeping patient. In order to gain information from these signals, sleep technicians look manually look through the study and score events and sleep staging (Figure 2). These events and sleep stages are used to diagnose sleep apnea.[7]

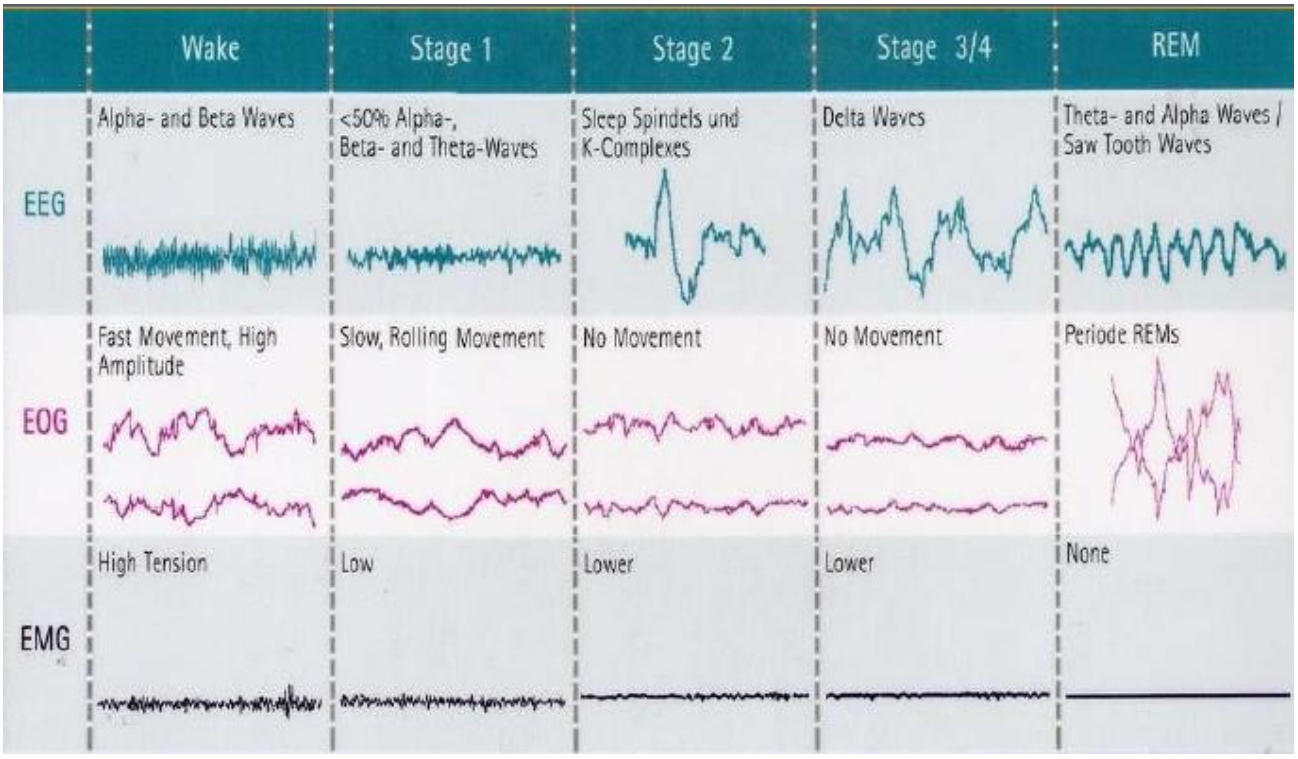

Figure 2. Sleep stage summary. The EEG channel describes brain activity, EOG channel describes eye activity, and EMG channel describes muscle movement.[47]

Signal data can be processed and analyzed or used as input features for other models. Generally, signal data are described or analyzed by measuring the frequency of the signals. These frequencies are associated with events that are occurring within the channel. Transforming the data from the signal domain into the frequency domain using Fourier or Wavelet transformations can summarize the frequencies of the signal and is helpful in analysis. These summaries are often used as inputs into models, because they contain a condensed version of the important information from the signal.

Machine learning

Machine learning uses probabilistic mathematics to recognize data patterns by inspecting many examples rather than by following explicit programming. There are two types of machine learning: supervised and unsupervised. Supervised machine learning uses labeled data to train models that predict an output (label) for the associated input. The goal for these models is to be able to predict the label for new, unseen input. To do this, supervised machine learning models learn patterns and relationships between the input and output labels. Unsupervised machine learning uses unlabeled data and attempts to find order, patterns or structure between the inputs. Its goal is more exploratory; it attempts to cluster the inputs in order to summarize, explain or identify noteworthy patterns.[48]

Machine learning analyses are not based on a priori clinical definitions, and therefore have the potential to learned existing and unrecognized phenotypes. Specifically for sleep data, we choose to use deep neural networks, a type of model which mimics the neuronal connections of the human brain. These systems have only become practical within the last decade as computational power has increased to the point where models can be trained on large datasets.[48] They rapidly analyze large volumes of data through layers of interconnected processing units to find patterns that many be too complex for human interpreters to recognize. Deep learning systems have been shown to scale well for large datasets and build highly accurate classifiers out of noisy, real-world datasets.

Deep learning and neural networks

A standard neural network consists of a number of simple connected processors called neurons that mathematically transform an input signal into an output. The relative strength, or weight, of each neuron is iteratively adjusted during model training to maximize the accuracy between the network output and the expected value. Deep neural networks have many layers of neurons, where the output of one layer provides the input to the next layer, enabling discovery of nonlinear and hierarchical relationships within the data.[49] An example of this hierarchical learning can be seen in the facial recognition problem (Figure 3).

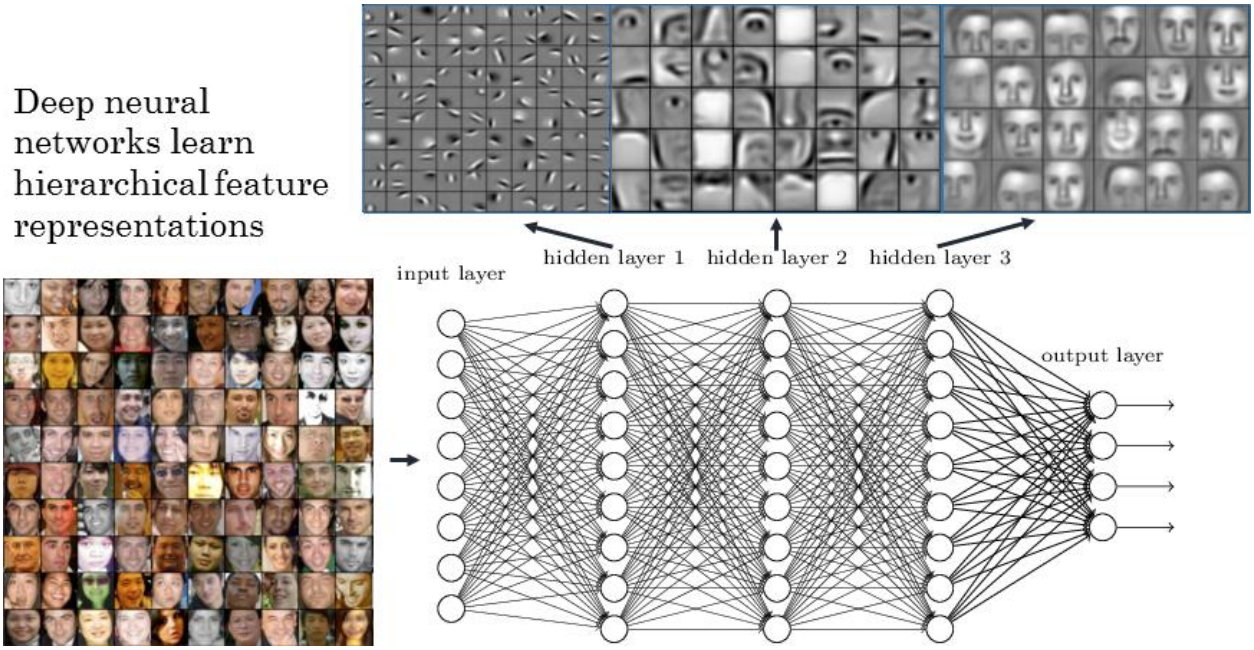

Figure 3. Example neural network used to solve facial recognition problem.[49]

Specialized neural networks use convolutional and recurrent layers to take advantage of naturally existing structures within data. Convolutional neural networks emphasize patterns in close spatial proximity and are well-suited to problems in the image classification and recognition space.[48] Recurrent neural networks function well with information contained in sequences such as natural language, where the next word or character depends on the immediately preceding data.[50]

Convolutional layers differ from the regular fully-connected network layers in that their connections are limited to other network structures in spatial proximity, emphasizing locality or

co-occurrence within the data. They have been used in various image and signal applications, ranging from image classification to brain-mapping using EEG signals.[51-52] (Figure 4).

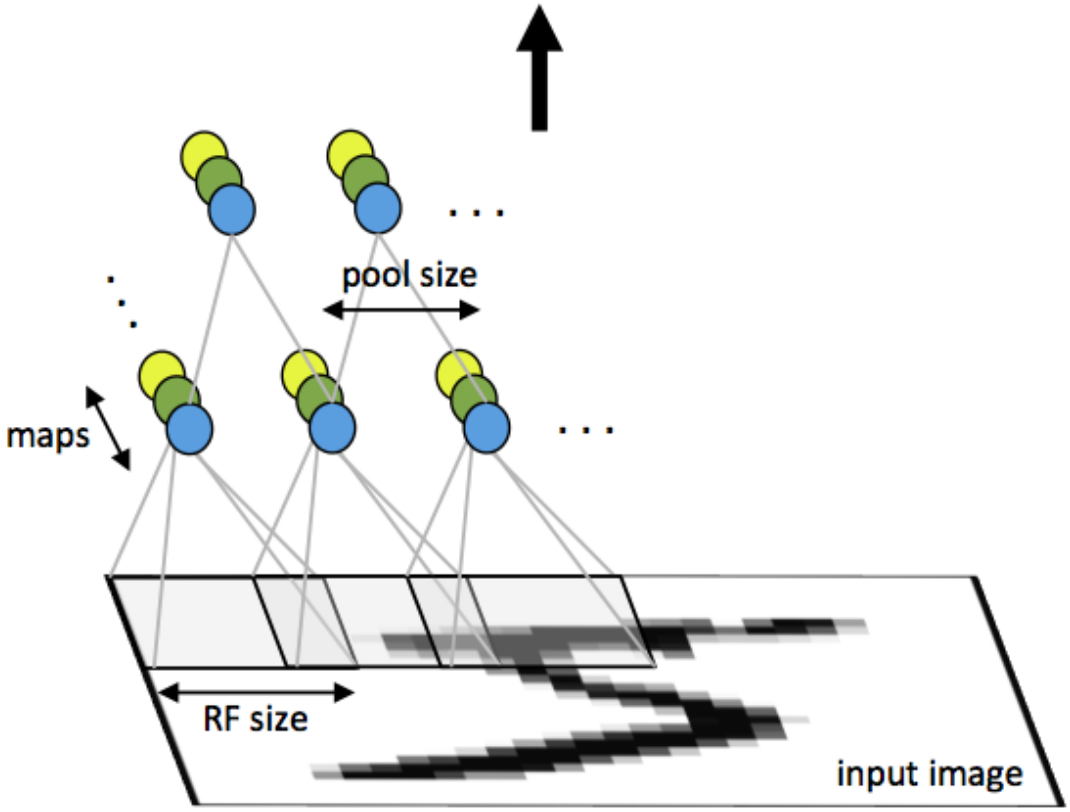

Figure 4. Convolutional neural network used to solve image recognition.[53]

Recurrent layers feature unidirectional feedback mechanisms from downstream network structures. This construction takes advantage of continuous sequences, where memories of past events contribute to future decision making. Recurrent neural networks have yielded dramatic improvements in speech recognition and wind forecasting.[50,54] The activating features of these networks can be visualized through different methods including visualizing layer activations and maximally-activating inputs, thus allowing for human interpretation of the learned model (Figure 5).

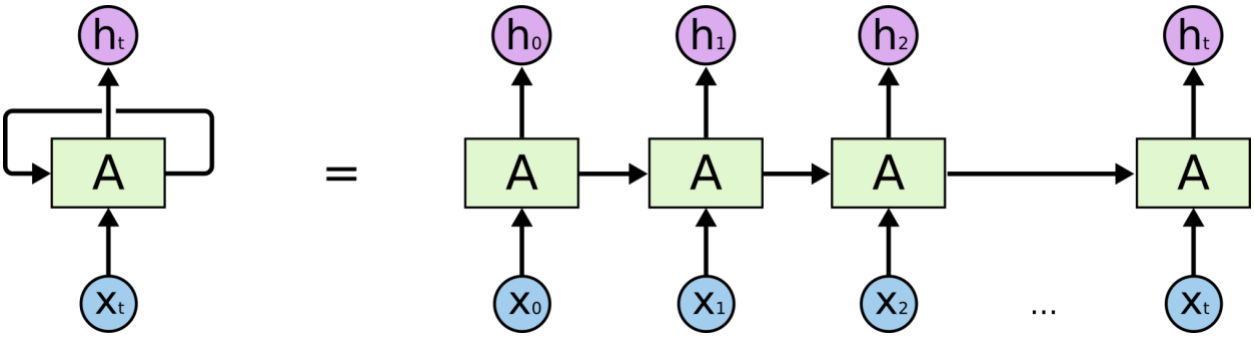

Figure 5. Recurrent neural network structure.[55]

Machine learning in sleep medicine

Machine learning is a field of computer science where classifiers discover novel patterns within a dataset without the traditional explicit encoding of all rules. Because PSG data are complex, different machine learning methods for detecting sleep stages have been trialed over the last twenty years. Published models have utilized hand-tuned feature extraction techniques such as spectral power, time domain analysis, and time-frequency domain (wavelet) analysis.[56-59] Other systems employ fuzzy logic, support vector machines, hidden Markov models, or artificial neural networks.[60-69] Most of these systems do not achieve human-level inter-rater agreement or are tested against a small set of preselected, high-quality PSGs that do not reflect realistic testing environments. Few have been validated against large clinical datasets. In recent years, deep neural networks have rapidly found favor for signal analysis. They have proven to be remarkably robust in developing classifier systems for noisy, "real-world" datasets: the type of data represented by PSGs. PSGs are well suited for convolutional and recurrent processing methodologies as they consist of spatially- and temporally-related signal data.

The increase in available computing power and publicly available PSG datasets over the last several years has brought the era of Big Data and machine learning to sleep medicine and made deep neural network processing of PSGs feasible.[70,71] Successful development of a reliable and accurate automated scoring system using machine learning will ease the burden of PSG scoring and will reduce sleep staging inter-rater variability that affects Sleep Medicine research and clinical practice.

Machine learning in sleep staging

Staging historically followed the Rechtschaffen and Kales criteria until the American Academy of Sleep Medicine (AASM) published updated criteria in 2007.[8,9] The AASM rules divide sleep into five stages: Wake, Non-Rapid Eye Movement stages 1, 2, and 3 (N1, N2, and N3), and Rapid Eye Movement (REM). PSG scoring is a labor-intensive process that requires up to two hours for a sleep technologist to complete.[10] Inter-rater and intra-rater reliability of PSG staging and event scoring is also known to suffer from considerable variability.[11-20]

Significant effort has been invested in developing computer-assistive or automated staging technologies, but they have struggled to achieve human-level performance.[10, 21-32] These studies use a wide variety of signal processing and machine learning methods. In order for a staging system to have clinical utility it should be at least as accurate and reliable as a trained human scorer. Therefore, a practical non-inferiority threshold for staging algorithms is an overall agreement of 82.0% (Cohen's kappa = 0.76), which is the overall inter-rater agreement between trained scorers at eight European centers using the 2007 AASM PSG scoring rules.

Machine learning in apnea-hypopnea prediction

There has been much work on apnea-hypopnea event prediction in the past. Koley & Dey (2013) automatically detect apnea and hypopnea events from a single oronasal airflow channel using support vector machines.[72] Huang et al. (2017) detect apnea and hypopnea events based on the respiratory nasal airflow signal and the oximetry signal.[73] They use a sliding window and short time slice method to eliminate systematic and sporadic noise of the airflow signal for improving the detection precision. Choi et al. (2018) use convolutional neural networks in a sliding window method to detect events.[74] Yu et al. (2019) use random forests on hand-engineered features for detection.[75] A number of these methods perform fairly well on the event detection problem, using a wide variety of techniques and methods of data representation. The problem is handled differently by almost every study in this problem space, because there is no defined period of time in which an apnea or hypopnea event can occur; the event can occur at any point, and only has the requirement of lasting a minimum of 10 seconds.

Because of this, we decide to explore different ways of representing the signal data in different window periods of time, and its effect on apnea/hypopnea event detection and the resulting AHI.

Phenotyping sleep apnea

Different methodologies have been applied in recent years in an attempt to better elucidate factors related to OSA pathogenesis. A top-down, structured approach to phenotyping based on PSG measurements of multiple clinician-defined physiologic traits was described in 2013 by Eckert and colleagues as the PALM Scale (airway critical closure pressure [P_{crit}], arousal threshold, loop gain, muscle responsiveness of the upper airway).[77] These features are measured by actively alternating airway pressures in patients wearing continuous positive airway pressure tolerance (CPAP) and then measuring physiologic responses according to pre-specified criteria. Patients showed significant trait heterogeneity, with over half displaying abnormal non-anatomic features such as hypersensitive loop gain, decreased arousal thresholds, and inadequate pharyngeal dilator muscle responsiveness. Further work has shown that patient populations display different combinations of PALM traits.[78,79] Other research confirms that phenotypes can have differential responses to targeted therapies,[80-85] lending greater weight to the concept of undiscovered OSA subtypes with different underlying disease risks and responses to therapy. There is a need for more work in this area, which we dive into with Aim 3.

CHAPTER 3

AUTOMATED SLEEP STAGE SCORING OF THE SLEEP HEART HEALTH STUDY USING DEEP NEURAL NETWORKS

Statement of significance

Sleep staging is an important part of evaluating overnight polysomnograms. Sleep stages are scored by technicians and physicians based on visual examination of neurophysiologic signal patterns. This process is labor intensive and suffers from variability between scorers. In this study, large amounts of publicly available PSG data were used to train a sleep staging classifier. Sleep staging classification by the model achieved better agreement than human agreement in literature. Generalizability of the model to other unseen datasets from different public projects is also demonstrated.

Introduction

Overnight polysomnography (PSG) is central to the diagnosis and management of many sleep disorders. The clinical standard for PSG sleep staging requires visual inspection of the data by trained sleep technicians and physicians. Staging historically followed the Rechtschaffen and Kales criteria until the American Academy of Sleep Medicine (AASM) published updated criteria in 2007.[7,8] The AASM rules divide sleep into five stages: Wake, Non-Rapid Eye Movement stages 1, 2, and 3 (N1, N2, and N3), and Rapid Eye Movement (REM). PSG scoring is a labor-intensive process that requires up to two hours for a sleep technologist to complete.[3] Inter-rater and intra-rater reliability of PSG staging and event scoring is also known to suffer from considerable variability.[11-20]

Significant effort has been invested in developing computer-assistive or automated staging technologies, but they have struggled to achieve human-level performance.[10, 21-32] In order for a staging system to have clinical utility it should be at least as accurate and reliable as a trained human scorer. Therefore, a practical non-inferiority threshold for staging algorithms is an overall agreement of 82.0% (Cohen's kappa = 0.76), which is the overall inter-rater

agreement between trained scorers at eight European centers using the 2007 AASM PSG scoring rules.[12]

Machine learning is a field of computer science where classifiers discover novel patterns within a dataset without the traditional explicit encoding of all rules. Because PSG data are complex, different machine learning methods for detecting sleep stages have been trialed over the last twenty years. Published models have utilized hand-tuned feature extraction techniques such as spectral power, time domain analysis, and time-frequency domain (wavelet) analysis.[56-59] Other systems employ fuzzy logic, support vector machines, hidden Markov models, or artificial neural networks.[60-69] Most of these systems do not achieve human-level inter-rater agreement or are tested against a small set of preselected, high-quality PSGs that do not reflect realistic testing environments. Few have been validated against large clinical datasets. In recent years, deep neural networks have rapidly found favor for signal analysis. They have proven to be remarkably robust in developing classifier systems for noisy, "real-world" datasets: the type of data represented by PSGs.

A standard neural network consists of a number of simple connected processors called neurons that mathematically transform an input signal into an output. The relative strength, or weight, of each neuron is iteratively adjusted during model training to maximize the accuracy between the network output and the expected value. Deep neural networks have many layers of neurons, where the output of one layer provides the input to the next layer, enabling discovery of nonlinear and hierarchical relationships within the data. Convolutional neural networks emphasize patterns in close spatial proximity and are well-suited to problems in the image classification and recognition space.[51,52] Recurrent neural networks function well with information contained in sequences such as natural language, where the next word or character depends on the immediately preceding data.[50] PSGs are well suited for convolutional and recurrent processing methodologies as they consist of spatially- and temporally-related signal data. For example, a k-complex may signal onset of N2 sleep, even though subsequent EMG data may be low-amplitude mixed-frequency data visually identical to N1.

The increase in available computing power and publicly available PSG datasets over the last several years has brought the era of Big Data and machine learning to sleep medicine and

made deep neural network processing of PSGs feasible.[70,71] Successful development of a reliable and accurate automated scoring system using machine learning will ease the burden of PSG scoring and will reduce sleep staging inter-rater variability that affects Sleep Medicine research and clinical practice.

Methods

This study was designed as a retrospective analysis of PSG data collected through several multicenter cohort studies made available through the National Sleep Research Resource (NSRR).[71,86,87] The study design was approved by the Vanderbilt University Medical Center Institutional Review Board (#171186) and data access was approved by the NSRR.

Study datasets

A deep neural network model was trained and tested on 5,804 Type II PSGs from multiple centers containing patients with and without sleep-disordered breathing collected for the Sleep Heart Health Study (SHHS; Table 1).[71,86,87]

Table 1. Sleep Heart Health Study summary statistics

Category	Mean	Median	Min, Max
Age	63.1	63	[39, 90]
Body Mass Index	28.2	27.5	[18, 50]
Apnea Hypopnea Index	17.9	13.2	[0, 161.8]
Sleep Time (minutes)	359.8	367.0	[34.5, 519]

Two additional unrelated datasets available through the NSRR were used to test the generalizability of the model: the Study of Osteoporotic Fractures (SOF) and the Osteoporotic Fractures in Men study (MrOS; Table 2).

Table 2. Summary of datasets used in study

Dataset	Polysomnography studies (n)	Study population	W (%)	N1 (%)	N2 (%)	N3 (%)	R (%)
SHHS	5,793	Adults aged 40 and older	28.8%	3.7%	40.9%	12.6%	13.9%
MrOS	2,907	Men 65 years or older	46.1%	3.7%	33.9%	5.8%	10.6%
SOF	461	Women ages 65-89 years	41.9%	2.9%	32.5%	11.9%	10.7%

SHHS = Sleep Heart Health Study; SOF = Study of Osteoporotic Fractures; MrOS = Osteoporotic Fractures in Men study

Polysomnography Data

All PSG files were downloaded in the European Data Format (EDF) which contained the raw time series data of physiologic signals from each PSG as well as human scored sleep stages and apneic events. For the training phase, 5,213 PSGs were randomly selected from the SHHS dataset, providing 42,560 hours of sleep data in 5,107,200 30-second epochs. PSGs in all three datasets were recorded as Type II unattended home studies previously scored using modified Rechtschaffen & Kales (RK) criteria. [71,86,87] PSG signal data and sleep stage labeling (Wake, N1, N2, N3, N4 or REM) were extracted from each study cohort. RK stages 3 and 4 were combined into a single stage N3 label to more closely align with modern AASM scoring conventions and to aid comparison with previously published literature. The model was trained and tuned using 90% of the SHHS visit 1 data (5,213 patients). A 10% holdout set (580 patients) was taken and set aside to validate the model.

Input data and feature selection

Signal data from the electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) PSG channels were extracted for model analysis. The Type II PSGs across all three cohorts were recorded using a single central (C3) EEG channel. Sampling rates across data channels from SOF and MrOS were down- or up-sampled as indicated to match corresponding baseline data sampling rates from SHHS.

Two different methodologies for feature representation were tested. In the first method, raw PSG signal data was provided directly as input to the network in per-epoch units and tested under various model architectures. In the second method, short time Fourier transforms were used to generate a spectrogram for each epoch and then provided to the model as the input. Spectrograms were generated using 2-second sub-epochs formed by a Tukey window with 25% of the window inside the tapered cosine region (Figure 6). Signal normalization and filter signal preprocessing methods (median, FIR, IIR filters) were tested to evaluate the impact of noise and artifact reduction.

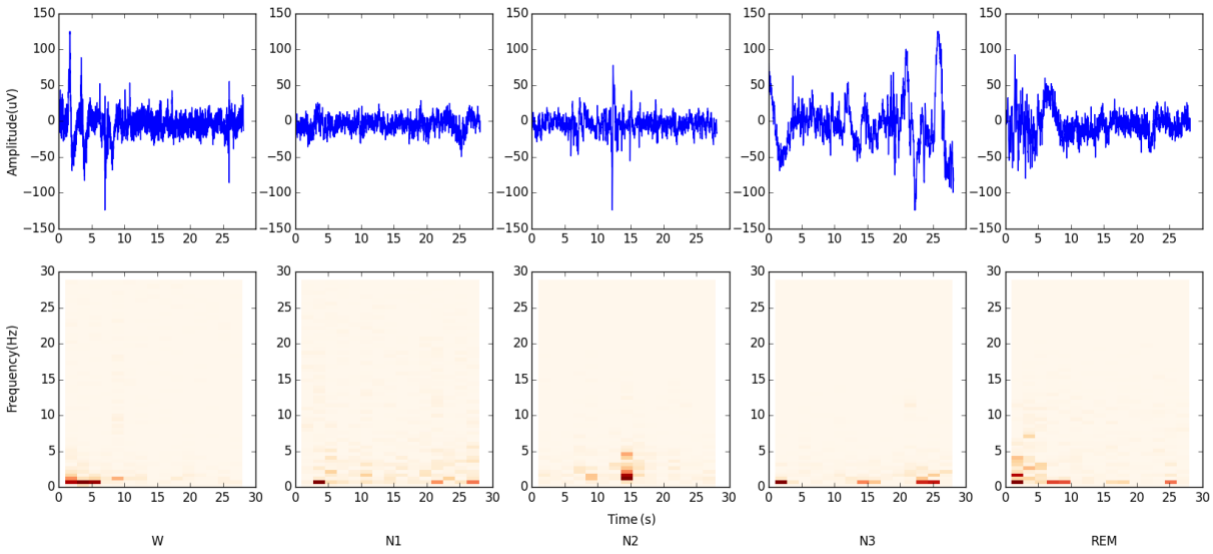

Figure 6. Representative raw data sample from each sleep stage with associated spectrogram.

All data preprocessing was performed using the signal module in the python package SciPy and scikit-learn. Model development was performed using Keras on a TensorFlow backend.

Model architecture

Convolutional and recurrent network layers were utilized to take advantage of the temporally-linked, sequential construction of PSG data. Convolutional layers were generated to evaluate the co-occurrence of signal patterns within one-dimensional PSG data channels or co-

occurrence of frequencies within single spectrograms. Recurrent layers were designed to take advantage of the temporal relationships in the data such as epochs of equivalent stage occurring in sequence. The deep neural network combined recurrent and convolutional structures to evaluate input spectrograms generated from the raw data (Figure 7). Multiple combinations of dense, convolutional, and recurrent layers were tested against the training set in the network architecture (Appendix A).

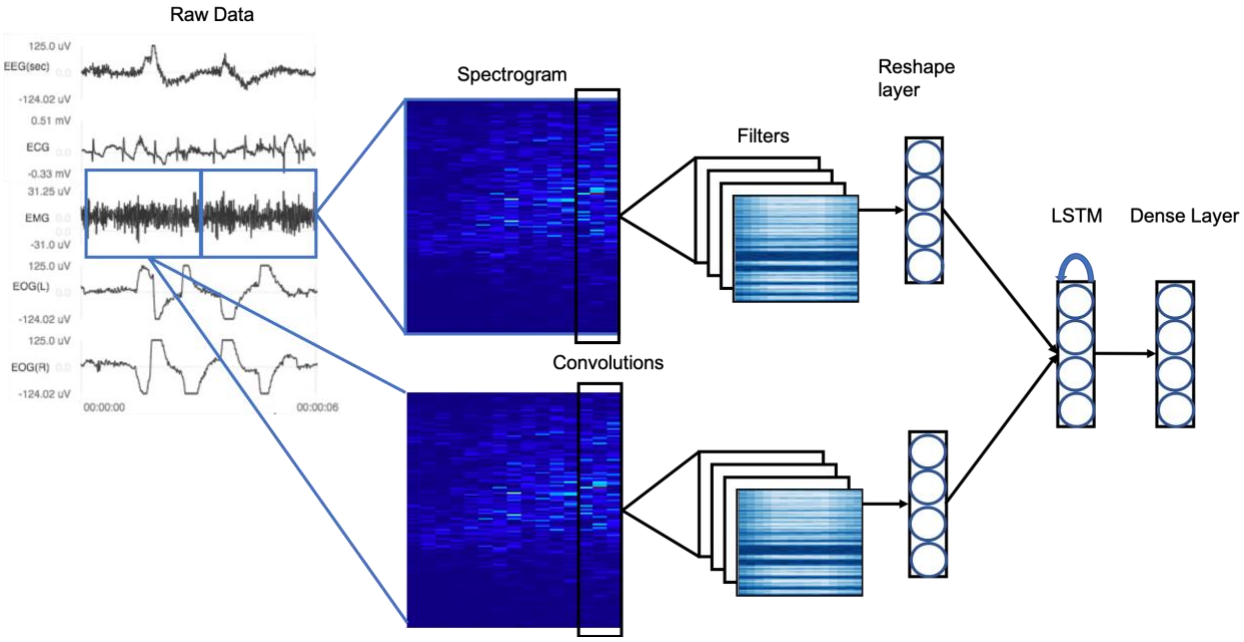

Figure 7. Simplified example model architecture for one data channel. LSTM = long short-term memory layer.

Model tuning

Deep neural networks contain tunable hyperparameters (i.e. number of layers, number of units in each layer, number of filters in convolutional layers, etc). A set of parameter search spaces were defined for each hyperparameter, and the best combination of hyperparameters were found using the python package hyperopt with a random search algorithm for parameter tuning.[88] Multiple hyperparameter configurations were evaluated using the training set.

Model evaluation

Model performance was evaluated with accuracy, F1-score, and Cohen's Kappa. Weighted and unweighted accuracy and F1-score were calculated to assess the effect of sleep stage class imbalances in the data. Weighted accuracy was calculated as the average of the per-class stage accuracies. Because the "ground truth" comparators are human-tagged PSG events with their own level of inter-rater reliability, model agreement was also assessed using inter-rater agreement statistics (Cohen's Kappa). Transition epoch F1-scores were calculated as scoring agreement is known to degrade during transition from one stage of sleep to another. Transition stages account for approximately 0.5% of the data, but were evaluated as they potentially convey physiologically relevant information.

Transfer learning

Generalizability was assessed using the SOF and MrOS datasets. These studies were conducted in different environments with various types of acquisition hardware and on different patient populations than SHHS.

Model performance was additionally evaluated on subsets of the SHHS population with mild, moderate and severe obstructive sleep apnea (OSA) to demonstrate model transferability between patients with different degrees of sleep-disordered breathing. A separate model was also trained and tested on only severe patients to demonstrate validity even when restricted to a subset of studied patients.

Results

The optimal sleep staging model's architecture consisted of a combination of separate networks for each signal channel. Spectrograms of each channel were fed into convolutional layers that examined the proximal relationships of the frequencies in time as well as recurrent layers that examined the sequential relationships of epochs (Table 3). The subnetworks for each signal channel were combined into two dense layers feeding into a final softmax output layer used to generate discrete stage predictions for each epoch.

Table 3. Base model architecture per data channel

Layer	Layer Type	Size	Output Size
Input			(2, 1, 129, 16)
C1	Convolutional	(32,64,3)	(2, 32, 66, 14)
C2	Convolutional	(32,64,3)	(2, 32, 2, 12)
P1	Max Pooling	(2,2)	(2, 32, 1, 6)
R1	Reshape		(2, 192)
L1	Long short-term memory	(256)	256
D1	Dense	(512)	512

Model testing

The SHHS dataset was split into a 90% training and 10% holdout set. The training set was further split into training and validation sets, which were used to train the model, select the optimal deep learning architecture (Appendix A), and tune the model hyperparameters (Appendix B). Model training required approximately 48 hours on an Nvidia GTX Titan X GPU. A learning curve plateauing around 1,000,000 training epochs demonstrated that the dataset was sufficiently large (Figure 8). Testing on the holdout set required approximately 30 minutes.

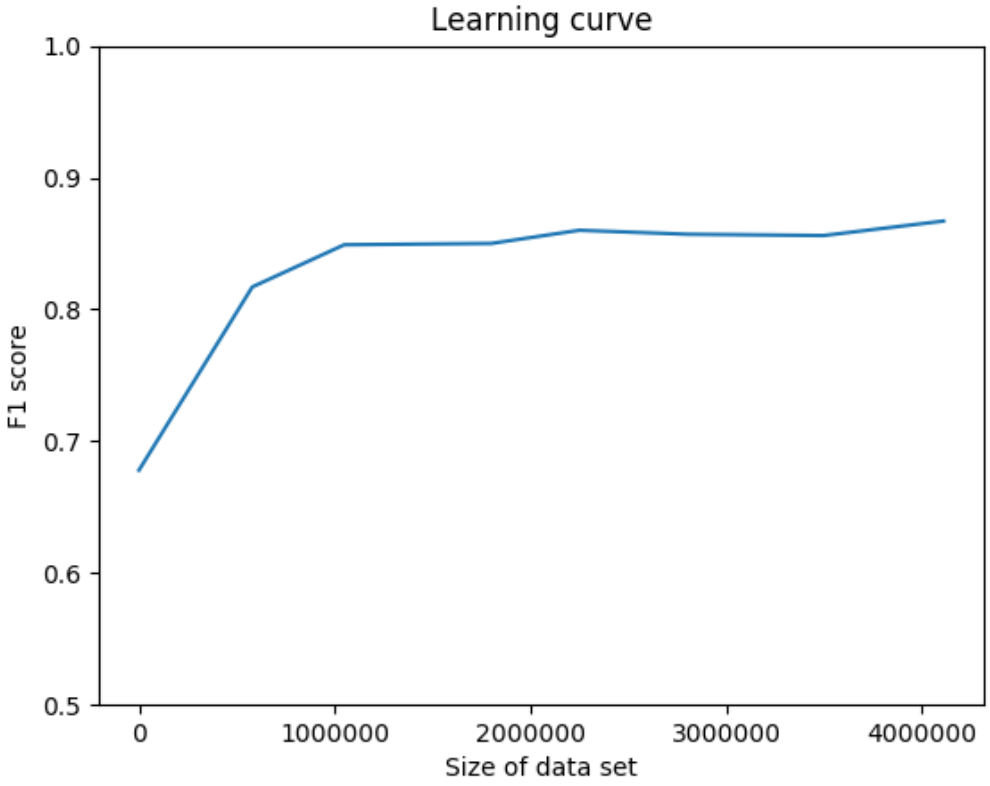

Figure 8. The deep neural network model learning curve begins to plateau at approximately 1,000,000

Model evaluation

Signal preprocessing methods were tested on the raw input signal. No significant improvement in accuracy or F1-scores were found using normalization or filters, so signal preprocessing was not used in the final pipeline (data not shown). Multiple model architectures were tested on the SHHS dataset. The first model was a simple baseline Markov Chain that predicted the next stage based on overall stage transition probabilities measured directly from SHHS. Because stages commonly occur in long chains with relatively rare transitions, this model has a high F1-score, but low transition F1-score. Following this baseline model, a convolutional neural network (CNN) was tested against raw PSG data, followed by separate CNN and long short-term memory (LSTM) models on the spectrogram data, and finally a combination of CNN + LSTM, which yielded the best performance (Figure 9).

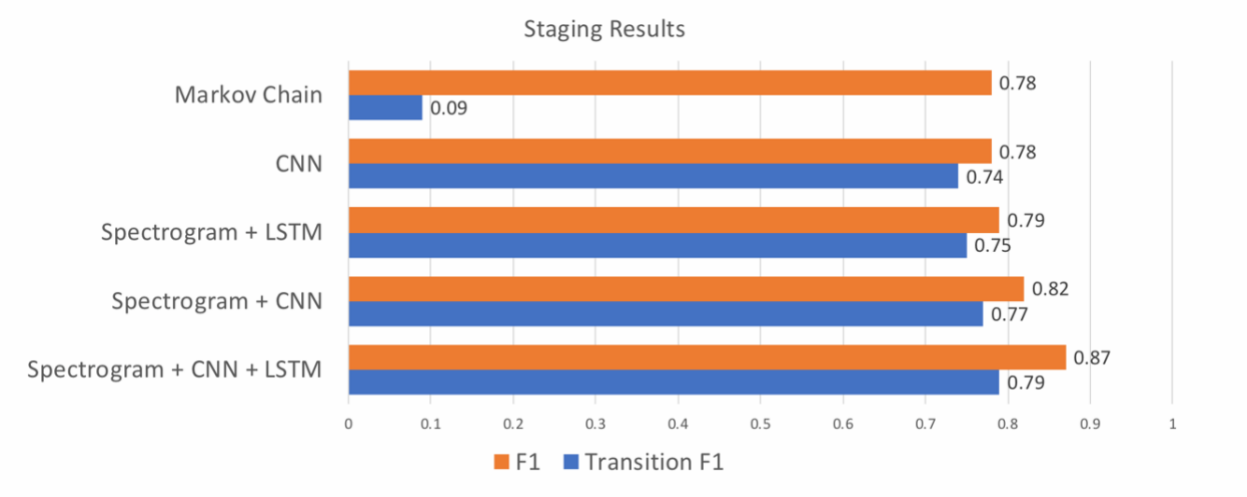

Figure 9. Model performance under various architectures against the SHHS dataset. CNN = convolutional neural network; LSTM = long short-term memory

The optimal neural network model was composed of spectrograms in the input layer feeding into CNN layers and an LSTM layer to achieve a weighted F1-score of 0.87 and Cohen's Unweighted Kappa of K = 0.82, higher than that of human agreement found in literature (K = 0.76).

A confusion matrix was generated for model performance against all tested epochs (Figure 10) as well as transition epochs (Figure 11).

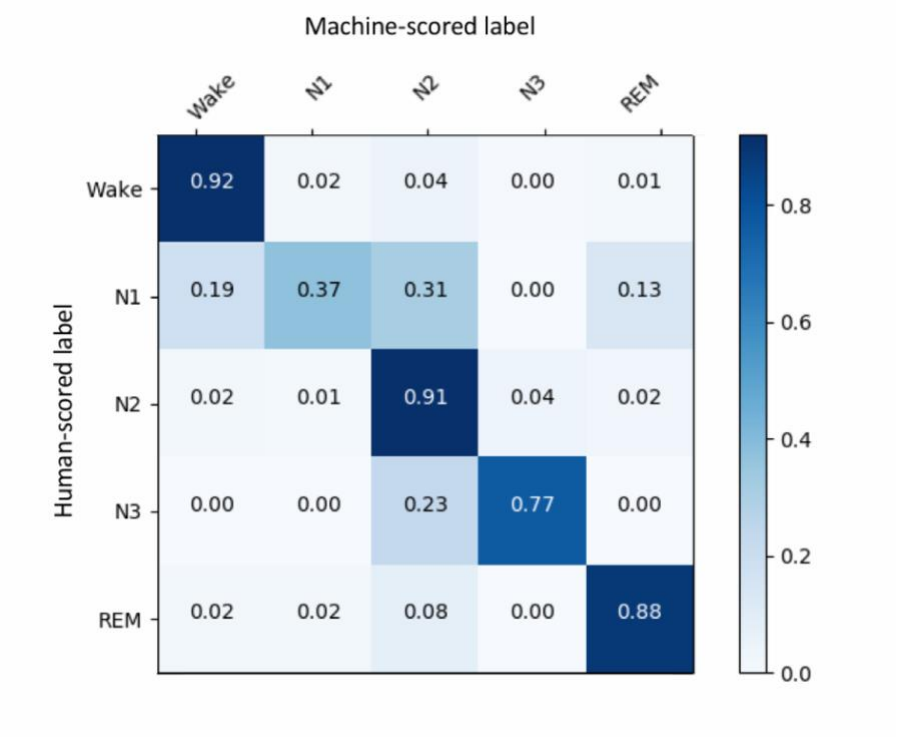

Figure 10. Confusion matrix for all epochs.

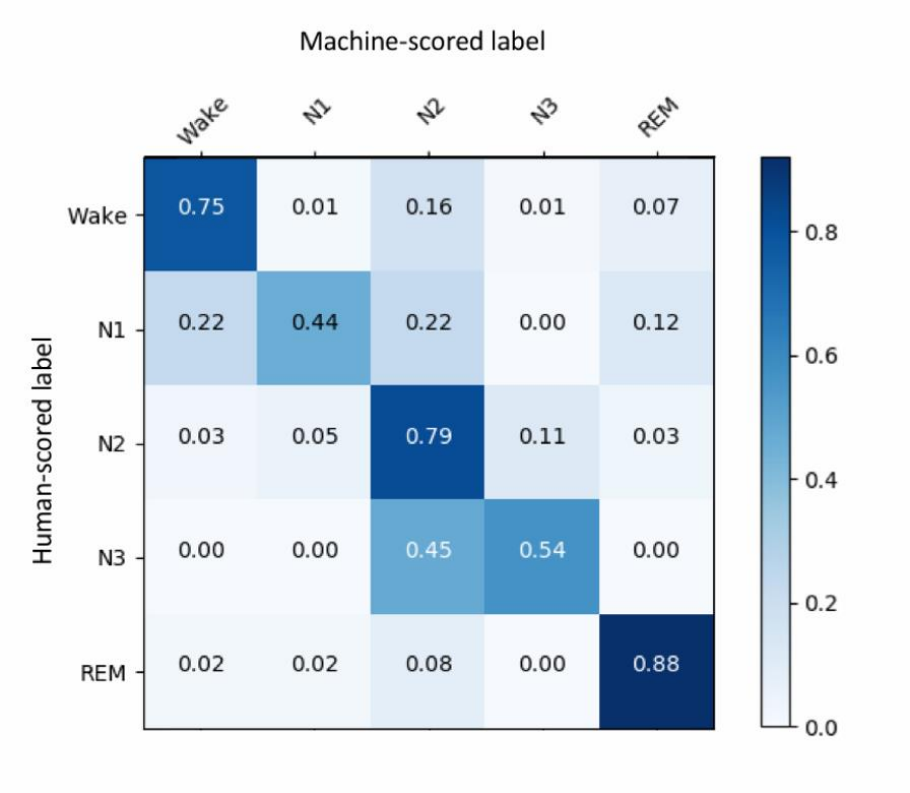

Figure 11. Transition epoch confusion matrix.

When considering all epochs, the model scored Wake, N1, N2, N3, and REM stages correctly 92%, 37%, 91%, 77%, and 88% of the time, respectively. During transition epochs correct staging was scored for Wake, N1, N2, N3, and REM 75%, 44%, 79%, 54%, and 88% of the time, respectively. Table 4 compares staging accuracy of this model to others published in the literature using the class imbalances present in the underlying dataset. Table 5 permits comparison to other models in the literature that used methods to balance the classes such that all classes contribute equally in model training. Figure 12 demonstrates agreement between a trained scorer and the automated scoring model in one example PSG hypnogram.

Table 4. Performance of class imbalanced model compared to other studies

Study	Sample size (studies)	Evaluation split	W Accuracy	N1 Accuracy	N2 Accuracy	N3 Accuracy	REM Accuracy	Overall Accuracy	Balanced Accuracy	Cohen's Kappa
Biswal et al. [89]	10,000	Train – validation – test	84.5%	56.2%%	88.4%	85.4%	92%	85.8%	81.3	0.795
Sors et al. [90]	5793	Training – validation – test	91%	35%	89%	85%	86%	87%	77.2%	0.81
Sharma et al. [91]	100	10-fold-CV	95%	17%	76%	57%	36%	91.7%	56.5%	N/A
Proposed model	5793	Train – validation - test	92%	37%	91%	77%	88%	87%	77%	0.82

Table 5. Performance of class balanced model compared to other studies

Study	Sample size (studies)	Evaluation split	W Accuracy	N1 Accuracy	N2 Accuracy	N3 Accuracy	REM Accuracy	Overall Accuracy	Balanced Accuracy	F1 Score	Cohen's Kappa
Supratak et al. [92]	62	31-fold cross validation	87.3%	43.5%	90.5%	77.1%	80.9%	86.2%	75.9%	0.817	0.8
Tsinalis et al. [93]	40	20-fold cross validation	70%	60%	73%	91%	74%	82%	74%	0.81	N/A
Chambon et al. [94]	62	5-fold cross validation	85%	52%	77%	91%	83%	79%	77.6%	0.72	N/A
Proposed model	5793	Train – validation - test	91%	46%	89%	77%	88%	86%	78%	0.81	0.82

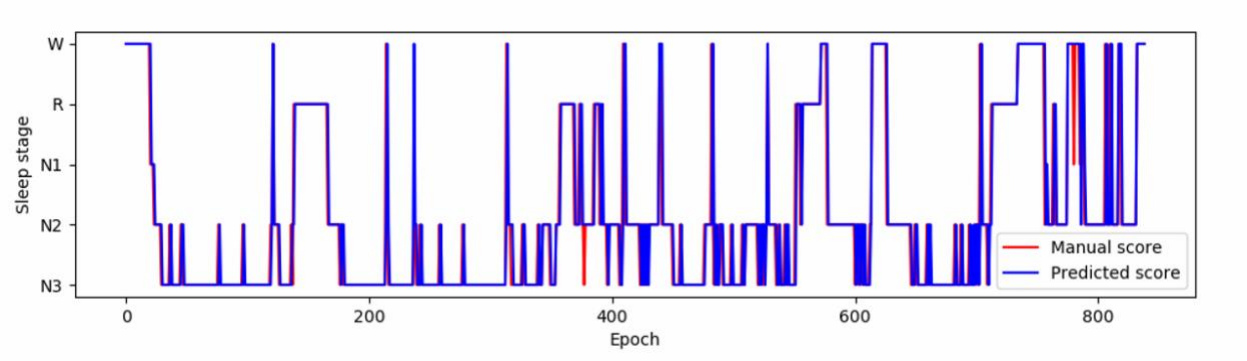

Figure 12. Example output hypnogram of a PSG scored by the model overlaid on the human manual scoring.

Performance on cohorts with and without sleep-disordered breathing

The model performs similarly on subsets of the holdout set with different apnea severity (Table 6). A model trained and tested on severe OSA patients only achieved an unweighted F1 score of 0.846, similar to the model trained on heterogeneous data.

Table 6. SHHS model performance on patient subgroups of varying obstructive sleep apnea severity2

Testing cohort	F1	Epochs (N)
All	0.872	621794
Normal (AHI < 5)	0.871	132742
Mild (5 < AHI < 15)	0.864	262426
Moderate (15 < AHI < 30)	0.853	168074
Severe (AHI > 30)	0.841	58552

SHHS = Sleep Heart Health Study; AHI = Apnea-Hypopnea Index

Transfer learning

After training on SHHS data, model generalizability was tested against two additional NSRR datasets. The microvolt mean and standard deviation of each included data channel was significantly different between studies, suggesting different signal architectures between datasets (Table 7).

Table 7. Mean and standard deviation of the channels for each dataset

Channel	SHHS	MrOS	SOF
EEG (uV)	-0.39 ± 30.31	2.5 ± 38.08 *	-8.87 ± 43.02 *
EMG (uV)	0.54 ± 9.68	-1.06 ± 58.49 *	10.05 ± 34.47 *
EOG(L) (uV)	-3.57 ± 30.60	-12.5 ± 49.28 *	-9.81 ± 35.60 *
EOG(R) (uV)	-4.19 ± 31.36	3.33 ± 50.81 *	5.32 ± 41.37 *

* indicates significant difference from SHHS data at $p < 0.05$. SHHS = Sleep Heart Health Study; SOF = Study of Osteoporotic Fractures; MrOS = Osteoporotic Fractures in Men study

F1-score and Cohen's Kappa scores on the MrOS and SOF datasets demonstrated moderate to strong inter-rater agreement between the model and trained scorers depending on the selected testing data and achieved high performance in the balance of precision and recall on sleep staging (Table 8).

Table 8. Generalizability of the SHHS model to novel datasets.

Model	F1-score (weighted)	Cohen's Kappa
Training data: SHHS Testing data: SHHS	0.87	0.82
Training data: SHHS Testing data: MrOS	0.79	0.70
Training data: SHHS Testing data: SOF	0.77	0.68
Training data: MrOS Testing data: SHHS	0.69	0.56
Training data: SOF Testing data: SHHS	0.66	0.53

SHHS = Sleep Heart Health Study; MrOS = Osteoporotic Fractures in Men study; SOF = Study of Osteoporotic Fractures

Discussion

The deep learning model presented here automatically predicts sleep stage with moderate to strong agreement compared with expert human scorers across multiple datasets. The optimal model utilized input consisting of spectrograms derived from the EEG, EMG, and EOG channels passed to a deep learning architecture with convolutional and recurrent layers. A learning curve demonstrated that sufficient data was available to train the model well. The model performs comparably or better than other models reported in literature and, when tested against studies with structure similar to the underlying training dataset, meets or exceeds the accepted benchmark of K = 0.76 between trained human scorers.

Spectrograms are used to represent the data provided to the model in the form of dimensionally-reduced input that retains important information for sleep stage classification. The Fourier transforms used to generate spectrograms organized PSG data into component frequencies more easily compared across different platforms than raw signal data, which contains baseline signal noise and variation due to different recording environments and hardware. Spectrogram construction also aided network throughput as the volume of input data were reduced without significant loss of key signal information.

Preprocessing raw signal data for noise and artifact reduction did not significantly impact classification results in preliminary testing. Prior performance analyses have demonstrated that deep learning models become more robust when trained on noisy data,[95] and we suspect that training on unfiltered data may be advantageous for accuracy and transferability when testing across clinical datasets.

PSGs have significant class imbalances between stage types due to the natural asymmetric distribution of sleep stages. The SHHS dataset is no exception, with large differences in representation between several of the stages. Accounting for class imbalances by over-representing minority classes (such as N1) can improve single class accuracy, but often at the expense of larger classes. For instance, in SHHS N1 is only 3.7% of the dataset, whereas N2 is 40.9%. The model presented here scored 31% of N1 and 91% of N2 epochs correctly with an overall accuracy of 87% when the native class imbalances are not adjusted.

When N1 was oversampled to balance class representation, accuracy of N1 increased to 45% at the expense of other stages, such as N2, which decreased to an accuracy of 88%. Class balancing decreased overall model scoring accuracy to 86%. Class imbalances also complicate comparison of performance metrics between published models. We believe that preserving native class imbalances best represents how the model would perform in a production setting. However, performance metrics for models trained on natural as well as balanced class distributions are provided in order to facilitate comparison with previously published models (Tables 4 and 5).

Accuracy in N1 scoring is worse than other sleep stages for this model, consistent with other published models.[90-94] This may be an artifact of PSG scoring rules, which allow for low-amplitude mixed pattern EEG signals identical to N1 to be scored as N2 if the preceding stage was also scored as N2. These rules, along with the large class imbalances between N1 and N2, likely compromise N1 accuracy.

Other issues may complicate scoring accuracy, such as patient movement artifacts contaminating W and N1 stages. Unlike many other published works, this model was not trained on a curated set of high-quality PSGs and contains studies partially contaminated by signal and motion artifacts. Contaminated epochs scored by humans theoretically contain enough signal information that they should be of value in training a machine learning algorithm that will be exposed to similar data in a production environment. The inclusion of this more ambiguous data may create systemic difficulties in scoring W and N1 in the same way that it would degrade inter-rater agreement between human scorers. To this point, Younes et al recently found an intra-class correlation coefficient of 0.69 (range: 0.30 – 0.86) in N1 scoring, suggesting only poor to moderate agreement between trained human scorers.[20]

This model presented in this work has several strengths. It meets or exceeds performance of other published works. A large and diverse training dataset increased transferability, demonstrated across several other large datasets. Significant differences existed in mean microvolt channel levels across the tested datasets (Table 6), suggesting significant underlying differences in dataset structure due to differences in recording hardware, environment, study populations, or other variables. Despite these differences, the model presented here could be

trained on one dataset and still perform with moderate to strong agreement on other datasets (MrOS F1 = 0.78, K = 0.68 and SOF F1 = 0.68, K = 0.55). The model also performed similarly on cohorts composed of subjects with varying degrees of sleep disordered breathing, with F1-scores ranging from 0.841 to 0.872, suggesting that sleep-disordered breathing does not significantly impact sleep stage classification patterns for the model. In comparison, a model trained only on patients with severe sleep apnea and tested on the same cohort performs only slightly better than one trained on all patients, demonstrating model transferability between different disease populations. Taken together, the transferability properties illustrated here suggest that automated deep learning classifiers have the potential for use in different clinical sleep laboratory environments without complete retraining on local data.

Few other studies test models on PSGs collected from a variety of recording environments and hardware platforms. Patanik et al.[97] did so, demonstrating generalizability by testing against two novel datasets with inter-rater agreement of K = 0.740 and K = 0.597. 54 However, their reported outcomes (accuracy) were obtained from model training data instead of separate holdout data, limiting inner-dataset comparability to the work presented here. The Kappa values are also not directly comparable to our inter-rater agreement of K = 0.70 and K = 0.56. The datasets in Patanik et al. were acquired using the same framework and pipeline, while the external test datasets presented here were acquired on a variety of different hardware platforms that were then down- or up-sampled to match SHHS dataset frequencies. Both studies demonstrate comparable performance on external datasets that the models were not trained on, demonstrating transferability.

This work is not without limitations. The datasets examined here are composed of Type II PSGs recorded in subject home environments with a limited, single EEG channel montage. Generalizability to more common Type I or Type III PSGs could not be evaluated; however, we suspect that training the model with additional EEG signals available in Type I PSGs would likely yield performance improvements from additional channel data. Retraining the model with additional channels while maintaining input from previously evaluated channels would be expected to improve performance, as deep neural networks generally perform better as more data is available.[98] Comparison with more limited montage datasets, such as consumer

wearables using actigraphy and heart rate monitoring, is limited by the lack of large, publicly available datasets. In addition, accuracy outcomes may differ between AASM sleep staging criteria and RK staging criteria.

In conclusion, this work suggests that automated PSG scoring systems can rapidly annotate PSG files with inter-rater agreement rivaling that of trained human scorers. Future work will require institutions and interested stakeholders to make available large libraries of high-quality datasets using modern scoring criteria in order for data scientists to develop robust, generalizable scoring models.

CHAPTER 4

APNEA AND HYPOPNEA EVENT DETECTION FOR APNEA-HYPOPNEA INDEX PREDICTION

Introduction

Overnight polysomnography (PSG) is central to the diagnosis and management of sleep disorders. Obstructive Sleep Apnea (OSA) Syndrome is a sleep disorder in which the upper airway collapses during sleep, obstructing ventilation. It has been independently linked to multiple health conditions including an increased risk of hypertension, diabetes, cardiovascular disease, stroke risk, and overall mortality.[1-4] OSA is diagnosed using an overnight sleep study measuring multiple high-resolution physiologic signals called a polysomnogram (PSG). A sleep technologist or clinician label diagnostic events such as apneas (near cessation of ventilation) and hypopneas (reduction in ventilation) to calculate the average number of respiratory events per hour: the Apnea-Hypopnea Index (AHI). An apnea is defined by the American Academy of Sleep Medicine (AASM) as decrease in baseline airflow amplitude by at least 90% for 10 or more seconds.[7] The preferred 2012 definition for hypopnea is:[7]

a. The peak signal excursions drop by ≥30% of pre-event baseline

b. The duration of the ≥30% drop in signal excursion is ≥10 seconds

c. There is a ≥3% oxygen desaturation from pre-event baseline and/or the event is associated with an arousal

The AHI is used to determine the severity of sleep-disordered breathing. Classification categories are as follows: normal (AHI < 5), mild (AHI = 5-14), moderate (AHI = 15-30) and severe (AHI > 30). PSG labeling, or scoring, is a labor-intensive process that requires up to two hours for a sleep technologist to complete.[10] In addition, inter-rater and intra-rater reliability of event scoring is also known to suffer from considerable variability.[13,14,18,19]. Despite its reported variability, AHI has been found to correlate with cardiovascular risk and overall mortality.[100-103]

A wide range of methodologies have been previously employed to create automated event detection models. The earliest models include rule-based models, followed by machine learning and deep learning models. Koley & Dey (2013) developed a combined SVM and rule-

based system from 36 patients to achieve an overall accuracy of 96.5% for event detection.[72] Lee et al. (2016) created a rule-based algorithm that achieved an 86.4% positive predictive value (PPV) and 84.5% sensitivity in AHI prediction[73], while Huang et al. (2017) created a rule-based algorithm using sliding windows and short time slices to test on cohorts of 30 and 28 patients, achieving PPVs of 97.6% and 92.3% and sensitivities of 95.7% and 92.3% in predicting AHI, respectively.[74] Choi et al. (2018) trained convolutional neural networks on 179 recordings for an overall accuracy of 96.6%[75], while Yu et al. (2019) developed a random forest model trained on 24 subjects that achieved an 88.3% accuracy.[76]

PSG data is composed of high-resolution biophysiological signal data including measurements of airflow, thoracic and abdominal movement (belts), brain activity via electroencephalography (EEG), and oxygen saturation that are well-suited to analysis with deep learning methods because of the inherent spatial and temporal relationships. Deep learning methods have been successfully applied to automating sleep stage labeling in PSGs.[104] A standard neural network consists of simple connected processors called neurons that mathematically transform an input signal into an output. Deep neural networks have many layers of interconnected neurons enabling discovery of nonlinear and hierarchical relationships within the data. Convolutional networks emphasize patterns in close spatial proximity and are well-suited to problems in the image classification and recognition space.[51,52] Moreover, compared to rule-based methods where humans define the rules for event prediction, deep learning models discover underlying patterns in data without explicit programming. The use of deep learning models has resulted in more robust and better performing models in many medical fields.[105]

While previous studies have found generally high prediction performance, they are trained and tested on small datasets, limiting conclusions that can be made regarding model robustness and generalizability. In addition, direct comparisons between rule-based and machine learning models on the same datasets have not been performed. This rule-based comparison is important in particular, because definite rules for apnea and hypopnea have been defined previously and are considered standard-of-care. In this study, a deep learning model was trained and tested on a large dataset and directly compared to a rule-based model.

The performance of the event prediction on calculating AHI was evaluated. Successful development of a reliable and accurate automated scoring system using machine learning will ease the labor burden of PSG analysis and will reduce inter-rater variability in event scoring affecting Sleep Medicine research and clinical practice.

Methods

This study was designed as a retrospective analysis of PSG data collected through several multicenter cohort studies available through the National Sleep Research Resource (NSRR). [71,86,87] Study design was approved by the Vanderbilt University Medical Center Institutional Review Board (#171186) and data access was approved by the NSRR.

Study datasets

A deep neural network model was trained on 5,804 Type II PSGs from multiple centers containing patients with and without sleep-disordered breathing collected for the Sleep Heart Health Study (SHHS). The study contains two sets of visits, with the first and larger set used here (SHHS 1; Table 9).

Table 9. SHHS 1 summary statistics

Category	Mean	Median	Min, Max
Age	63.1	63	[39, 90]
Body Mass Index	28.2	27.5	[18, 50]
Apnea Hypopnea Index	17.9	13.2	[0, 161.8]
Sleep Time (minutes)	359.8	367.0	[34.5, 519]

Polysomnography Data

All PSG files were downloaded in the European Data Format containing the raw time series data of physiologic signals from each PSG as well as human-scored sleep stages and apneic events. For the training phase, 5,213 PSGs were randomly selected from the SHHS dataset. The model was trained and validated using 90% of the SHHS visit 1 data (5,213 patients) and a 10% holdout set (580 patients) was taken and set aside to test the model.

Input data and representation

Signal data from the airflow, abdominal and thoracic belt channels were used for the apnea model. Data from the airflow, saturated oxygen channels the EEG channel were tested as input for the hypopnea model. The raw signal data were normalized before segmentation and model input. All data preprocessing was performed using the python packages SciPy and scikit-learn.[106] Model development was performed using Pytorch.[107]

Apnea and hypopnea events are defined as having a decrease in airflow for a minimum of 10 seconds, but there is no limit on the event length. Due to the fact that event length is unrestricted, windows, or segments of set length, of the PSG data are selected as input for classification. Differing window sizes and length of overlap between windows for representation of input data were tested (Appendix C). Window sizes ranging from 5 to 60 seconds were tested with overlap between the windows on a range from 0 to 5 seconds.

A single input window was considered apnea- or hypopnea-positive if at least 5 seconds of a segment less than 30 seconds belonged to a labeled event, or at least 10 seconds for a segment of length 30 seconds or greater. These thresholds represent the minimum segment processing (5 seconds) and event definition (10 seconds) lengths, and therefore require considerable overlap in order to register as a positive detection event.

Model architecture

Multiple combinations of dense, convolutional, and recurrent layers were tested against the training set in the network architecture (Appendix D). Convolutional layers were used on multiple channels of data with the same frequency and on single channels of data. Recurrent layers were tested only on non-overlapping segments of input data, which represent sequences. The best architecture for each the apnea and hypopnea models were selected based on per-segment accuracy metrics, and the best performing model used convolutional layers on normalized signal data (Figure 13, Figure 14).

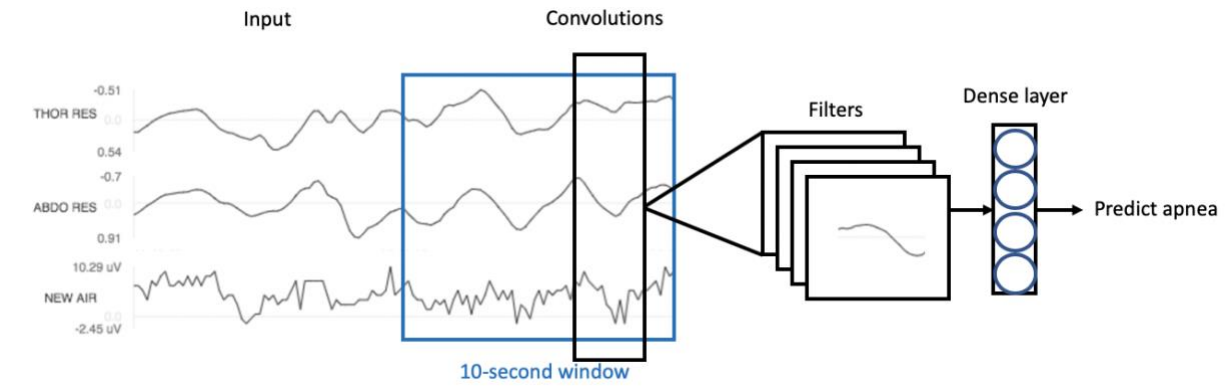

Figure 13. Simplified apnea model architecture

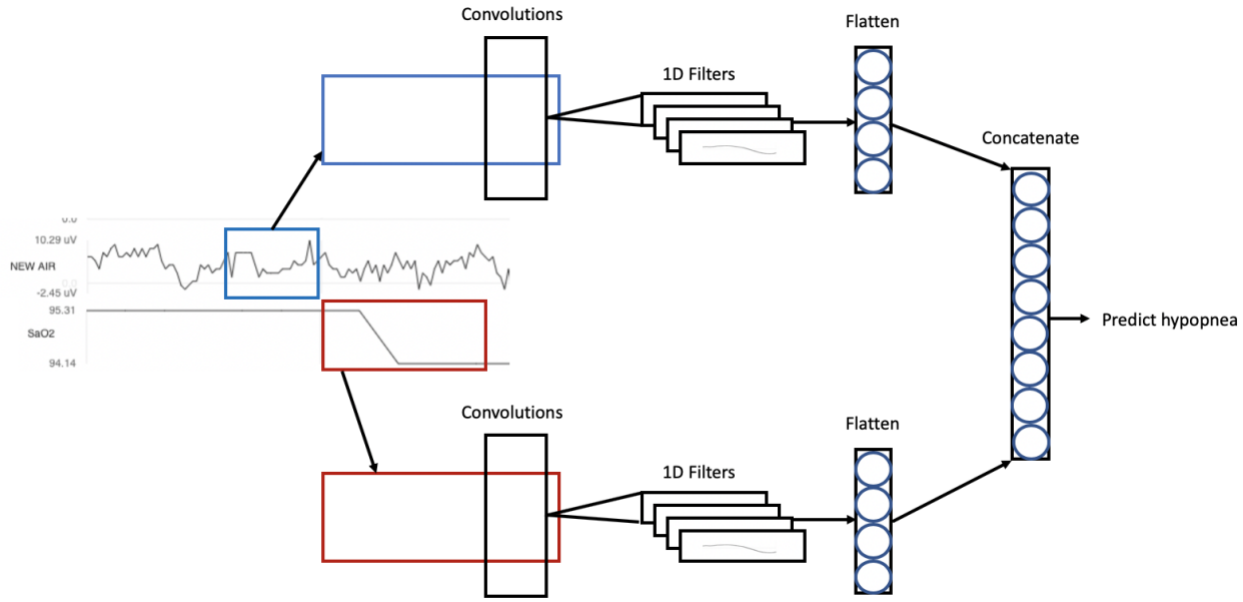

Figure 14. Simplified hypopnea model architecture

The apnea and hypopnea neural network models assign labels to each segment. However, those labels only provide information that an event has occurred within that period of time. To detect the specific seconds in time where an event has occurred, the models use the per-segment labels to calculate a majority vote for each second, where the votes are counted from all segments containing the second under evaluation (Figure 15). Ten or greater seconds with apnea or hypopnea labels are scored as an apnea or hypopnea event, respectively. If both apnea and hypopnea models identify the same section of the PSG, the section is counted as a single respiratory event when calculating AHI.

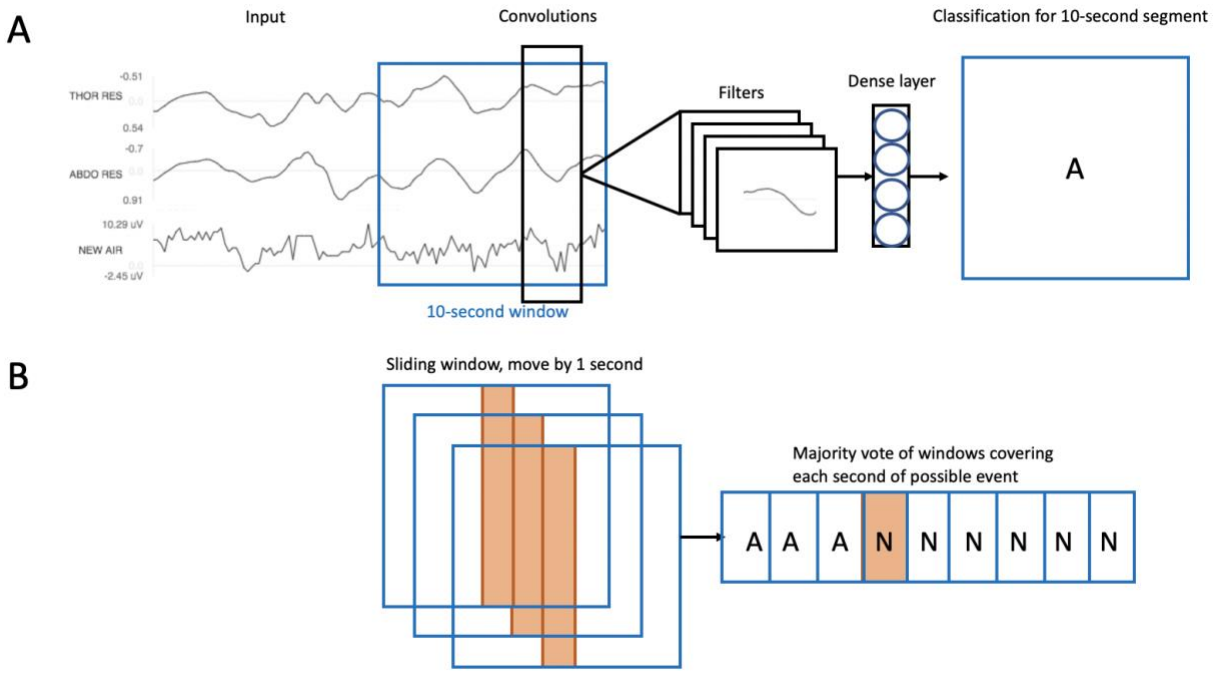

Figure 15. A. Apnea model is used to classify each 10-second segment within the polysomnogram. B. Classification of each 1-second segment is obtained by a majority vote of the all of the classified 10-second windows containing each 1-second segment.
A = Apnea, N = No event

Model tuning

Deep neural networks contain tunable hyperparameters (e.g. number of layers, number of units in each layer, number of filters in convolutional layers, etc). A set of parameter search spaces were defined for each hyperparameter, and the best combination of hyperparameters were found using the python package hyperopt with a random search algorithm for parameter tuning.[108] Multiple hyperparameter configurations were evaluated using the training set. Additionally, models were tested with and without signal preprocessing/denoising.

Model evaluation

Model performance was evaluated with accuracy, area under the receiver operating characteristic, and Cohen's Kappa. Performance of model event detection was calculated using the per-segment accuracy and ROC. AHI prediction accuracy was evaluated by comparing estimated per-patient AHI to human-scored AHI and predicted versus human-scored apnea severity class.

Comparison to rule-based methods

A rule-based event detection model was implemented from the AASM definitions as a baseline for comparison to data-driven models. The rule-based model's pipeline consists of signal preprocessing, amplitude computation from peak/trough point detection, and event detection. Signal preprocessing was performed using a median filter to reduce signal noise followed by normalization of the signal. Peaks and troughs were detected using scipy's find_peaks method. This method finds peaks using the neighbors comparison method, enabling easy calculation and comparison of signal amplitude that comprises the definitions for apneas and hypopneas.

The amplitude was calculated from detected peaks and troughs, and a rule-based model implemented from Lee et a. (2016) was used for event detection. The model uses the amplitude to determine event occurrence, and the duration to keep track of events. An outline of the rules used can be seen in Appendix F and pseudocode can be found in the paper.[73]

Thresholds for the rule-based model were selected based on the contemporary AASM definitions for apneas and hypopneas for the years in which the dataset was scored (follow the 2007 rules). The hypopnea threshold was 0.5 (50% drop in signal) and hypopnea duration was 10 seconds based on the previous 2007 hypopnea definition. The apnea threshold was 0.1, based on the unchanged apnea definition.

The raw AHI predicted values from the rule-based model and deep-learning models were also binned into severity classes such as normal (AHI < 5), mild (AHI = 5-14), moderate (AHI = 15-30) and severe (AHI > 30), and severity class prediction quality was measured.

Results

The optimal architecture for apnea detection consisted of an input of the airflow, chest and thoracic belt channels fed into a convolutional neural network (Table 10, Figure 13). The best model for hypopnea detection consisted of an input of the airflow and SaO_2 channels fed separately into 1-D convolutional layers, ultimately flattened and concatenated to a dense layer (Table 11, Figure 14). Hypopnea model input was structured differently: the airflow channel corresponded with the location in the PSG where an event could be detected, while the

relevant SaO_2 channel data generally followed 30 seconds after the airflow channel window due to the delay in oxygen desaturation after airway obstruction.

Table 10. Apnea model architecture

Layer	Layer Type	Size	Output Size
Input			(1, 100, 3)
C1	Convolutional	(16, 20, 3)	(16, 81, 1)
C2	Convolutional	(16, 20, 1)	(16, 62, 1)
P1	Max Pooling	(2,1)	(16, 31, 1)
R1	Reshape		(496)
D1	Dense	(512)	512
D2	Dense	(256)	256

Table 11. Hypopnea model architecture

Layer	Layer Type	Size	Output Size
Input1			(1, 100)
C1	Convolutional	(16,32)	(16, 69)
C2	Convolutional	(16,32)	(16, 38)
P1	Max Pooling	(2,2)	(8, 18)
R1	Reshape		(144)
D1	Dense	(512)	512
Input2			(1, 100)
C3	Convolutional	(16,32)	(16, 69)
C4	Convolutional	(16,32)	(16, 38)
P2	Max Pooling	(2,2)	(8, 18)
R2	Reshape		(144)
D2	Dense	(512)	512
D1+D2	Concatenate	(1024)	1024

Model testing

The SHHS dataset was split into a 90% training/validation and 10% holdout set. The training/validation was split into training and validation sets, which were used to train the model, select the best input representation (Appendix C) with its optimal deep learning architecture (Appendix D), and tune the model hyperparameters (Appendix E).

Model evaluation

The performance of the models in predicting each event was evaluated (Table 12).

Table 12. Model event detection performance

Model	**Accuracy**	**AUC**	**PPV**	**Sensitivity**
Apnea	97%	0.95	94%	85%
Hypopnea	92%	0.88	90%	83%

The model achieves a high accuracy overall accuracy of 96.1% when predicting apnea severity class based on AHI calculated from the model (Table 13).

Table 13. Model AHI prediction performance

Class	**Accuracy**
Normal	98.2%
Mild	95.5%
Moderate	92.1%
Severe	96.7%

The model predicts overpredicts AHI (mean: 16.3) compared to human-scored AHI (14.6) with a 2.89 mean squared error (MSE). A breakdown of the AHI class predictions is presented in Table 14.

Table 14. Model per-class AHI prediction performance

Class	**Estimated AHI**	**AHI**
Normal	2.2	2.8
Mild	10.1	11.7
Moderate	19.9	22.4
Severe	44.6	50.5

A rule-based model for apnea and hypopnea detection developed in 2016 was implemented for comparison to generated deep learning models (Table 15).[73]

Table 15. Comparison of accuracy of rule-based[73] vs. deep learning model performance in predicting apnea severity class by AHI

Model	Rule-based	Deep learning
No preprocessing	61%	91%
Noise reduction/normalization	87%	96%

The performance of the rule-based model was found to vary based on the quality of preprocessing as well as rule thresholds. The thresholds for apnea and hypopnea (0.1 and 0.5 of amplitude respectively) were chosen based on the contemporary rules used for human event scoring at the time SHHS was conducted. A hypopnea threshold of 0.7 was originally used (based on AASM 2012 definitions) and documented a worse performance. Algorithm-specific thresholds for counting amplitude (thres_count_amp = 6), skipping amplitude calculation after an event (count_skip = 4), and rescaling amplitude when a significant change occurred (thres_amp_over = 1.2) were tested as well (Appendix G). For these selected and tuned thresholds, the deep learning model performed better than the rule-based model.

Discussion

The deep learning models developed in this work achieve a high accuracy and AUC when predicting segmented windows of respiratory events and achieve an accurate AHI prediction for patients with the labeled events. The accuracy and estimation of the predicted AHI is on par with or slightly better than previous works within existing literature and performs better than an expert-designed rule-based method with less user input.

Originally, the airflow, SaO_2 and EEG channels were used as input to predict hypopnea because of the AASM definition of a hypopnea. We found that the EEG channel did not improve detection, which was likely due to the fact that in the SHHS scoring criteria for PSGs, a hypopnea was labeled if the amplitude of any respiratory signal is reduced by 30% of the amplitude of "baseline" and if this change lasts for ≥ 10 s and for >2 breaths, with more subtle changes in breathing (not clearly reduced by 30% or more from baseline) requiring at least a 2% desaturation. Because of this definition, arousals in the labeling of the PSG would not have included arousals, and the definition used to label the PSG may make a significant difference in model performance and affect model design.

Due to the non-limited length of apnea and hypopnea events, a number of different representations for the input data were explored, as well as different architectures for the models. While non-overlapping segments of signal are easier and faster to process, they generally seem to result in an underestimate of respiratory events, resulting in an underestimate of AHI (especially for window sizes of 30+ seconds). On the other hand, overlapping segments greatly increases the number of segments to train and predict on, but resulted in a more accurate prediction. Neural networks consisting of convolutional layers produced the better accuracy when compared to models with recurrent layers, which may be due to the fact that the base rules for apnea and hypopnea events are fairly simple, and recurrent layers are not needed for these simple patterns.

Compared to the implemented rule-based model, the deep learning model performs better on per-segment apnea and hypopnea event detection. The deep learning model required less manual tuning (thresholds and/or variables) and less signal preprocessing than the rule-

based method for a better result. A limitation in this comparison is that only one rule-based algorithm was implemented for comparison. An algorithm by Huang et al. (2017) cites higher performance, but we were unable to implement that model and achieve comparable results based on its written description. More importantly, the deep learning model performs much better on un-processed signal data than the rule-based model. In addition, the model performs better than or on par with other models using traditional machine learning methods[14-17,19] that require human feature generation. To create features to feed into these models, such as SVMs or random forests, expert knowledge of the AASM rules and signal processing is generally required. Engineered features include averaging signal data, calculating statistical moments, or performing other types of transforms. Practically, this means that the data-driven deep learning model requires less human intervention in preprocessing and tuning for a comparable or better result.

There are a few limitations with this work. Since the predicted hypopnea events were scored using a specific criterion, the model may perform worse when compared to newer criteria. Training the model on a dataset scored with the new criteria would mitigate this issue. Another limitation is the method in which AHI events are detected. Events were detected based on whether there was a minimum of 10 consecutive seconds having an event label; if there is a single second gap between seconds without the event label, the chain is broken. This may have contributed to the under-estimation of AHI.

Conclusion

Models were developed to predict apnea, hypopnea and AHI with high accuracy. These deep-learning based models perform as well as or better than rule-based models with less manual labor and human bias. Automated scoring systems may ease the burden of PSG scoring and reduce inter-rater variability in event scoring, improving the PSG scoring process for sleep medicine research and clinical practice.

CHAPTER 5

ENGINEERING PSG-DERIVED FEATURES TO PREDICT SLEEP APNEA-ASSOCIATED OUTCOMES

Introduction

Obstructive sleep apnea (OSA) is a sleep disorder in which multiple interacting pathophysiologic mechanisms lead to recurrent upper airway obstruction and associated physiologic stress. It has been independently linked to multiple health conditions including an increased risk of hypertension, diabetes, cardiovascular disease, stroke risk, and overall mortality.[1-4] It affects approximately 6% of women and 13% of men in the United States.[2] OSA is diagnosed using an overnight sleep study measuring multiple high-resolution physiologic signals called a polysomnogram (PSG). The management of OSA and associated disease risks is largely dependent on a single disease metric: the apnea-hypopnea index (AHI), defined as the average number of airflow limitations (hypopneas) and cessations (apneas) observed per hour of sleep. This disease metric is used to categorize the severity of sleep apnea (Table 16).

Table 16. AHI severity classes

AHI Class	Criteria
None/Minimal	AHI < 5
Mild	$5 \leq$ AHI < 15
Moderate	$15 \leq$ AHI < 30
Severe	AHI ≥ 30

However, overall predictive value of AHI for complications from OSA is low, and a need exists for better prognostic metrics.[37–39] Significant variation in OSA presenting symptoms, disease mechanisms, associated comorbidities, and treatment outcomes has been reported in patients with similar AHI.[109–115] Three flaws in AHI contribute to poor prognostic value: 1) differences in threshold criteria defining respiratory events,[116] 2) variability in human event scoring accuracy,[117,118] and 3) the loss of important physiologic data associating with comorbid disease risk.[37]

In this paper, we develop a model that predicts cardiovascular outcomes with data from the Sleep Heart Health Study (SHHS). SHHS is a large, high-quality dataset within the National Heart, Lung, and Blood Institute that monitored cardiovascular outcomes in 5,804 subjects with and without untreated OSA. [71,86,87] This model we develop includes not only existing clinical and cardiac features, but engineered features from polysomnogram channels. To date, multiple studies have examined cardiovascular disease risk stratification in association with single PSG metrics, such as AHI, but few have tried engineering new features.

A growing body of evidence suggests that a common definition of AHI may be insufficient for predicting the many complications of OSA. This may be due, in part, to the fact that repetitive airway obstructions can variably precipitate other physiologic events such as tachycardia, blood oxygen desaturation, and neurologic arousal. While these related risks are likely important for quantifying associated disease risk, they are only coarsely captured by defining threshold criteria for apneas and hypopneas, if at all. In SHHS, prior work has demonstrated conflicting associations between cardiovascular outcomes and AHI. The prevalence of atrial fibrillation (AF) was demonstrated to be 4-fold higher in those with OSA compared to unaffected individuals with a temporal relationship between respiratory and AF events.[119,120] More recent work identified central, but not obstructive, sleep apnea as having an association with incident AF, while other AHI analyses demonstrated associations with incident heart failure, but not incident coronary disease in SHHS.[121,122] The risk of stroke in the SHHS cohort was 3-fold higher in men in the highest quartile of OSA severity compared to the lowest quartile, but the relationship was less robust in women.[123]

This study helps clarify the relationship between OSA and cardiovascular outcomes as the epidemiologic findings are somewhat discordant with the pathophysiology shared by these two conditions (i.e. inflammation, sympathetic nervous system activation, platelet aggregation) and may be bi-directional (as also suggested by SHHS data).[124] Overall, SHHS data demonstrate that severe OSA doubles the hazard ratio of cardiovascular disease risk in certain subpopulations, but in other groups the prognostic power of AHI appears more limited.[2,37,39]

AHI severity categories ultimately do not correlate well with symptom burden or comorbidity outcomes.[39] This has led to other signal patterns in PSG data being investigated for their relationships with outcomes of interest. Two-percent oxygen desaturations have been found to predict insulin resistance, while 4% oxygen desaturations and REM-predominant OSA predict hypertension. [41-43] The time spent with an oxygen saturation below 90% predicts platelet aggregation, while an increased arousal index predicts memory impairment. [44-46] Definitions of disease burden may therefore need to change depending on a given patient's history, the outcome of interest, and the specific pathophysiologic mechanisms underpinning their OSA.[38]

Different methodologies have been applied in recent years in an attempt to better elucidate factors related to OSA pathogenesis. A top-down, structured approach to phenotyping based on PSG measurements of multiple clinician-defined physiologic traits was described in 2013 by Eckert and colleagues as the PALM Scale (airway critical closure pressure [P_{crit}], arousal threshold, loop gain, muscle responsiveness of the upper airway).[77] These features are measured by actively alternating airway pressures in patients wearing continuous positive airway pressure tolerance (CPAP) and then measuring physiologic responses according to pre-specified criteria. Patients showed significant trait heterogeneity, with over half displaying abnormal non-anatomic features such as hypersensitive loop gain, decreased arousal thresholds, and inadequate pharyngeal dilator muscle responsiveness. Further work has shown that patient populations display different combinations of PALM traits.[78,79] Other research confirms that phenotypes can have differential responses to targeted therapies,[80-85] lending greater weight to the concept of undiscovered OSA subtypes with different underlying disease risks and responses to therapy.

These clinical subtypes, and the mechanisms that underlie them, may only become apparent with the application of more sophisticated mathematical tools. For example, electrocardiographic research has demonstrated changes in heart rate variability and autonomic/respiratory interactions (called cardiopulmonary coupling) in SHHS that correlate with EEG and respiratory event markers of sleep.[125–128] Cardiopulmonary coupling has been significantly associated with comorbidities including hypertension and stroke risk,[129] but it is

difficult for humans to rapidly visually recognize and is unlikely to be the only significant determinant of disease risk. Other clinically meaningful information, not captured by visual scoring methods, may therefore lie within PSG signals. OSA is a distinctly different disease in different populations, and there is a great need for prognostic data beyond the AHI.

This study develops a model that uses clinical, cardiac, and new engineered features to predict cardiovascular outcomes, serving as a proof of concept. This study will help move towards more personalized sleep medicine by outlining important phenotypic elements from PSG data.

Methods

This study was designed as a retrospective analysis of PSG data collected through several multicenter cohort studies available through the National Sleep Research Resource (NSRR). [71,86,87] Study design was approved by the Vanderbilt University Medical Center Institutional Review Board (#171186) and data access was approved by the NSRR.

Study datasets

A logistic regression model was trained on 5,804 Type II PSGs from multiple centers containing patients with and without sleep-disordered breathing collected for the Sleep Heart Health Study (SHHS). The study contains two sets of visits, with the first and larger set used here (SHHS 1; Table 17).

Table 17. SHHS 1 summary statistics

Category	Mean	Median	Min, Max
Age	63.1	63	[39, 90]
Body Mass Index	28.2	27.5	[18, 50]
Apnea Hypopnea Index	17.9	13.2	[0, 161.8]
Sleep Time (minutes)	359.8	367.0	[34.5, 519]

SHHS data

The data included in the SHHS study ranges from clinical data such as age and gender, full night polysomnogram data that has been labeled by experts, and cardiac outcomes over the following 10 years.

Polysomnography data

All PSG files were downloaded in the European Data Format containing the raw time series data of physiologic signals from each PSG as well as human-scored sleep stages and apneic events. For this study, 5,213 patients were randomly selected from the SHHS1 dataset. The model was trained and validated using 5-folds cross validation with the SHHS1 data.

Outcomes data

The patients in the SHHS study were kept track of and followed up with for outcomes information. The patients were targeted from the following studies: Atherosclerosis Risk in Communities Study (1,750 participants), Cardiovascular Health Study (1,350 participants), Framingham Heart Study (1,000 participants), Strong Heart Study (600 participants), New York Hypertension Cohorts (1,000 participants), and Tucson Epidemiologic Study of Airways Obstructive Diseases and the Health and Environment Study (900 participants).

Outcomes data for SHHS subjects belonging to these cohorts were provided to the SHHS by the parent cohorts. ARIC, CHS, FHS, and SHS studies had mechanisms in place for determining CVD outcomes since the start of SHHS. The SHHS subjects recruited in Tucson and New York were members of research cohorts that did not include ongoing assessment of CVD outcomes. In these two sites, SHHS investigators have implemented their own procedures for ascertaining and adjudicating CVD outcomes among SHHS participants. The outcomes that we use are shown in Table 18.

Table 18. Number of positive cases of each cardiac event

Cardiac event	Positive cases for those age > 65	Positive cases for those age < 65
Angina	344 (9.8%)	17 (0.7%)
Congestive heart failure	584 (16.7%)	43 (1.7%)
Myocardial infarction	310 (8.8%)	55 (2.2%)
Stroke	263 (7.5%)	26 (1.1%)
Fatal cardiovascular disease	340 (9.7%)	18 (0.7%)
Fatal coronary heart disease	221 (6.3%)	12 (0.5%)

Model design

We use machine learning methods create a model that predicts if a cardiac event will occur in the future (within 10 years). The model is built from various features available within the SHHS data and the polysomnogram data. This includes:

1. AHI
2. Clinical features
    - Demographic
    - Laboratory
    - Cardiac
3. Engineered features

Age ended up being a dominant factor in predicting cardiac risk. Due to this, we decided to split the data into two groups: age > 65 and age < 65, and not include age as a feature. The different channels from the polysomnogram that we used included: EEG (arousals), airflow (breathing).

Feature selection

We selected a number of clinical features from those available from the SHHS. These features included the demographic feature gender, as well as a number of laboratory values (Table 19) and cardiologists' ECG labeled determinations (described in the next section).

Table 19. Clinical demographic and laboratory features

Feature	Description
AHI	Apnea-Hypopnea Index
Gender	Patient gender
Weight	Patient weight
Waist	Patient waist circumference
BMI	Body-Mass Index
Chol	Cholesterol
HDL	HDL
Trig	Triglycerides
FEV1	Forced expiratory volume
FVC	Forced vital capacity
DiasBP	Diastolic blood pressure
SystBP	Systolic blood pressure

Clinical ECG features

ECGs have been previously used to find features associated with respiratory signals. ECGs have been found to provide information to identify apneic epochs.[129,130] With signal processing, respiration can be extracted out of the ECG signals with a reasonable degree of reliability. This can be valuable information when studying patients with OSA.[131] Extracting respiration from the ECG signal has been found to be possible by three mechanisms: the physical effect of respiration causes displacement of the ECG electrodes, ventilation changes

the volume of air within the lung which alters the amplitude of the ECG signal, and respiration causes heart rate variability (HRV).[132]

ECG features were based on the ECG channel measured in the polysomnogram. The ECGs from the SHHS study were labeled by expert cardiologists; we use these expert labels (Table 20). Recently, there's been a lot of work in using deep learning to find these labels.

Table 20. Clinical cardiac features

Feature	Description
antlatmi	Anterolateral myocardial infarction (MI)
antsepmi	Anteroseptal myocardial infarction (MI)
apbs	Atrial bypasses
av1deg	First degree atrioventricular block
av3deg	Third degree atrioventricular block
ilbbb	Incomplete left bundle-branch block
infmi	Inferior myocardial infarction (MI)
irbbb	Incomplete right bundle-branch block
iventblk	Indeterminate intraventricular block pattern
lah	Left Atrial Hypertrophy
lbbb	Left bundle-branch block
lvh3_1	Left Ventricular Hypertrophy: Voltage 3-1
lvh3_3	Left Ventricular Hypertrophy: Voltage 3-3
lvhst	Left Ventricular Hypertrophy with ST and T-wave abnormalities
mob1	Mobitz Type-1 Heart Block
mob2	Mobitz Type-2 Heart Block
nodal	Nodal rhythm
nonsp_st	Nonspecific ST wave abnormalty
nonsp_tw	Nonspecific T wave abnormalty
paced	Paced rate

part2deg	Partial second degree atrioventricular block
qrs	QRS Axis
rbbb	Right bundle-branch block
rtrial	Right Atrial Enlargement
rvh	Right Ventricular Hypertrophy
st4_1_3	ST and T-wave 4-1 to 4-3
st5_1_3	ST and T-wave 5-1 to 5-3
truposmi	True posterior myocardial infarction (MI)
ventrate	Ventricular rate
vpbs	Ventricular bypasses
wpw	Wolff-Parkinson-White Syndrome

Engineered features

We engineer features from the sleep studies of the patients using both clinical and signal processing knowledge. These features are derived from data solely contained within PSGs. In engineering these features, we enlisted clinician help for ideas on sleep-related aspects of specific data channels (i.e., respiratory, ECG) that are not well described nor often used to predict cardiac outcomes. We then create features that describe these aspects of the data.

Engineered respiratory features

Respiratory features were based on the airflow channel in a polysomnogram. This channel measures the airflow during sleep, generally used to find and label apneas and hypopneas. Traditionally, these are then used to calculate AHI. While AHI gives a good estimate of the number of respiratory events that occur over a period of time, it doesn't give information about those events themselves.

A number of physiological features have been shown to occur in response to respiratory events, like oxygen saturation drops or arousals. Although the immediate consequences of drops in oxygen saturation throughout the night is not yet clear, it has been associated with various conditions such as carotid wall thickening and plaque occurrence[133], excessive daytime

sleepiness[134], cancer progression[135], and neurobehavioral and autonomic alterations[136]. This suggests that other physiological parameters besides those measured by AHI, such as the magnitude of oxygen desaturation, event length, or clustering of events may contribute differently to OSA. With this in mind, we selected a few features that encompass this information (Table 21).

Table 21. Engineered respiratory features

Feature	Description
Average apnea length	Average apnea event length
Average hypopnea length	Average hypopnea event length
Average time between apneas	Average time between apnea events
Average time between hypopneas	Average time between hypopnea events
SASHB	Area under the curve (SaO_2) of respiratory event to baseline

We selected the average event length to better characterize respiratory events, since the AHI only takes into account the number of events that have occurred. The average time between events was selected to better characterize the clustering of those events.

The Sleep Apnea-Specific Hypoxic Burden (SASHB) is the area under the curve of the oxygen saturation for the oxygen desaturation following those respiratory events.[137] It is calculated by determining the baseline for SaO_2, and then finding the area between the SaO_2 curve and the baseline. This concept has been recently used to predict cardiovascular disease-related mortality[138] and incident HF[137]. For each identified respiratory event, the pre-event baseline saturation was defined as the maximum SaO_2 during the 100 seconds before the end of the event. The area under this baseline value was calculated over the desaturation that followed the respiratory event as labeled by sleep technicians (Figure 16).

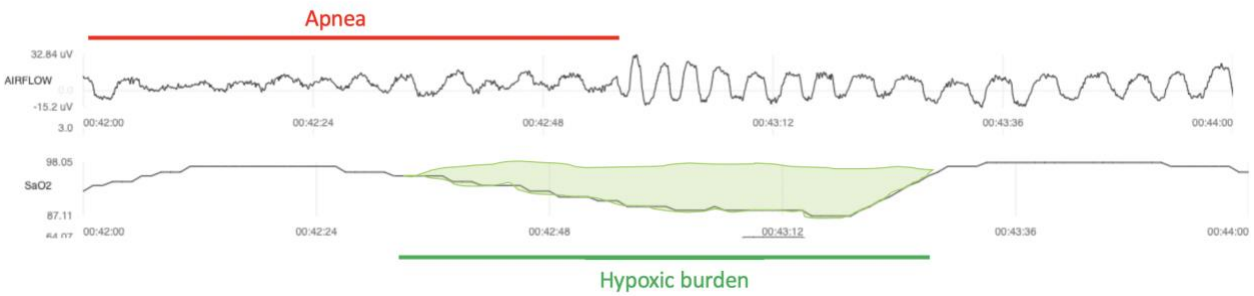

Figure 16. Sleep Apnea-Specific Hypoxic Burden

Engineered EEG Features

EEG recordings are widely used in sleep medicine because of its ability to detect variations in cortical activity and differentiate wake from different stages of sleep. Guidelines for scoring wake and sleep stages standardize and improve upon previous ad-hoc methods.[39,40] Because sleep is currently assessed visually by sleep technicians, there is possibly more information to be gained from EEG routinely collected EEG signal. Efforts to further process these stored data will be helpful in better characterizing sleep physiology in patients with OSA.

Simple summary EEG features were designed to capture information about arousals, which aren't currently widely used to evaluate OSA (Table 22).

Table 22. Engineered EEG features

Feature	Description
Arousal duration	Duration of arousal
Arousal average time between	Average time between arousals

Engineered ECG Features

Several ECG features were tested and selected based on literature of QT interval implications on cardiac death[139], P-wave differences in OSA patients[140], and HRV relationships with OSA[141] (Table 23).

Table 23. Engineered ECG features

Feature	Description
QT interval	Length of Q-T interval
P-wave area difference	Difference in area under P-wave between event and non-event breathing
R-R interval	Average interval between R peaks

Model validation and testing

We tested a number of different models including logistic regression, SVMs, random forests, and deep learning models. We settled on a logistic regression model with the features as input, because of performance and interpretability of the model. We oversample the minority class to ensure a class balance within the imbalanced data, and measure model performance using accuracy and area under the receiver operating characteristic (ROC) curve. Models are tuned using a random walk search over a set of parameters. We use 5-fold cross validation to measure performance for the final result.

Results

The final model consists of a combination of AHI (1), clinical (42) and engineered (5 respiratory, 2 EEG, 3 ECG) features. We plot the training and testing accuracy vs the number of samples to ensure the models converge properly (Figures 17-22).

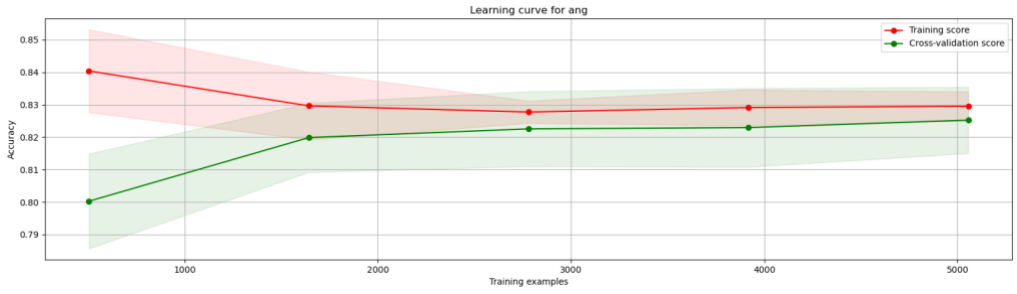

Figure 17. Angina learning curve

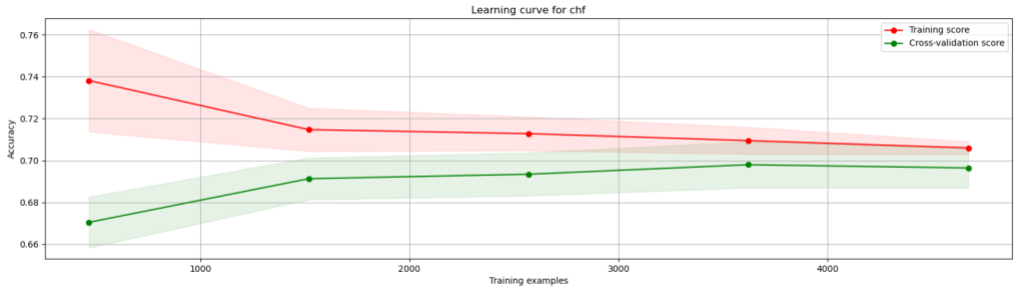

Figure 18. CHF learning curve

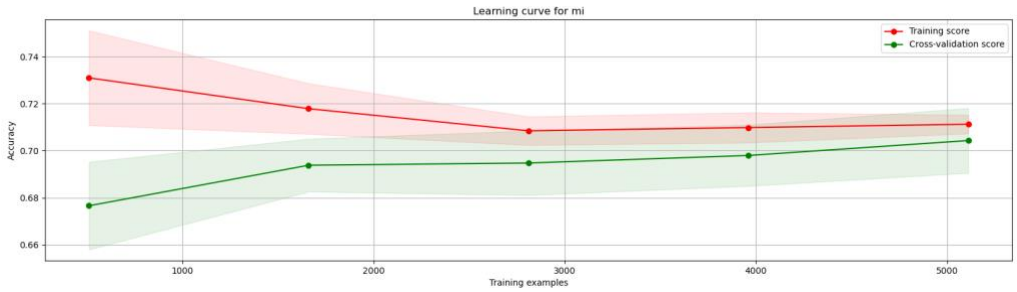

Figure 19. Myocardial Infarction learning curve

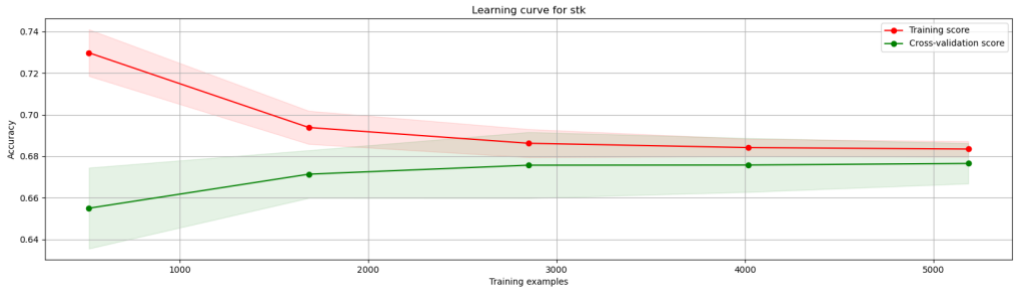

Figure 20. Stroke learning curve

Figure 21. CVD learning curve

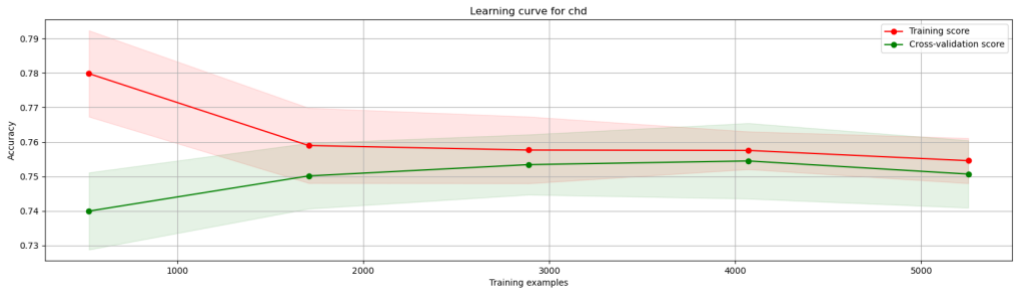

Figure 22. CHD learning curve

We show the model performance against a baseline model using only AHI as the feature, as well as the final combined model compared to AHI and clinical features (Figures 23-28). A significant increase in AUC ($p < 0.05$) is marked by an asterisk.

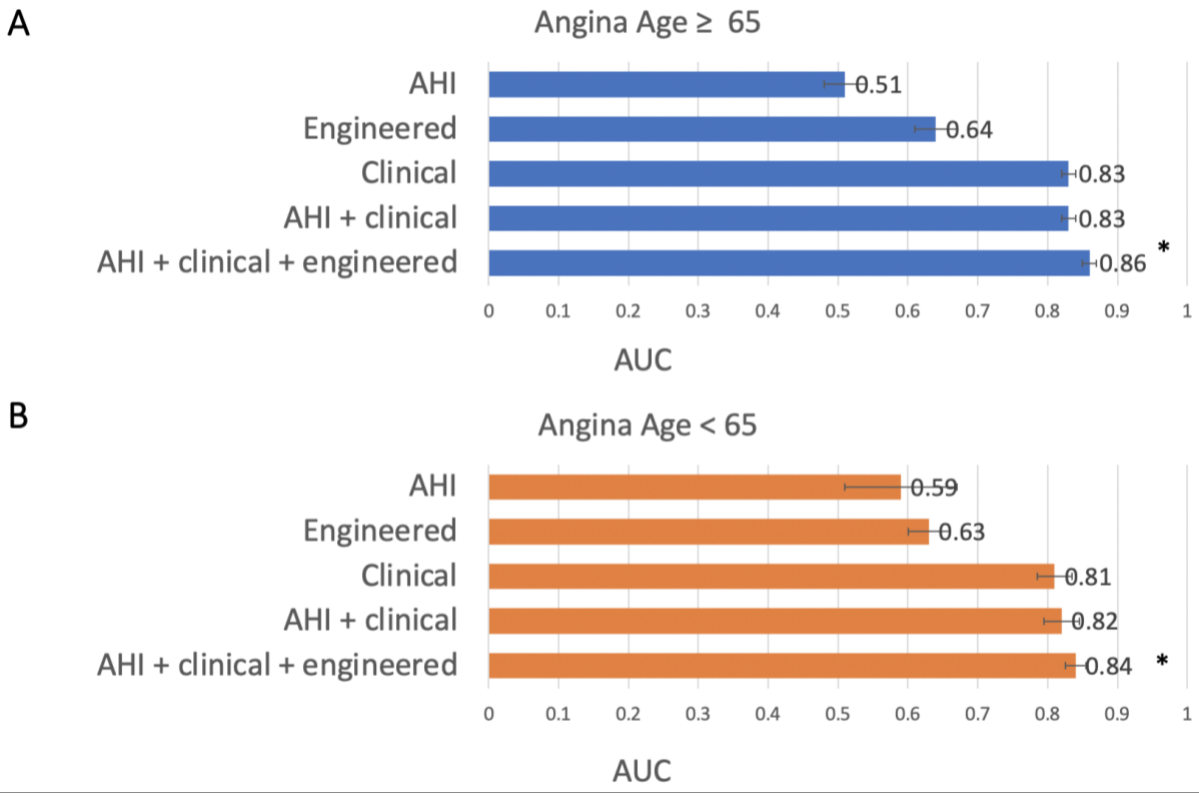

Figure 23. Angina model

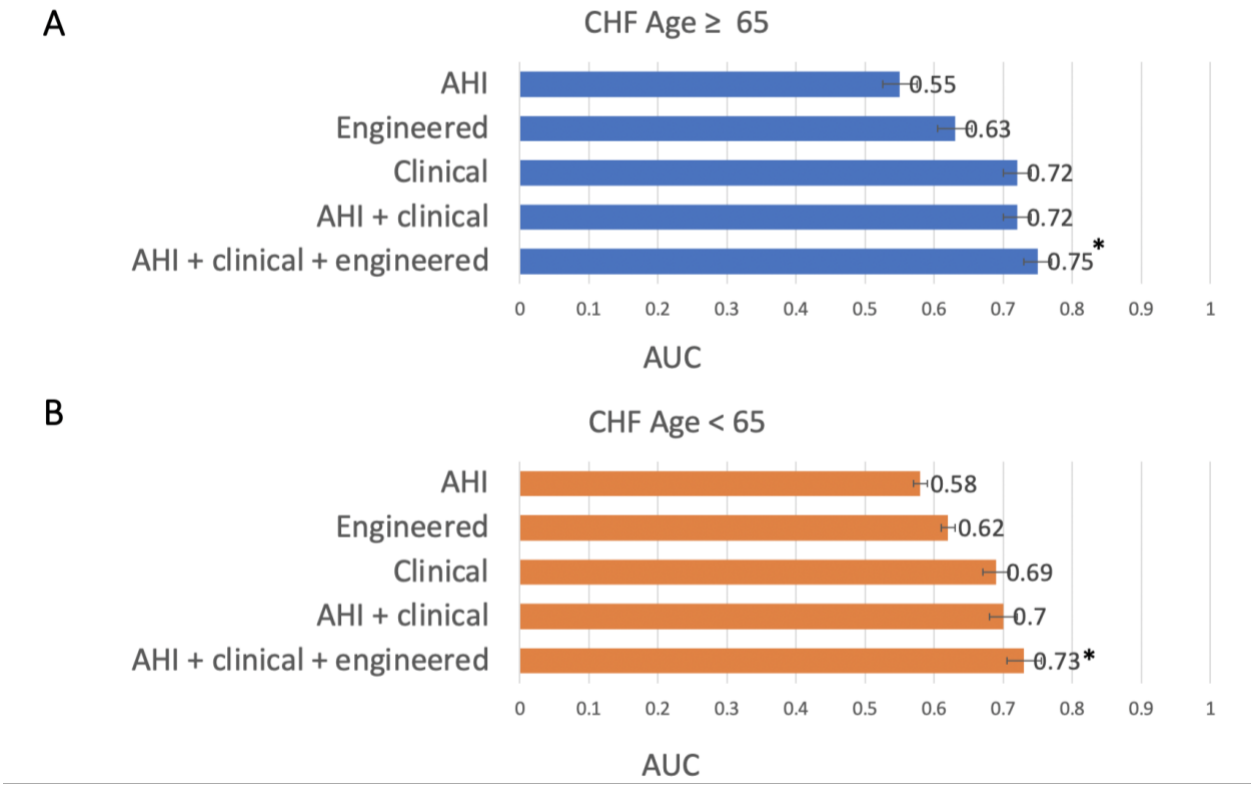

Figure 24. Congestive heart failure model

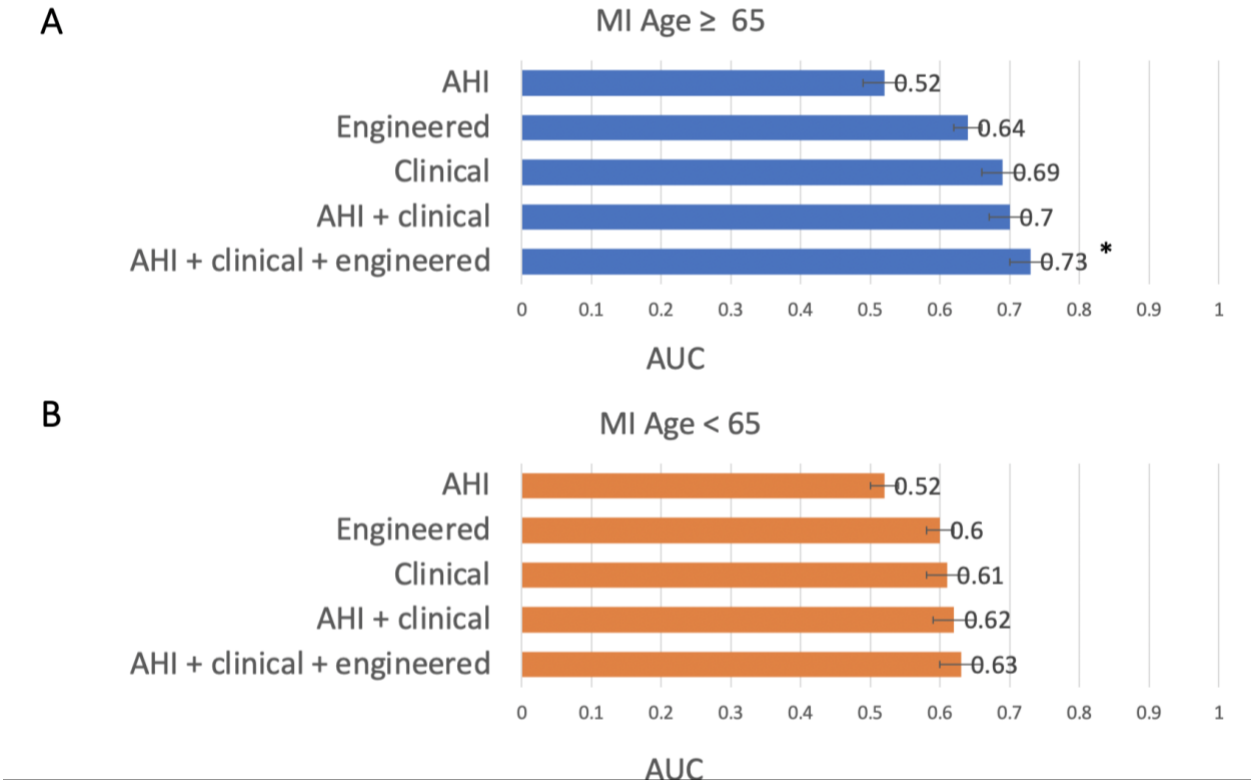

Figure 25. Myocardial infarction model

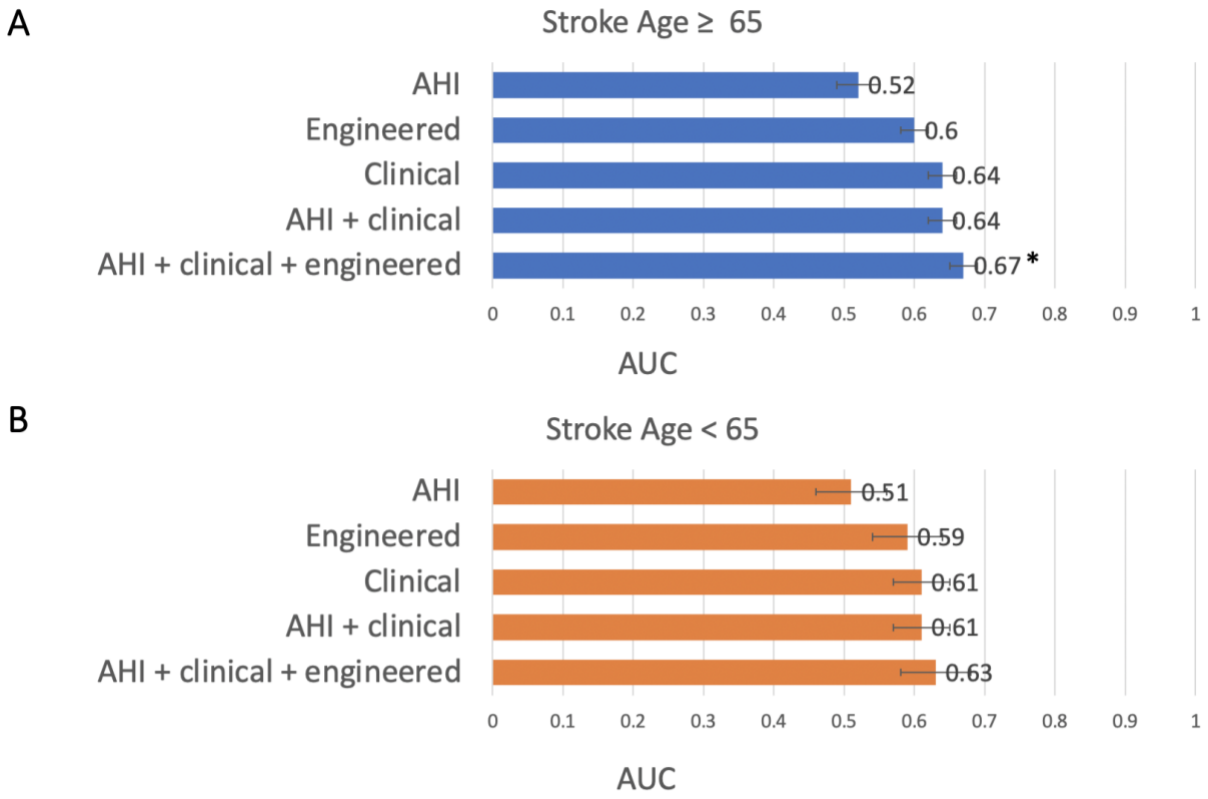

Figure 26. Stroke model

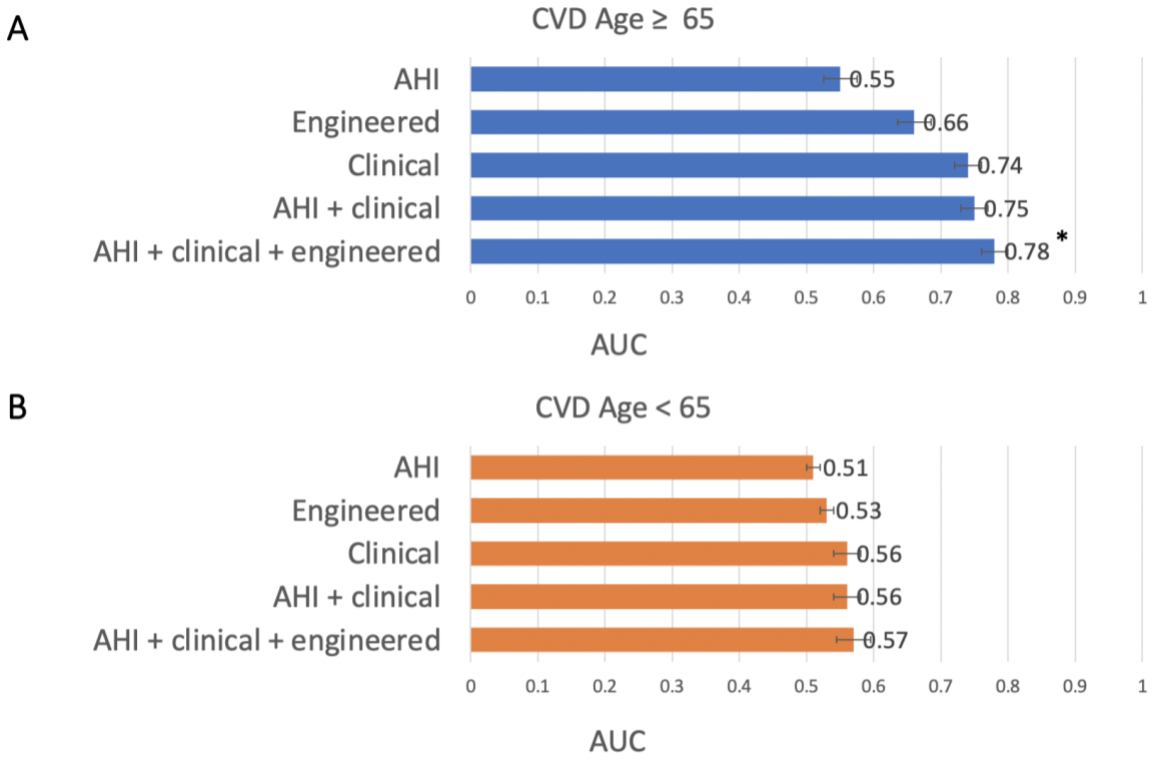

Figure 27. Cardiovascular disease model

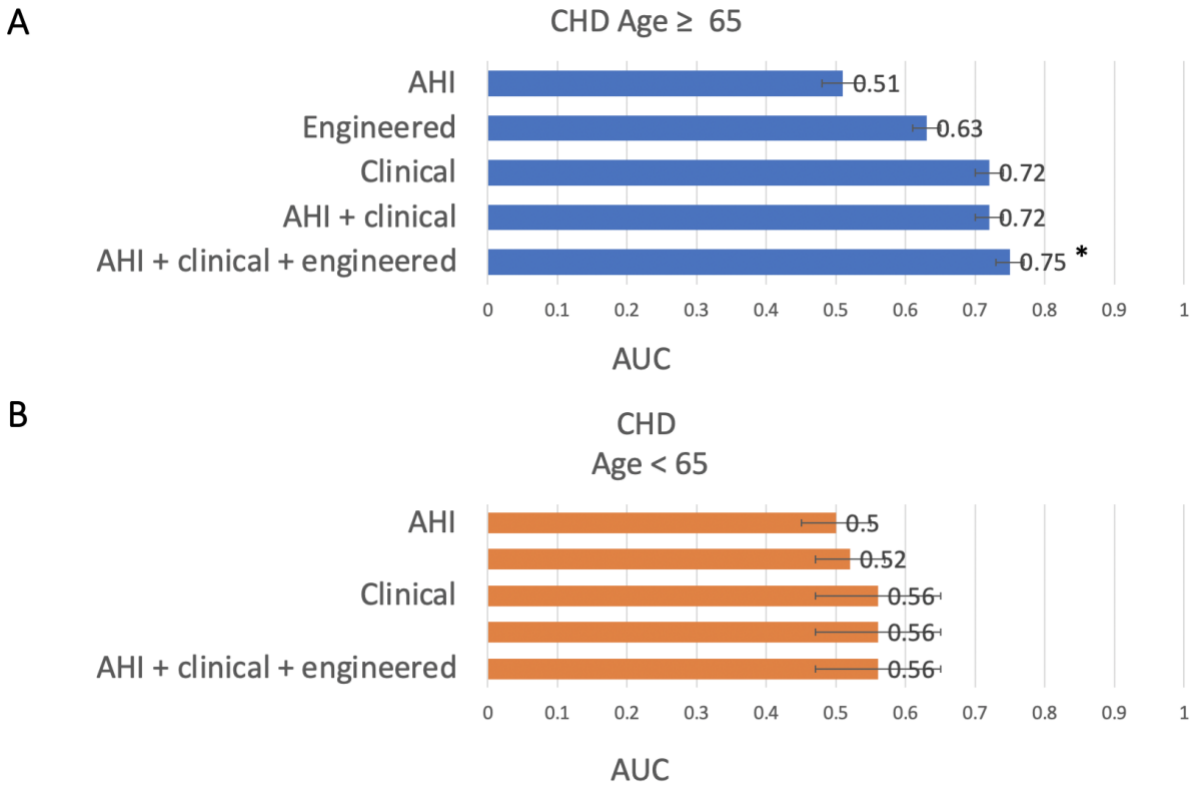

Figure 28. Coronary heart disease model

The results of each of the final models is summarized in Table 24. Additionally, we check for significance between the results of the models with and without the engineered features using a t-test to compare the 5-fold cross validated accuracies between the two models. Tables for age < 65 can be found in Appendix H.

Table 24. Summary of model outcomes for age ≥ 65

Model Outcome	**AUC**
Angina	0.86 +/- 0.013 *
CHF	0.76 +/- 0.023 *
MI	0.76 +/- 0.022 *
Stroke	0.67 +/- 0.018 *
CVD	0.78 +/- 0.020 *
CHD	0.75 +/- 0.025 *

* Indicates significance at $p < 0.05$

The impact of the features in the model can be inferred by examining the weights of the features in the logistic regression model (Table 25). We examine the features for angina, the outcome that had the highest predictability with the feature set.

Table 25. Feature weights for angina model

Feature	**Weight**
ahi_a0h3a	-0.02
Gender	-0.33
Weight	-0.12
Waist	0.10
bmi_s1	0.15
Chol	-0.23
HDL	-0.08
Trig	0.16
FEV1	-0.05
FVC	-0.32
DiasBP	-0.42
SystBP	0.46
lvh3_1	-0.13
lvh3_3	-0.19
st4_1_3	-0.57
st5_1_3	0.37
lvhst	0.78
mob1	-0.97
part2deg	-1.40
mob2	-0.97
av3deg	-0.97
av1deg	0.20
lbbb	0.19
rbbb	-0.08
ilbbb	0.01
irbbb	-0.21
lah	0.02
iventblk	-0.02
wpw	1.07
antsepmi	0.10
infmi	-0.05
antlatmi	0.13
nonsp_st	0.37

nonsp_tw	-0.23
rtrial	0.58
rvh	0.69
VENTRATE	0.48
QRS	0.01
AFIB	0.31
PACED	2.86
nodal	-0.53
apbs	-0.82
vpbs	-0.48
Average apnea length	-0.63
Average hypopnea length	0.55
Average time between apneas	0.38
Average time between hypopneas	-0.34
SASHB	0.87
Arousal duration	-0.27
Arousal average time between	-0.58
QT interval	0.41
P-wave area difference	-0.21
R-R interval	0.18

To summarize the most important features for each of the cardiac outcomes, we find the top 5 features (Table 26).

Table 26. Top 5 features by weight per cardiac outcome

Angina	**CHF**	**MI**	**Stroke**	**CVD**	**CHD**
PACED	Arousal time btwn	Arousal duration	nodal	av3deg	PACED
part2deg	lah	wpw	SASHB	mob1	vpbs
mob1	nodal	PACED	Hypopnea duration	FVC	rvh
mob2	rvh	Arousal time btwn	wpw	mob2	iventblk
av3deg	part2deg	part2deg	vpbs	PACED	av3deg

Discussion

We developed machine learning models to predict the occurrence of six different cardiovascular events within the following 10 years. We compare the performance breakdown of each group of features to the engineered features that we implement and to the gold standard, AHI. We show that AHI has a weak predictive performance for cardiac outcomes, and features that better describe the apneas, hypopneas, and arousals outperform it.

In examining the features that have higher weights in the logistic regression model, we can infer the features that have a higher impact on predicting the outcome. We find that cardiac features that describe atrioventricular blocks, heart blocks, bypasses and pace rate, plus the SASHB feature have the highest weights in predicting angina. The AHI feature appears to have one of the lowest weights within the model for angina.

The results for the age < 65 may be affected by the low number of actual cardiovascular events that occur. In particular, the stroke, CVD and CHD have the lowest number of cases and the worst performance. While we oversample the minority class for the model, this means that the samples are all similar and likely does not fully describe cohort at risk for those outcomes.

There are several limitations within this study. One limitation is that we predicted cardiac outcomes as occurrence within 10 years. This was selected due to the available cardiovascular outcomes data but could be better refined. In addition, there is possibly missing data in the follow up for the SHHS patients; there are possible missing cardiovascular events.

AHI has in the past been shown to have conflicting associations with cardiovascular outcomes and not correlate well with symptom burden or comorbidity outcomes. We find that using AHI as a feature for 6 different cardiovascular outcomes mirrors this finding; for angina and CHF, AHI is predictive, but for myocardial infarction, stroke, CVD and CHD it has very little predictive power. We show that engineered features that better describe the apneas, hypopneas, and arousals that occur during sleep in most cases make better features for predicting those cardiovascular outcomes.

Future work in unsupervised phenotype discovery and PSG analysis

Future work in phenotype discovery may be aided by deep learning, which has been shown to have success in fields such as natural language processing and audio. In this future work section, we present an idea we used to attempt to summarize and organize PSG data with. Work in the implementation of this model allowed us to learn about the difficulties in summarizing PSG data.

PSG2VEC

In the field of natural language processing, the sentence and document representation problems have been solved using models called WORD2VEC and DOC2VEC[142,143]. WORD2VEC is a well-known language model that converts words into a vector representation of term semantics that can be mathematically manipulated (e.g. king - man + women = queen). DOC2VEC leverages a corpus of WORD2VEC-transformed terms to create a vector representation of a document for easy comparison between documents. Our idea was to create a similar model, which we dubbed PSG2VEC, in which representations of signal (analogous to words) could be used to describe a PSG.

Autoencoders are a specific type of neural network that consist of an input layer, hidden layer, and output layer. The weights of the hidden layer are updated iteratively during training to produce an output that is as similar as possible to the input. After training, the hidden layer can be used to transform the input into a compact representation. For PSG data, the encoding layer will consist of convolutional layers to identify spatial patterns. It has been shown that CNN-based autoencoders can find features that outperform other methods in classification tasks, including principal component analysis and sparse random projection in EEG signal.[58–60] The compact autoencoder representation will be used to capture a dictionary of features from raw PSG data. Relevant features may be simple, easily recognizable single channel patterns such as k-complexes or sleep spindles, but can also be far more complex, spanning multiple channels and physiologic signal interactions over large time intervals. These encoded features would then be used in a sequence learning model to generate a single vector representation of a PSG, termed PSG2VEC.

The PSG2VEC approach analogously leverages the auto-encoded features as “words”, and sequences of those PSG “words” to learn a “document” representation (Figure 29). The resulting PSG “document” representation would allow for the comparison and stratification of PSG data.

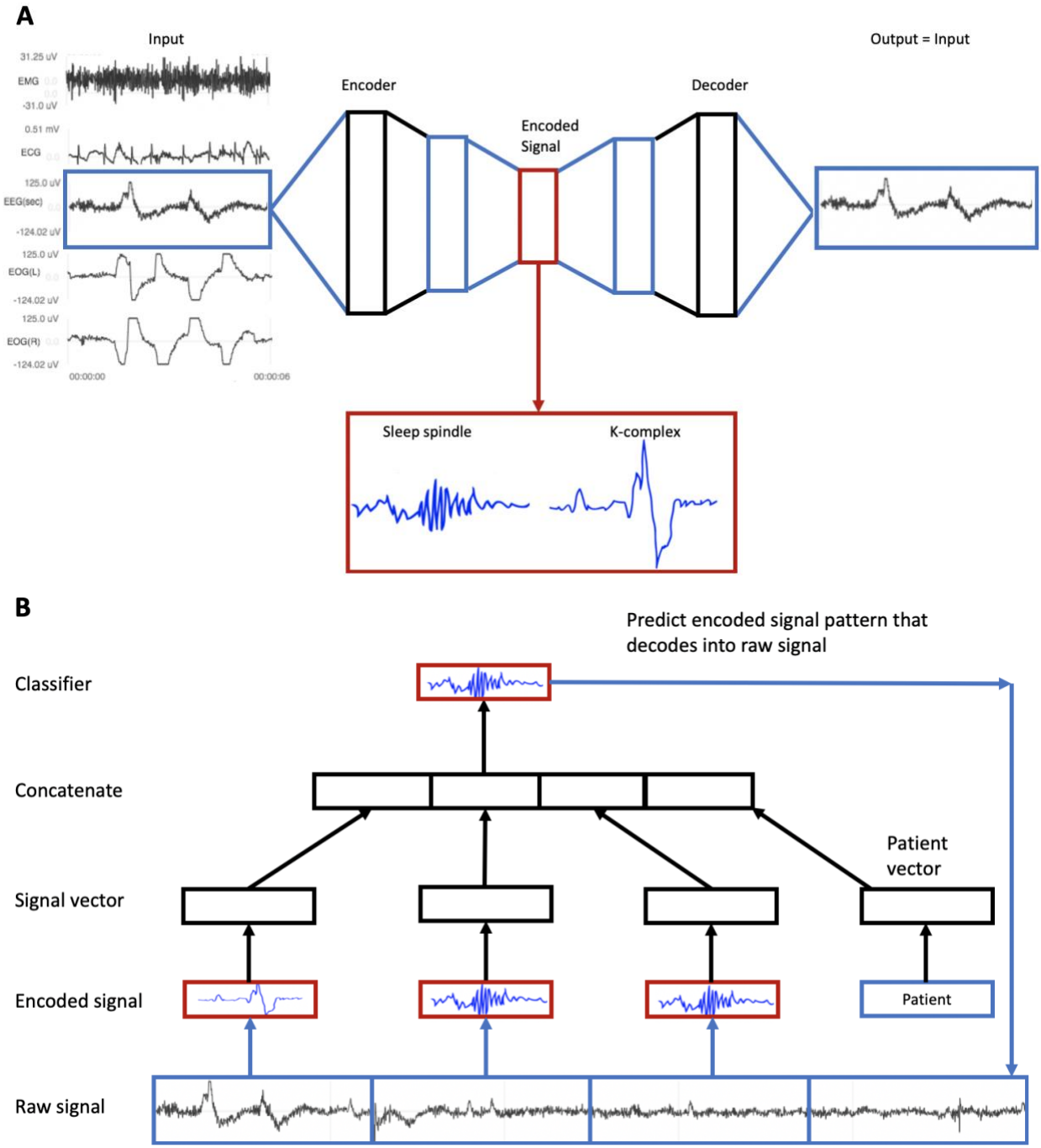

Figure 29. A. PSG data are used to train an autoencoder that capture important sleep

characteristics. The resulting encoded vectors results in a sequence of compressed features that represents the signal. B. The PSG2VEC model takes a sequence of encoded PSG features and a patient ID to predict the next pattern/signal encoding in the sequence. The trained network outputs an encoding for patient a PSG to a PSG2VEC vector that summarizes the characteristics of an entire PSG exam, which can be used for patient stratification.

These PSG2VEC vectors could be leveraged as predictors for cardiovascular outcomes by themselves and in combination with clinician-defined features. Moreover, PSG2VEC vectors from multiple patients can be compared and correlated with other patient characteristics.

Our work in this area did not yield a working model, but allowed us to learn about the difficulties that must be solved in in future work in order to do so. PSG data is much more complex than in natural language in that signals are noisy and not so well-defined as words. Autoencoders were used to attempt to create more defined units analogous to words, but the variation in signal makes the task difficult. There is the disentanglement problem: how to represent the information present in a compact and interpretable structure. In addition, we want the representation to be meaningful to our goal, phenotyping sleep apnea (and correlation to its outcomes). Some ideas for future work in this area include using inductive bias, or using labels and related metadata to help direct the formation of those representations.

The other problem we found was that the depth and breadth of the PSG data was hard to summarize succinctly. Essentially, we want to make a metric or data representation like AHI that includes more relevant information. The idea of a model similar to DOC2VEC is hindered without the smaller units that have distinct meaning (i.e. words). In the audio domain, autoencoders have been used to categorize songs, but here the PSG data is too much to process in a similar manner. For future work, the biggest challenge is finding a way to summarize or analyze the data in a meaningful manner.

CHAPTER 6
CONCLUSIONS

Sleep apnea is a highly preventable disease that can be more efficiently treated if we have better tools to measure, characterize and phenotype it. With the rise of machine and deep learning methods in the last decade, we have the capability to more efficiently and quickly categorize and analyze polysomnogram data that has previously been laborious to label and complex to evaluate. The work presented in this thesis takes a step towards the improvement of sleep study analysis systems and provides insight and suggestions for better phenotyping sleep apnea.

Contribution and innovation

The management of OSA and associated disease risks currently remain largely dependent on the single disease metric, the apnea-hypopnea index. Literature has shown, and we have mirrored the finding, that predictive value of AHI for complications from OSA is low (cardiac, in our work), and a need exists for better prognostic metrics. Significant variation in OSA presenting symptoms, disease mechanisms, associated comorbidities, and treatment outcomes has been reported in patients with similar AHI. We describe three flaws in AHI that contribute to poor prognostic value: 1) differences in threshold criteria defining respiratory events, 2) variability in human event scoring accuracy, and 3) the loss of important physiologic data associating with comorbid disease risk. We used these three flaws as a baseline upon which we developed three aims for different studies.

To enhance the efficiency and accuracy of PSG analysis, we developed and evaluated staging and event scoring models. The evaluation of these models demonstrated that deep-learning based models perform as well as or better than humans with less manual labor and human bias. Automated scoring systems may ease the burden of PSG scoring and reduce inter-rater variability in staging and event scoring, improving and expediting the PSG scoring process for sleep medicine research and clinical practice. Automated systems are our answer for the first two flaws with AHI that we determined.

In the third aim, we developed sleep-derived features from sleep studies to predict cardiac outcomes associated with sleep apnea. These additional features derived from polysomnogram data were found to have a higher predictive contribution than AHI and added additional information to aid prediction on top of AHI, demographic and clinical data. This is the first step in defining better prognostic measures and discovering and defining phenotypes for obstructive sleep apnea.

Limitations

The work presented is limited in several dimensions that require further development and testing in order to better apply these models and features to practice.

Effect of different event and staging definitions

One flaw with AHI is the differences in threshold criteria defining respiratory events. In a similar fashion, there are different criteria for determining sleep stage: AASM and RK. Both Aim 1 and 2 studies used only one set of criteria, and the effect of additional data from a different set is unknown. Model accuracy outcomes may differ between new vs. old AASM respiratory event definitions, or AASM sleep staging criteria and RK staging criteria. In the future, the newest set of criteria should be chosen, or the models should be trained on a mixture of newer and older criterion's data. This also brings up another question: is the current definition the best definition of a respiratory event or a sleep stage (both are human-defined), or should the definitions be more data-driven and unbiased?

Transferability across different PSG types and additional data sources

The datasets examined in these works are composed of Type II PSGs recorded in subject home environments with a limited, single EEG channel montage. Generalizability to more common Type I or Type III PSGs could not be evaluated, and we suspect that training the model with additional EEG signals available in Type I PSGs would likely yield performance improvements to models using that channel because more data from those additional channels would be available. As availability and usage of consumer wearables becomes higher,

comparison with more limited montage datasets, such as consumer wearables utilizing actigraphy and heart rate monitoring, becomes possible. Work in this area is limited by the lack of large, publicly available datasets with this type of data source.

Timeframe and labeling in PSG event or staging detection and analysis

In the apnea and hypopnea event detection models, we defined a respiratory event as occurring when we find 10 or more seconds successively defined as belonging to an event. A single negatively defined second in 10+ seconds breaks the chain, and a possible event would be left undetected. This may have contributed to the under-estimation of AHI for our model. This type of continuity problem always must be dealt with when using timeseries data.

Another issue is that it is known that there is variability in the labeling of apnea and hypopnea events. This variability exists in selecting existing events, as well defining the start point and duration of those events. This leads to a limitation in the standard for event detection; the human-labels cannot make up a true gold standard. Despite that, however, since there is a high degree of agreement between human-labels, the labels that make up this silver standard are still useful. The same applies to stage labeling, minus the variability in staging duration since those are defined to exist in 30 second periods, which brings us to another limitation.

The definition of staging being limited to 30 second periods is a human definition that was created for simplicity. Sleep staging describes a continuous cycle of the stages of sleep, which are not contained physiologically in 30 second periods. There is inaccuracy in describing the sleep stages in such a manner. Future work may explore continuous sleep staging, which could more accurately describe that area of analysis.

PSG data summarization and feature selection for phenotyping

In our work to attempt phenotyping obstructive sleep apnea using PSG data, we found that the breadth and depth of PSG data coupled with the lack of a method of summarization hindered our efforts. The positive point about AHI is that it is a singular metric that describes some information pertaining to sleep apnea, though it is lacking in that information. It is

difficult to find a similar metric that is more meaningful. Sleep study data contains hours of multiple channels of physiological signal data, which can be hard to analyze and even harder to summarize.

Prior to engineering specific features using previous knowledge from literature, we attempted to use methods used in other timeseries data to summarize a PSG. These methods included models similar to WORD2VEC, or models used in the auditory domain. We found that PSG data is more complex in that signals are noisy and not so well-defined as words, and that a single PSG held too much data to be summarized in a similar manner to a song. Future work may want to attempt this feat again, but engineering features using clinical knowledge was our path towards that grander goal.

Clinical and informatics implications

The methods developed in this thesis demonstrate that advancements in computer science and informatics can be applied to and improve the solutions to clinical problems. We developed models using deep-learning methods that can be directly applied to PSG labeling in a clinical setting, enabling more efficient and less labor-intensive sleep study analysis. While there may be some limitations with these models, we demonstrated generalizability for both the event detection and sleep stage scoring models. Automated scoring systems can largely improve the productivity of sleep labs, for example reducing scoring time of sleep staging from 2 hours per PSG to 2 minutes.

The engineered features and methods for selection presented in our cardiac prediction models demonstrate a strong first step towards sleep apnea phenotyping and prognostic development. Clinical knowledge can be combined with signal processing and machine learning to find features that help to better describe and predict relevant outcomes to sleep apnea. We also detail the limitations currently inherent in dealing with and attempting to summarize PSG data, and some areas to look into for future work.

Future work

This work is a step towards better defining sleep apnea phenotypes and prognostics. The event detection and sleep staging models help process and analyze PSGs in an efficient, less variable manner. Further work in this area includes exploring and solidifying those definitions, possibly in a more data-driven and precise manner. Our work in engineered features is the start of forming better descriptors of sleep apnea correlated with relevant outcomes and complications. Further work can be done to discover more descriptors using complications and outcomes other than the cardiac ones used in our study. These descriptors can be used to cluster groups of sleep apnea cohorts and start to define different phenotypes.

Apart from using features derived from clinical knowledge, deep learning may be used to perform unsupervised exploration of the PSG data. As mentioned in the limitations section, work in other areas of research such as document summarization and song categorization may lend some relevant techniques to this endeavor. We have proved in our modeling of respiratory event detection and sleep staging that deep-learning lends itself well to signal data and can be used to predict human-defined labels. This indicates that deep learning may also be able to perform well in the more unsupervised space and is the next area of exploration in this field.

Appendix A

Tested sleep staging machine learning model architectures

Model	Description
Markov Chain	Simple Markov model, predict next stage based on prior probabilities observed
CNN	Raw signal as input fed into a convolutional neural network
Spectrogram + LSTM	Spectrogram for 30 seconds of signal data as input, fed into LSTM network
Spectrogram + CNN	Spectrogram for 30 seconds of signal data as input, fed into convolutional neural network
Spectrogram + CNN + LSTM	Spectrogram for 30 seconds of signal data as input, fed into convolutional layers whose features are fed into LSTM layers

Appendix B

Sleep staging machine learning model hyperparameter search space

Parameter	Search space
Number of convolutional layers	1,2,3,4,5
Number of filters	32,64,96,128
Filter width	1,2,3,4
Filter height	32,64,128
Number of pooling layers	1,2,3
Width/height of pooling layer	2,3,4
Number of LSTM layers	1,2,3
Number of LSTM units	256, 512, 768, 1024
Number of dense layers	1,2,3
Number of units in dense layer	256, 512, 768, 1024

Appendix C

Apnea and hypopnea window search parameters

Input representation
Window = 5, overlap = 0
Window = 10, overlap = 0
Window = 15, overlap = 0
Window = 20, overlap = 0
Window = 25, overlap = 0
Window = 30, overlap = 0
Window = 60, overlap = 0
Window = 5, overlap = 1
Window = 10, overlap = 1
Window = 15, overlap = 1
Window = 20, overlap = 1
Window = 5, overlap = 3
Window = 10, overlap = 3
Window = 15, overlap = 3
Window = 20, overlap = 3
Window = 10, overlap = 5
Window = 15, overlap = 5
Window = 30, overlap = 5

Appendix D

Apnea and hypopnea event detection architectures and performance

Input representation and architecture type effect on apnea model performance – most significant input representation and model architectures are listed, though more combinations were tested.

Model	**Performance (Patient AHI Class Accuracy)**
Window = 5, overlap = 0 Convolutional layers	0.72
Window = 10, overlap = 0 Convolutional layers	0.87
Window = 15, overlap = 0 Convolutional layers	0.83
Window = 20, overlap = 0 Convolutional layers	0.85
Window = 25, overlap = 0 Convolutional layers	0.81
Window = 30, overlap = 0 Convolutional layers	0.80
Window = 60, overlap = 0 Convolutional layers	0.75
Window = 5, overlap = 0 Convolutional+recurrent layers	0.74
Window = 10, overlap = 0 Convolutional+recurrent layers	0.87
Window = 15, overlap = 0 Convolutional+recurrent layers	0.82
Window = 20, overlap = 0 Convolutional+recurrent layers	0.86
Window = 25, overlap = 0 Convolutional+recurrent layers	0.81
Window = 30, overlap = 0	0.80

Convolutional+recurrent layers	
Window = 60, overlap = 0 Convolutional+recurrent layers	0.72
Window = 5, overlap = 1 Convolutional layers	0.80
Window = 10, overlap = 1 **Convolutional layers**	**0.96**
Window = 15, overlap = 1 Convolutional layers	0.94
Window = 20, overlap = 1 Convolutional layers	0.91
Window = 5, overlap = 3 Convolutional layers	0.90
Window = 10, overlap = 3 Convolutional layers	0.94
Window = 15, overlap = 3 Convolutional layers	0.91
Window = 20, overlap = 3 Convolutional layers	0.90
Window = 10, overlap = 5 Convolutional layers	0.90
Window = 15, overlap = 5 Convolutional layers	0.87
Window = 30, overlap = 5 Convolutional layers	0.85

Appendix E

Apnea and hypopnea machine learning model hyperparameter search space

Parameter	Search space
Number of convolutional layers	1,2,3,4,5
Number of filters	16, 32,64,128
Filter width	1,2,3,4,5,10
Filter height	2,3
Number of pooling layers	1,2,3
Width/height of pooling layer	2,3
Number of dense layers	1,2,3
Number of units in dense layer	256, 512, 768, 1024

Appendix F
Rule-based model summary

- For signal x, check if point is trough:
- If trough:
    1. Compute mean amplitude based on the last six peak-to-trough measurements if the hypopnea threshold is exceeded
    2. Determine if the point is in a hypopnea event by keeping count of such points
    3. Check threshold duration of event
    4. Determine if an event has ended based on current amplitude of signal and count of event duration
    5. Stop computation of mean amplitude if event has ended, or adjust if necessary
- If not a trough:
    1. Determine if point is in an apnea event by checking amplitude compared to mean amplitude and apnea count
    2. If apnea event is detected using the apnea threshold, override hypopnea count

Appendix G
Rule-based model parameters

Final algorithm parameters:
apnea_thres = 0.1; hypopnea_thres = 0.5; apnea_duration = 10 s; hypopnea duration = 10 s
thres_count_amp = 6; count_skip = 4; thres_amp_over = 1.2

Performance of algorithm different parameters

Parameter	AUC
Model with the above parameters	0.87
hypopnea_thres = 0.7	0.84
thres_count_amp = 4	0.85
thres_count_amp = 5	0.86
thres_count_amp = 7	0.84
count_skip = 3	0.86
count_skip = 5	0.86
thres_amp_over = 1.1	0.85
thres_amp_over = 1.3	0.86

Appendix H
Summary of model outcomes for age $<$ 65

Model Outcome	**AUC**
Angina	0.84 +/- 0.017 *
CHF	0.73 +/- 0.026 *
MI	0.63 +/- 0.042
Stroke	0.63 +/- 0.031
CVD	0.57 +/- 0.023
CHD	0.56 +/- 0.054

* Indicates significance at $p < 0.05$

REFERENCES

1. Marin JM, Carrizo SJ, Vicente E, Agusti AGN. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet Lond Engl. 2005;365(9464):1046-1053. doi:10.1016/S0140-6736(05)71141-7.
2. Punjabi NM, Caffo BS, Goodwin JL, et al. Sleep-disordered breathing and mortality: a prospective cohort study. PLoS Med. 2009;6(8):e1000132. doi:10.1371/journal.pmed.1000132.
3. Terán-Santos J, Jiménez-Gómez A, Cordero-Guevara J. The association between sleep apnea and the risk of traffic accidents. Cooperative Group Burgos-Santander. N Engl J Med. 1999;340(11):847-851. doi:10.1056/NEJM199903183401104.
4. Young T, Finn L, Peppard PE, et al. Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. Sleep. 2008;31(8):1071-1078.
5. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol. 2013; 177:1006–14.
6. Polysomnography (2018). https://aystesis.com/polysomnography/. Accessed Jan 2019.
7. Berry RB, Brooks R, Gamaldo, CE, et al. The AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specifications (American Academy of Sleep Medicine, Darien, IL, 2012).
8. Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. National Institutes of Health Publication no. 204; 1968.
9. Iber C. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology And Technical Specifications. American Academy of Sleep Medicine; 2007 IL.
10. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep. 2013; 36(4): 573–582.

11. Silber MH, Ancoli-israel S, Bonnet MH, et al. The visual scoring of sleep in adults. J Clin Sleep Med. 2007;3(2):121-31.
12. Danker-hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res. 2009;18(1):74-84.
13. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med. 2002;3:43–7.
14. Loredo JS, Clausen JL, Ancoli-Israel S, Dimsdale JE. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep. 1999;22:916–20.
15. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver Agreement among sleep scorers from different centers in a large dataset. Sleep. 2000;23:901–8.
16. Bliwise D, Bliwise NG, Kraemer HC, Dement W. Measurement error in visually scored electrophysiological data: respiration during sleep. J Neurosci Meth. 1984;12:49–56.
17. Lord S, Sawyer B, Pond D, et al. Inter-rater reliability of computer-assisted scoring of breathing during sleep. Sleep. 1989;12:550–8.
18. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep. 1998;21:749–57.
19. Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med. 1998;158:358–62.
20. Younes M, Kuna ST, Pack AI, et al. Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice. J Clin Sleep Med. 2018;14(2):205-213.
21. Schaltenbrand N, Lengelle R, Toussaint M, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep. 1996; 19(1):26–35.
22. Anderer P, Gruber G, Parapatics S, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. Neuropsychobiology. 2005; 51(3): 115–133.

23. Berthomier C, Drouot X, Herman-Stoïca M, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep. 2007; 30(11): 1587–1595.
24. Anderer P, Moreau A, Woertz M, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. Neuropsychobiology. 2010; 62(4): 250–264.
25. Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. Methods Inf Med. 2010; 49(3): 230–237.
26. Liang SF, Kuo CE, Hu YH, Cheng YS. A rule-based automatic sleep staging method. J Neurosci Methods. 2012; 205(1): 169–176.
27. Lajnef T, Chaibi S, Ruby P, et al. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi- class support vector machines. J Neurosci Methods. 2015; 250: 94–105.
28. Wang Y, Loparo KA, Kelly MR, Kaplan RF. Evaluation of an automated single-channel sleep staging algorithm. Nat Sci Sleep. 2015; 7: 101.
29. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. Sleep. 2015; 38(10): 1555–1566.
30. Hassan AR, Bhuiyan MIH. A decision support system for automatic sleep staging from EEG signals using tunable q-factor wavelet transform and spectral features. J Neurosci Methods. 2016; 271: 107–118.
31. Younes M, Younes M, Giannouli E. Accuracy of automatic polysomnography scoring using frontal electrodes. J Clin Sleep Med. 2016; 12(5): 735–746.
32. Younes M, Soiferman M, Thompson W, Giannouli E. Performance of a new portable wireless sleep monitor. J Clin Sleep Med. 2017; 13(2): 245–258.
33. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med. 2002;3:43–7.
34. Loredo JS, Clausen JL, Ancoli-Israel S, Dimsdale JE. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep. 1999;22:916–20.

35. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep. 1998;21:749–57.
36. Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med. 1998;158:358–62.
37. Malhotra A, Orr JE, Owens RL. On the cutting edge of obstructive sleep apnoea: where next? *Lancet Respir Med*. 2015;3(5):397-403. doi:10.1016/S2213-2600(15)00051-X
38. Dempsey JA, Veasey SC, Morgan BJ, O'Donnell CP. Pathophysiology of sleep apnea. *Physiol Rev*. 2010;90(1):47-112. doi:10.1152/physrev.00043.2008.
39. Shahar E, Whitney CW, Redline S, et al. Sleep-disordered breathing and cardiovascular disease: cross-sectional results of the Sleep Heart Health Study. *Am J Respir Crit Care Med*. 2001;163(1):19-25. doi:10.1164/ajrccm.163.1.2001008.
40. Hudgel DW. Sleep apnea severity classification – revisited. *SLEEP* 2016;39(5):1165–1166.
41. Mokhlesi B, Finn LA, Hagen EW, et al. Obstructive sleep apnea during REM sleep and hypertension. results of the Wisconsin Sleep Cohort. *Am J Respir Crit Care Med*. 2014;190(10):1158-1167. doi:10.1164/rccm.201406-1136OC.
42. Punjabi NM, Newman AB, Young TB, Resnick HE, Sanders MH. Sleep-disordered breathing and cardiovascular disease: an outcome-based definition of hypopneas. *Am J Respir Crit Care Med*. 2008;177(10):1150-1155. doi:10.1164/rccm.200712-1884OC.
43. Punjabi NM, Beamer BA. Alterations in Glucose Disposal in Sleep-disordered Breathing. *Am J Respir Crit Care Med*. 2009;179(3):235-240. doi:10.1164/rccm.200809-1392OC.
44. Rahangdale S, Yeh SY, Novack V, et al. The influence of intermittent hypoxemia on platelet activation in obese patients with obstructive sleep apnea. *J Clin Sleep Med JCSM Off Publ Am Acad Sleep Med*. 2011;7(2):172-178.
45. Djonlagic I, Guo M, Matteis P, Carusona A, Stickgold R, Malhotra A. Untreated sleep-disordered breathing: links to aging-related decline in sleep-dependent memory consolidation. *PloS One*. 2014;9(1):e85918. doi:10.1371/journal.pone.0085918
46. Djonlagic I, Saboisky J, Carusona A, Stickgold R, Malhotra A. Increased sleep fragmentation leads to impaired off-line consolidation of motor memories in humans. *PloS One*. 2012;7(3):e34106. doi:10.1371/journal.pone.0034106

47. Krishnan, P. Polysomnography: recording and sleep staging. Slideshare (2009). https://www.slideshare.net/drpramodkrishnan/polysomnography. Accessed Jan 2017.
48. Krizhevsky, A, Sutskever, I, Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012: 1106-1114.
49. Lee, Honglak & Grosse, Roger & Ranganath, Rajesh & Ng, Andrew. (2011). Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. Commun. ACM. 54. 95-103. 10.1145/2001269.2001295.
50. Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013, pp. 6645–6649.
51. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc.; 2012:1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed February 13, 2020.
52. Cecotti H, Graser A. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans Pattern Anal Mach Intell*. 2010;33(3):433–445.
53. UFLDL Tutorial: Convolutional neural network. http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/. Accessed Jan 2017.
54. Barbounis TG, Theocharis JB, Alexiadis MC, Dokopoulos PS. Long-term wind speed and power forecasting using local recurrent neural network models. *IEEE Trans Energy Convers*. 2006;21(1):273–284.
55. Olah, C. Understanding LSTM Networks. 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed Jan 2017.
56. Agarwal R, Gotman J. Computer-assisted sleep staging. IEEE Trans Biomed Eng. 2001;48(12):1412-23.
57. Zoubek, Lukáš, et al. Feature selection for sleep/wake stages classification using data driven methods. Biomedical Signal Processing and Control 2.3 (2007): 171-179.

58. Fraiwan, L., N. Khaswaneh, and Khaldon Y. Lweesy. Automatic sleep stage scoring with wavelet packets based on single EEG recording. World Academy of Science, Engineering and Technology 54.3 (2009): 485-88.
59. Bajaj, Varun, and Ram Bilas Pachori. Automatic classification of sleep stages based on the time-frequency image of EEG signals. Computer methods and programs in biomedicine 112.3 (2013): 320-328.
60. Chapotot, Florian, and Guillaume Becq. Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules. International Journal of Adaptive Control and Signal Processing 24.5 (2010): 409-423.
61. Helland, VC Figueroa, et al. Investigation of an automatic sleep stage classification by means of multiscorer hypnogram. Methods of Information in Medicine 49.05 (2010): 467-472.
62. Helland, VC Figueroa, et al. Investigation of an automatic sleep stage classification by means of multiscorer hypnogram. Methods of Information in Medicine 49.05 (2010): 467-472.
63. Jo, Han G., et al. Genetic fuzzy classifier for sleep stage identification. Computers in Biology and Medicine 40.7 (2010): 629-634.
64. Güneş, Salih, Kemal Polat, and Şebnem Yosunkaya. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. Expert Systems with Applications 37.12 (2010): 7922-7928.
65. Doroshenkov, L. G., V. A. Konyshev, and S. V. Selishchev. Classification of human sleep stages based on EEG processing using hidden Markov models. Biomedical Engineering 41.1 (2007): 25-28.
66. Dong, Jing, et al. Automated sleep staging technique based on the empirical mode decomposition algorithm: a preliminary study. Advances in Adaptive Data Analysis 2.02 (2010): 267-276.
67. Koley, B., and D. Dey. An ensemble system for automatic sleep stage classification using single channel EEG signal. Computers in biology and medicine 42.12 (2012): 1186-1195.

68. Hsu, Yu-Liang, et al. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. Neurocomputing 104 (2013): 105-114.
69. Krakovská, Anna, and Kristína Mezeiová. Automatic sleep scoring: A search for an optimal combination of measures. Artificial intelligence in medicine 53.1 (2011): 25-33.
70. Budhiraja R, Thomas R, Kim M, Redline S. The Role of Big Data in the Management of Sleep-Disordered Breathing. Sleep Med Clin. 2016;11(2):241-55.
71. Dean DA, Goldberger AL, Mueller R, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016;39(5):1151-64.
72. Koley BL, Dey D. Real-time adaptive apnea and hypopnea event detection methodology for portable sleep apnea monitoring devices. IEEE Trans Biomed Eng. 2013 Dec;60(12):3354-63. doi:10.1109/TBME.2013.2282337.
73. Lee H, Park J, Kim H, et al. New rule-based algorithm for real-time detecting sleep apnea and hypopnea events using a nasal pressure signal. J Med Syst 40, 282 (2016). https://doi.org/10.1007/s10916-016-0637-8.
74. Huang W, Guo B, Shen Y, Tang X. A novel method to precisely detect apnea and hypopnea events by airflow and oximetry signals. Comput Biol Med. 2017 Sep 1;88:32-40. doi: 10.1016/j.compbiomed.2017.06.015.
75. Choi SH, Yoon H, Kim HS, Kim HB, Kwon HB, Oh SM, Lee YJ, Park KS. Real-time apnea-hypopnea event detection during sleep by convolutional neural networks. Comput Biol Med. 2018 Sep 1;100:123-131. doi: 10.1016/j.compbiomed.2018.06.028.
76. Yu H, Deng C, Sun J, et al. Cascading detection model for prediction of apnea-hypopnea events based on nasal flow and arterial blood oxygen saturation. Sleep Breath (2019). https://doi.org/10.1007/s11325-019-01886-4.
77. Eckert DJ, White DP, Jordan AS, Malhotra A, Wellman A. Defining phenotypic causes of obstructive sleep apnea: Identification of novel therapeutic targets. *Am J Respir Crit Care Med*. 2013;188(8):996-1004. doi:10.1164/rccm.201303-0448OC.
78. Bradley A. Edwards, Andrew Wellman, Scott A. Sands, Robert L. Owens, Danny J. Eckert DPW, Malhotra A. Obstructive Sleep Apnea in older adults is a distinctly different

physiological phenotype. Sleep (6). Am J Respir Crit Care Med. 2015;192(9):1128. doi:10.5665/sleep.3844.

79. Joosten SA, Edwards BA, Wellman A, et al. The Effect of Body Position on Physiological Factors that Contribute to Obstructive Sleep Apnea. Sleep. 2015;38(9):1469-1478. doi:10.5665/sleep.4992.
80. Owens RL, Edwards BA, Eckert DJ, et al. An Integrative Model of Physiological Traits Can be Used to Predict Obstructive Sleep Apnea and Response to Non Positive Airway Pressure Therapy. Sleep. 2015;38(6):961-970. doi:10.5665/sleep.4750.
81. Eckert DJ, Owens RL, Kehlmann GB, et al. Eszopiclone increases the respiratory arousal threshold and lowers the apnoea/hypopnoea index in obstructive sleep apnoea patients with a low arousal threshold. Clin Sci Lond Engl 1979. 2011;120(12):505-514. doi:10.1042/CS20100588.
82. Edwards BA, Sands SA, Eckert DJ, et al. Acetazolamide improves loop gain but not the other physiological traits causing obstructive sleep apnoea. J Physiol. 2012;590(5):1199-1211. doi:10.1113/jphysiol.2011.223925.
83. Edwards BA, Sands SA, Owens RL, et al. Effects of hyperoxia and hypoxia on the physiological traits responsible for obstructive sleep apnoea. J Physiol. 2014;592(20):4523-4535. doi:10.1113/jphysiol.2014.277210.
84. Schwartz AR, Gold AR, Schubert N, et al. Effect of weight loss on upper airway collapsibility in obstructive sleep apnea. Am Rev Respir Dis. 1991;144(3 Pt 1):494-498. doi:10.1164/ajrccm/144.3_Pt_1.494.
85. Schwartz AR, Schubert N, Rothman W, et al. Effect of uvulopalatopharyngoplasty on upper airway collapsibility in obstructive sleep apnea. Am Rev Respir Dis. 1992;145(3):527-532. doi:10.1164/ajrccm/145.3.527.
86. Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., & Wahl, P. W. (1997). The Sleep Heart Health Study: design, rationale, and methods. Sleep, 12,1077–1085.
87. Redline, S., Sanders, M. H., Lind, B. K., Quan, S. F., Iber, C., Gottlieb, D. J., Bonekat, W. H., Rapoport, D. M., Smith, P. L., & Kiley, J. P. (1998). Methods for obtaining and analyzing

unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. Sleep, 7, 759–767.

88. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010; 22 (10): 1345-1359.
89. Biswal S, Kulas J, Sun H, Goparaju B, Westover M, Bianchi M, Sun J. SLEEPNET: Automated Sleep Staging System via Deep Learning. arXiv. 2017; 1-17. https://arxiv.org/pdf/1707.08262.pdf
90. M. Sharma, D. Goyal, P. Achuth, U. R. Acharya. An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank. Computers in Biology and Medicine., vol 98, pp. 58-75, June (2018).
91. Sors, A, Bonnet S, Mirek S, Vercueil L, Payen, JF. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomedical Signal Processing and Control. 2018; 42: 107–114.
92. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. 2017; 25 (11): 1998–2008.
93. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv. 2016; 1–10. https://arxiv. org/abs/1610.01683
94. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2018; 26(4): 758-769.
95. Nazaré TS, da Costa GB, Contato WA, Ponti M. Deep convolutional neural networks and noisy images. InIberoamerican Congress on Pattern Recognition 2017 Nov 7 (pp. 416-424). Springer, Cham.
96. Stephan Z, et al. Improving the robustness of deep neural networks via stability training. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Jun 26, 2016 – Jul 1, 2016; IEEE, Piscataway, NJ. 4480–4488.

97. Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MW. An end-to-end framework for real-time automatic sleep stage classification. Sleep. 2018; 41 (5): 1-11.

98. Sun C, et al. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: ICCV. Oct 22-29, 2017; IEEE, Piscataway, NJ.

99. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med. 2002;3:43–7.

100. Marin JM, Carrizo SJ, Vicente E, Agusti AG. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet. 2005;365(9464):1046–1053.

101. Campos-Rodriguez F, Martinez-Garcia MA, de la Cruz-Moron I, Almeida-Gonzalez C, Catalan-Serra P, Montserrat JM. Cardiovascular mortality in women with obstructive sleep apnea with or without continuous positive airway pressure treatment: A cohort study. Ann Intern Med. 2012;156(2):115–122.

102. Martinez-Garcia MA, Campos-Rodriguez F, Catalan-Serra P, et al. Cardiovascular mortality in obstructive sleep apnea in the elderly: role of long-term continuous positive airway pressure treatment: a prospective observational study. Am J Respir Crit Care Med. 2012;186(9):909–916.

103. Yaggi HK, Concato J, Kernan WN, Lichtman JH, Brass LM, Mohsenin V. Obstructive sleep apnea as a risk factor for stroke and death. N Engl J Med. 2005;353(19):2034–2041.

104. Zhang L, Fabbri D, Upender R, Kent D. Automated sleep stage scoring of the sleep heart health study using deep neural networks. Sleep. 2019;42(11). doi:10.1093/sleep/zsz159.

105. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinforma. 2018;19:1236–1246. doi: 10.1093/bib/bbx044.

106. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

107. Paszke A, Gross S, Massa F et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 2019; 32: 8024-8035.

108. James B, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a Python library for model selection and hyperparameter optimization. Comput. Sci. Disc. 2015; 8:014008. https://doi.org/10.1088/1749-4699/8/1/014008.

109. Ayas NT, Owens RL, Kheirandish-Gozal L. Update in Sleep Medicine 2014. *Am J Respir Crit Care Med*. 2015;192(4):415-420. doi:10.1164/rccm.201503-0647UP

110. Ye L, Pien GW, Ratcliffe SJ, et al. The different clinical faces of obstructive sleep apnoea: a cluster analysis. *Eur Respir J*. 2014;44(6):1600-1607. doi:10.1183/09031936.00032314

111. Eckert DJ, White DP, Jordan AS, Malhotra A, Wellman A. Defining phenotypic causes of obstructive sleep apnea: Identification of novel therapeutic targets. *Am J Respir Crit Care Med*. 2013;188(8):996-1004. doi:10.1164/rccm.201303-0448OC

112. Vavougios GD, George D G, Pastaka C, Zarogiannis SG, Gourgoulianis KI. Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. *J Sleep Res*. 2016;25(1):31-38. doi:10.1111/jsr.12344

113. Barbé F, Durán-Cantolla J, Sánchez-de-la-Torre M, et al. Effect of continuous positive airway pressure on the incidence of hypertension and cardiovascular events in nonsleepy patients with obstructive sleep apnea: a randomized controlled trial. *JAMA*. 2012;307(20):2161-2168. doi:10.1001/jama.2012.4366

114. Mokhlesi B, Finn LA, Hagen EW, et al. Obstructive sleep apnea during REM sleep and hypertension. results of the Wisconsin Sleep Cohort. *Am J Respir Crit Care Med*. 2014;190(10):1158-1167. doi:10.1164/rccm.201406-1136OC

115. Roca GQ, Redline S, Claggett B, et al. Sex-Specific Association of Sleep Apnea Severity With Subclinical Myocardial Injury, Ventricular Hypertrophy, and Heart Failure Risk in a Community-Dwelling Cohort: The Atherosclerosis Risk in Communities-Sleep Heart Health Study. *Circulation*. 2015;132(14):1329-1337. doi:10.1161/CIRCULATIONAHA.115.016985

116. Punjabi NM, Newman AB, Young TB, Resnick HE, Sanders MH. Sleep-disordered breathing and cardiovascular disease: an outcome-based definition of hypopneas. *Am J Respir Crit Care Med*. 2008;177(10):1150-1155. doi:10.1164/rccm.200712-1884OC

117. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res*. 2009;18(1):74-84. doi:10.1111/j.1365-2869.2008.00700.x

118. Younes M, Kuna ST, Pack AI, et al. Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice. *J Clin Sleep Med JCSM Off Publ Am Acad Sleep Med*. 2018;14(2):205-213. doi:10.5664/jcsm.6934

119. Mehra R, Benjamin EJ, Shahar E, et al. Association of nocturnal arrhythmias with sleep-disordered breathing: The Sleep Heart Health Study. *Am J Respir Crit Care Med*. 2006;173(8):910-916. doi:10.1164/rccm.200509-1442OC

120. Monahan K, Storfer-Isser A, Mehra R, et al. Triggering of nocturnal arrhythmias by sleep-disordered breathing events. *J Am Coll Cardiol*. 2009;54(19):1797-1804. doi:10.1016/j.jacc.2009.06.038

121. Tung P, Levitzky YS, Wang R, et al. Obstructive and Central Sleep Apnea and the Risk of Incident Atrial Fibrillation in a Community Cohort of Men and Women. *J Am Heart Assoc*. 2017;6(7). doi:10.1161/JAHA.116.004500

122. Gottlieb DJ, Yenokyan G, Newman AB, et al. Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure: the sleep heart health study. *Circulation*. 2010;122(4):352-360. doi:10.1161/CIRCULATIONAHA.109.901801

123. Redline S, Yenokyan G, Gottlieb DJ, et al. Obstructive sleep apnea-hypopnea and incident stroke: the sleep heart health study. *Am J Respir Crit Care Med*. 2010;182(2):269-277. doi:10.1164/rccm.200911-1746OC

124. Chami HA, Resnick HE, Quan SF, Gottlieb DJ. Association of incident cardiovascular disease with progression of sleep-disordered breathing. *Circulation*. 2011;123(12):1280-1286. doi:10.1161/CIRCULATIONAHA.110.974022

125. Thomas RJ, Mietus JE, Peng C-K, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. *Sleep*. 2005;28(9):1151-1161. doi:10.1093/sleep/28.9.1151

126. Thomas RJ, Mietus JE, Peng C-K, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. *Sleep*. 2007;30(12):1756-1769. doi:10.1093/sleep/30.12.1756

127. Ibrahim LH, Jacono FJ, Patel SR, et al. Heritability of abnormalities in cardiopulmonary coupling in sleep apnea: use of an electrocardiogram-based technique. *Sleep*. 2010;33(5):643-646. doi:10.1093/sleep/33.5.643

128. Gagnadoux F, Le Vaillant M, Paris A, et al. Relationship Between OSA Clinical Phenotypes and CPAP Treatment Outcomes. *Chest*. 2016;149(1):288-290. doi:10.1016/j.chest.2015.09.032

129. Penzel T, McNames J, Murray A, de Chazal P, Moody G and Raymond B 2002. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings Med. Biol. Eng. Comput 40 402–7

130. Moody GB, Mark RG, Goldberger A and Penzel T 2000. Stimulating rapid research advances via focused competition: the Computers in Cardiology Challenge 2000 (IEEE) pp 207–10.

131. Langley P, Bowers EJ, Murray A. Principal component analysis as a tool for analyzing beat-to-beat changes in ECG features: application to ECG-derived respiration. *IEEE Trans Biomed Eng. 2010 Apr; 57(4):821-9.*

132. Mazzotti, Diego R et al. “Opportunities for utilizing polysomnography signals to characterize obstructive sleep apnea subtypes and severity.” *Physiological measurement* vol. 39,9 09TR01. 13 Sep. 2018, doi:10.1088/1361-6579/aad5fe

133. *Baguet JP, Hammer L, Lévy P, Pierre H, Launois S, Mallion JM, Pépin JL.* The severity of oxygen desaturation is predictive of carotid wall thickening and plaque occurrence. *Chest. 2005 Nov; 128(5):3407-12.*

134. *Jacobsen JH, Shi L, Mokhlesi B.* Factors associated with excessive daytime sleepiness in patients with severe obstructive sleep apnea. *Sleep Breath. 2013 May; 17(2):629-35.*

135. Cao J, Feng J, Li L, Chen B. Obstructive sleep apnea promotes cancer development and progression: a concise review. *Sleep Breath. 2015 May; 19(2):453-7.*

136. Idiaquez J, Santos I, Santin J, Del Rio R, Iturriaga R. Neurobehavioral and autonomic alterations in adults with obstructive sleep apnea. *Sleep Med. 2014 Nov; 15(11):1319-23.*

137. Azarbarzin A, Sands SA, Taranto-Montemurro L, Vena D, Sofer T, Kim SW, Stone KL, White DP, Wellman A, Redline S. The Sleep Apnea-Specific Hypoxic Burden Predicts Incident Heart Failure. Chest. 2020 Aug;158(2):739-750. doi: 10.1016/j.chest.2020.03.053. Epub 2020 Apr 13. PMID: 32298733; PMCID: PMC7417383.

138. Ali Azarbarzin, Scott A Sands, Katie L Stone, Luigi Taranto-Montemurro, Ludovico Messineo, Philip I Terrill, Sonia Ancoli-Israel, Kristine Ensrud, Shaun Purcell, David P White, Susan Redline, Andrew Wellman, The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study, *European Heart Journal*, Volume 40, Issue 14, 07 April 2019, Pages 1149–1157.

139. Del Rosario ME, Weachter R, Flaker GC. Drug-induced QT prolongation and sudden death. *Mo Med*. 2010;107(1):53-58.

140. Monahan K, Hodges E, Agrawal A, Upender R, Abraham RL. Signal-averaged P wave area increases during respiratory events in patients with paroxysmal atrial fibrillation and obstructive sleep apnea. Sleep Breath. 2019 Dec;23(4):1275-1281. doi: 10.1007/s11325-019-01823-5. Epub 2019 Mar 18. PMID: 30887227.

141. Gula LJ, Krahn AD, Skanes A, Ferguson KA, George C, Yee R, Klein GJ. Heart rate variability in obstructive sleep apnea: a prospective study and frequency domain analysis. Ann Noninvasive Electrocardiol. 2003 Apr;8(2):144-9. doi: 10.1046/j.1542-474x.2003.08209.x. PMID: 12848796; PMCID: PMC6932147.

142. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*. ; 2013:3111–3119.

143. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. ; 2014:1188–1196.