A Diagnostic and Prescriptive Conformal Prediction Framework: Applied to Sleep Disorders

by

Faduma Khalif

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

MASTERS OF ENGINEERING IN COMPUTATION AND COGNITION

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by: Faduma Khalif
Department of Brain and Cognitive Sciences
January 29, 2024

Certified by: Regina Barzilay
Professor of Electrical Engineering and Computer Science, Thesis Supervisor

Accepted by: Sierra Vallin
Graduate Officer, Department of Brain and Cognitive Sciences

A Diagnostic and Prescriptive Conformal Prediction Framework: Applied to Sleep Disorders

by

Faduma Khalif

Submitted to the Department of Brain and Cognitive Sciences
on January 29, 2024 in partial fulfillment of the requirements for the degree of

MASTERS OF ENGINEERING IN COMPUTATION AND COGNITION

ABSTRACT

We propose a novel predictive framework for the future diagnoses and treatments of patients with neurological conditions, specifically patients with sleep disorders, given their clinical history. Via the use of a conformal algorithm with a classifier as its base model, we are able to utilize a patients history of diagnoses, pharmacy dispensing, and other features to produce a set of possible final sleep disorder diagnoses and/or treatments with a definitive level of confidence and bounded level of uncertainty. We also utilize selective classification in order to allow the model to abstain from generating a prediction in cases where the algorithm's predictive confidence does not meet a given confidence threshold, and we further investigate variables that correlate with "abstain" model outcomes. In addition, we experiment with the use of additional machine learning methods such as no-regret learning to better address issues that arise in clinical decision-making. We find that even in cases where there is a limited level of accuracy produced by our base classifier, we are able to use minimal data and selective prediction to establish highly accurate predictive outcomes for certain subsets of our cohort. In developing and testing this framework, we attempt to propose a new standard for predictive algorithms that target clinical-use cases and to better understand uncertainty quantification in a multitude of dimensions.

Thesis supervisor: Regina Barzilay

Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to first thank my advisor, Prof. Regina Barzilay, for many things. For being very committed to truly being a mentor, for making her lab a welcoming space where I felt very seen, for reminding me that research, results and success are not always linear, and finally for exemplifying to me, as an early stage scientist, what the meaning of excellence in science is. I would like to thank my direct supervisor, Prof. Bracha Laufer-Goldshtein, for all of her help on my project, for sharing and allowing me to learn from her great level of expertise in machine learning, for her incisive level of attention to detail and creativity that greatly aided me through the form of very necessary and helpful constructive criticism, and for always kindly and gently coupling her feedback with positive feedback on what went well. I would also like to thank her for kindly continuing to make the time to mentor me after moving across the world to accept a faculty position. I would like to thank my other direct supervisor, Aziz Ayed, for supporting and encouraging me even when faced with roadblocks, for all of his clever suggestions as to how to produce maximally optimal work, and for his virtually unlimited guidance, without which I would not have been able to complete this work. I would like to thank the Takeda team, especially Dana Teltsch, Bilal Khokhar, and Kevin Galinsky (who was not on the team but kindly offered time to help with preprocessing) for all of their insight and help. Finally, I would like to thank my parents for their unending patience, love and emotional support.

Contents

Chapter 1

Introduction

Sleep-wake disorders are life-altering neurological conditions that can be disabling. It can take decades for patients to be properly diagnosed, and these challenges are even further magnified for minorities (for example, women take on average 12 extra years after symptom onset to be diagnosed with narcolepsy) [17]. In short, sleep-wake disorders are not only chronic and very difficult to live with conditions, but they each are also very hard to diagnose as a pathology. Therefore, we propose utilizing conformal prediction methods to aid in this area, so that models may at least suggest differential diagnoses so that they may be fully investigated in a clinical setting.

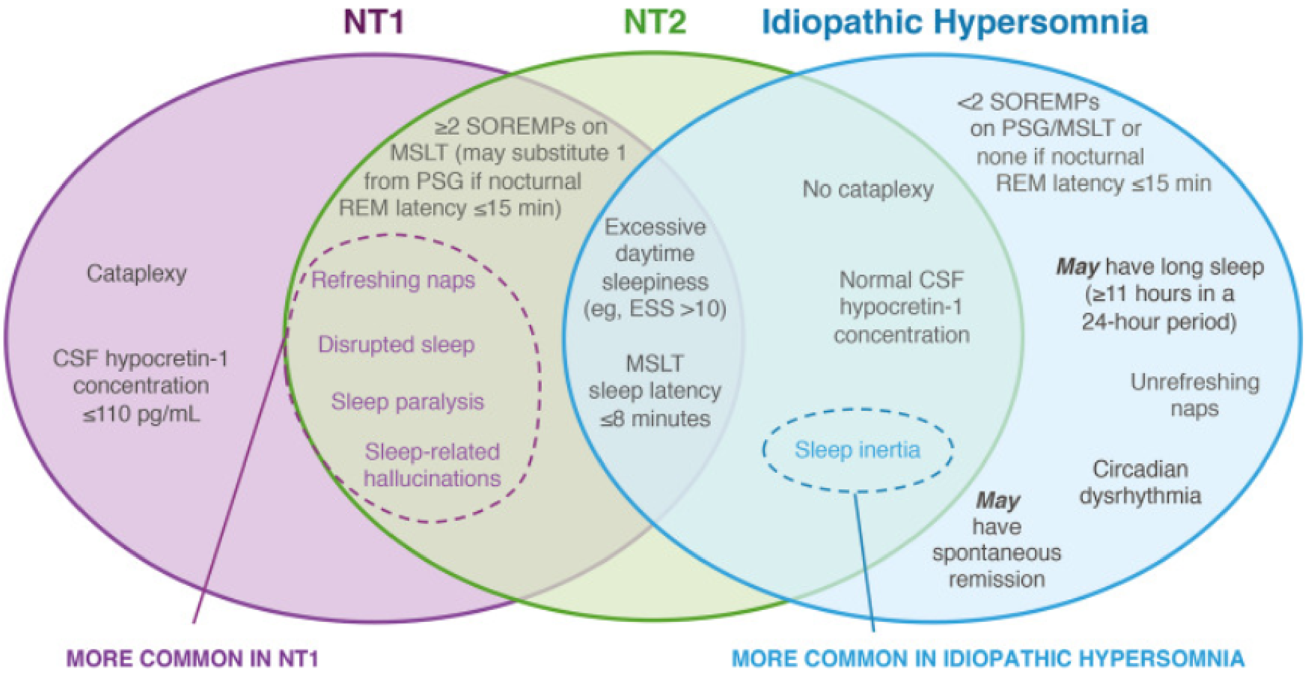
Predicting final outcomes in high risk settings where there is little information to differentiate between final outcomes, or difficulty distinguishing between final outcomes, is a very difficult task that we hypothesize is not fully addressable when only applying classical machine learning classification methods. Furthermore, it is necessary to investigate algorithmic certainty, have some level of control over the predictive confidence of our outputs, and produce "fair" outcomes. We explore this problem in the context of a family of closely related sleep-wake disorders (i.e. narcolepsy types I and II, and idiopathic hypersomnia), specifically their treatments and diagnoses, by utilizing tabular American health care claims data as our primary source for inputs within our framework, in order to develop and test our methods in a simulated clinical condition with what could be considered a high level of risk. We further develop, refine, and apply methods such as conformal prediction, calibrated and selective classification, in order to produce this body of work.

1.1 Sleep Disorders

We focus our application on addressing a specific subset of sleep-wake disorders. Sleep-wake disorders are a class of neurological disorders that manifest as patterns of sleep that deviate from the norm and cause a decrease in the health and quality of life of those affected. Our focus is on sleep disorders that result in an excess of sleep. We focus on the specific final diagnosis of patients with one of three similar sleep-wake disorders: narcolepsy type I, narcolepsy type II, and idiopathic hypersomnia (IH). Narcolepsy type I is a form of narcolepsy that is characterized by the presence of too little hypocretin, the neuropeptide responsible for promoting wakefulness, and/or occurrences of bouts of cataplexy, a condition in which the muscles suddenly weaken while awake. Narcolepsy type II is characterized by a lack of cataplexy. Additionally, the majority of type I narcoleptics experience visual hallucinations while only a minority of type II narcoleptics experience them. Finally, idiopathic hypersomnia is a sleep-wake disorder that presents as extreme drowsiness and/or sleepiness throughout the day regardless of the quality or quantity of sleep previously obtained, and also is often accompanied by difficulties waking up.
[4]

In this work, we set out to develop a highly certain framework that is able to utilize previous claims data to accurately predict the possible set of final diagnoses and/or treatments within a family of diseases.

Figure 1.1: Graphic Visualization of Distinctions Between Sleep Disorders Characterized by Excessive Sleep


NT1
NT2
Idiopathic Hypersomnia
Cataplexy
CSF hypocretin-1 concentration ≤110 pg/mL
≥2 SOREMPs on MSLT (may substitute 1 from PSG if nocturnal REM latency ≤15 min)
Refreshing naps
Disrupted sleep
Sleep paralysis
Sleep-related hallucinations
Excessive daytime sleepiness (eg, ESS >10)
MSLT sleep latency ≤8 minutes
No cataplexy
Normal CSF hypocretin-1 concentration
Sleep inertia
May have spontaneous remission
<2 SOREMPs on PSG/MSLT or none if nocturnal REM latency ≤15 min
May have long sleep (≥11 hours in a 24-hour period)
Unrefreshing naps
Circadian dysrhythmia
MORE COMMON IN NT1
MORE COMMON IN IDIOPATHIC HYPERSOMNIA

1.2 Conformal prediction and Selective Classification

Previous research has shown that predictive machine learning algorithms are capable of facilitating clinically helpful health diagnostics via training on populations of diagnosed patients' claims data. Often times such algorithms produce a single predictive diagnostic label for each patient instance, and these labels predictions are of a relatively high and often unbounded uncertainty. However, such methods are limited in that they generally produce a singular predictive outcome, and do not present predictions with any bounded or "certain" level of confidence. We seek to not only provide differential diagnoses and allow healthcare providers/patients the ability to explore other possible diagnoses and treatments, but also to quantify uncertainty, and produce predictions with a high level of certainty and bounded level of uncertainty. Furthermore, we set out to produce an algorithm that not only is highly confident, but also is able to appropriately abstain from producing a prediction when the desired level of confidence is not possible given a predictive task and base algorithm. Classification models such as k-nearest neighbor-backed classifiers and convolutional neural networks are the current field standard. These algorithms, however, produce limited predictions in that they are, as is, restricted to producing a single target output. We utilize conformal algorithms to output a more generalized and higher certainty prediction, as well as one with a greater accuracy. Conformal algorithms are a class of machine learning processes that make use of conformal prediction. Conformal prediction confers the ability to quantify uncertainty and predictive confidence, and therefore produce a set of solutions or predictive region with a chosen or predetermined level of certainty.

1.3 Research Overview

In the following chapter, we focus on building a base understanding of previous literature that is related to either diagnostic classification and/or clinical use-cases of large language models, conformal prediction, and selective classification in the context of conformal prediction. From there, in chapter 3, we discuss early work revolving around data processing, base classification benchmarking, exploratory test cases, and other tasks intended to initialize the research at hand. In chapter 4, we discuss (the intricacies of and motivations behind) our two main predictive tasks: a patients most recent or “final” diagnosis, as well as possible medical treatments. In chapter 5, we discuss the details of our conformal prediction implementation, and other related augmentations and architectural ideas. Resultantly, in chapter 6, we investigate the utilization of selective classification, and how it may serve both as a key addendum to our predictive framework and as a vehicle for better model interpretability, especially in the context of differing algorithmic confidence outcomes for different demographics and to better investigate algorithmically fair approaches. Lastly, in chapter 7, we draw conclusions from our results and summarize possible directions for future work.

Chapter 2

Relevant work

2.1 Previous work in Predictive Clinical Algorithms

In the work done by Maksabedian Hernandez et al., they devise a predictive framework for the diagnosis of rheumatoid arthritis, an autoimmune condition affecting the joints. Utilizing the MarketScan clinical dataset, they extract a variety of features such as simultaneously prescribed medication, the physician speciality of patients different healthcare providers, length of treatment times, etc. They further distill a patients prescription dispensing history to include whether or not they had taken each of the 500 most prescribed medications among the cohort members. From there, they train a variety of classifiers on medication data as an exposure variable, claims attributes as covariates, and output labels of whether or not a patient has rheumatoid arthritis to develop a pre-trained classifier that is capable of predicting whether or not a patient has rheumatoid arthritis given their pharmacy data. Finally they test and compare the outcomes of performance evaluations of different models on a test set of exposure variables and covariate data. They found that in the end, the best machine learning classifier for their dataset was a random forest classifier, but that it did not necessarily significantly outperform a simple logistic regression method because other evaluation metrics such as the f1 score were equivalent, with random forest having a single digit level of accuracy on average than the logistic regression method [9].

In the work of Rasmy et al, they employ a bidirectional transformer, specifically BERT, that is specially developed from electronic health records (EHR) to deploy in a variety of clinical use-cases. They call their framework med-BERT, with BERT being short for for Bidirectional Encoder Representations from Transformer. After seeing the efficacy of BERT at a multitude of natural language processing applications when trained upon an expansive language data source, it became apparent that perhaps similar success could be achieved in a medical context when training on a large, medicine specific corpus. The authors describe Med-BERT, as "a contextualized embedding model pretrained on a structured EHR dataset of 28,490,650 patients." [11]

2.2 Previous work in Conformal Prediction

In the work of Angelopoulos and Bates, they present the concept and necessity of conformal prediction is specific high risk contexts. The status quo outside of conformal prediction in the area of classification tasks is currently to utilize what they describe as "black box machine learning tools" in order to produce singular, oftentimes "low confidence" predictions for any given problem. This is not helpful in the context of high risk prediction tasks because in such cases, a high level of confidence and bounded level of uncertainty becomes necessary to optimize the outcomes that follow the prediction generation. In order to allow for algorithm outputs that can create a guarantee that, with a chosen probability level, there will be that definitively that level of of confidence (e.g. we can arbitrarily choose, say 10% as our theta value. In this case we can say that there is a 90% chance that the ground truth data solution will be within the given output set of possible classes for a given prediction). Moreover, these prediction sets are described by the authors to be "distribution-free". What is meant by this? In essence, there is no requirement to have access to assumptions regarding either the distributional aspects of the data, nor the model itself [1].

2.3 Previous work in Selective Classification

2.3.1 Selective Classification

In Linusson et al., the authors presented a novel idea with respect to classification result outputs when designing predictive algorithms. They present the idea of a model having the option to "reject" producing a specific outcome or prediction, and facilitate this new architectural aspect via conformal prediction. They define a framework wherein there is a test set, as well as a user-specified input value, which they call "k", that is defined as the maximum amount of errors the model can make on average when producing a as comprehensive of a prediction set as possible. What is of particular importance about their work is the fact that their model is able to "reject" classification and return no prediction (i.e. abstain) in cases where there is too great a level of uncertainty for a given test instance. This process, selective classification, is made possible by ordering the output prediction sets of a given conformal prediction classifier by the amount of confidence the model associates with them, then producing an approximation of the "cumulative error count" of these predictions. The authors describe the benefit of selective classification as making it possible to cap the number of inaccurate predictions produced for a given test set without requiring access to the ground truth labels of any of the members of the test set [8].

2.3.2 Calibrated Selective Classification

While selective classification has many benefits as previously described, it also comes with a few limitations. For one, selective classification as a method places a great amount of stake in the actual level of confidence or uncertainty of a predictive outcome, and this can sometimes cause issues. For example, in the work by Fisch et al., they describe cases in which frameworks that employ selective classification as is produce erroneous predictions such as in cases where an incorrect prediction yields a high measure of confidence, or cases where a correct prediction yields a low level of confidence. However, when calibrated uncertainty estimates of predictions are accounted for as opposed to uncalibrated uncertainty estimate, this can reduce errors that arise as a result confidence levels that may push forth inaccurate predictions or reject accurate ones. The underlying reason as to why this issue occurs is because initial uncertainty estimates can be off-target. A solution to this, then, can be to appraise the actual uncertainty yielded by a prediction and to measure the “uncertainty of an uncertainty”. From there, one can reject uncertainty values that are too uncertain, and prevent unhelpful predictions and rejections that result from faulty uncertainties [6].

2.3.3 Selective Classification with a Fairness Guarantee as Measured through Sufficiently Equal Performance

Although selective classification and related methods may allow for great improvement in the predictive outcomes of a given predictive algorithm, it unfortunately may also hamper the predictive accuracy of specific groups in datasets. In fact, Lee et al. claim that it not only can heighten the level of disparity between already unequal demographics in real life, but also in their level of receiving a valuable predictive outcome (i.e. a non-abstained, confident, and accurate prediction). In response to this issue, they propose a “sufficiency criterion” which is able to help minimize and avoid these predictive inequalities by making sure that when selective classification is employed, it is in such a way that it is able to increase the predictive accuracy for every subgroup in the data set. They also propose a method such that they are also able to avoid imbalances in the precision of the prediction set across groups as well [7].

Chapter 3

Data Exploration and Analysis

In terms of the datasets used, we had access to two different patient health records databases. We utilized the IBM Marketscan research database (IBM Watson Health, www.ibm.com/products/marketscan-research-databases, Armonk, NY, USA), which included medical claims data from insured adults. The data was sourced from a collection of medicaid agencies, health plan records, and employers with sizable insured populations of employees. We also utilized the Optum Clinformatics Data Mart datasets, which includes de-identified longitudinal patient data such as their member data (patient demographic information), their pharmacy and medical claims, their lab test results, their inpatient records, and their provider data.

3.1 Data Pre-Processing

In order to preprocess the data, we first had to manually create mappings of necessary pieces of information to their claims codes if there were not already available or too large and computationally expensive. We mapped things such as medication names with their ndc codes and diagnoses with their corresponding ICD9 and 10 codes. From there, we created SQL calls to query claims databases and create our cohorts of interest. Scripts were written to automatically and procedurally query the databases to find the full claims and record histories of patients within the defined cohorts. These queries were then aggregated into a smaller collection of records that could then be utilized to conduct experiments. We queried tables that included patients diagnostic histories, as well as their prescription claims histories, and their demographic information.

From there, we worked on feature selection. For our work on predicting a final sleep-wake disorder, we initially vectorized patients' chronological history of diagnoses, and used that alone as a starting point. We then utilized a vectorized history of their pharmaceutical claims, and kept a vector that chronologically listed the codes of drugs listed in their claims. When we coupled these features together in this way, we developed a dramatically higher accuracy. We then tried to incorporate a patients history of ordered labs as a feature, with the rationale being that a patients lab history is very descriptive in that it not only can point towards a test that is used to define their condition, but also in that it may also work as a proxy to represent any other conditions that doctors had previously considered diagnosing a patient with, which could lead to a more accurate diagnosis down the line. However, patients' history of labs ordered did not significantly better the accuracy of our models.

While this level of data processing produced a decent level of accuracy, it was still apparent that further processing was necessary. Simply listing a patients prescription history could yield too much noise. One method explored was to take the 500 most commonly used drugs, and create a matrix that included one hot encoded vectors of whether or not each patient had taken any of the 500 most common medications. However, this proved to be too reductive of a method as many patients had not been prescribed most of these 500 and had little continuity between one another within classes. After expanding the most common medication list and increasing the vectors to including the top 1500, and then 2000 most common prescriptions and still seeing too little data, we elected to try to rationally decide what prescriptions led to the most significant impact in learning for our models.

3.2 Sample Cohort Building

Before working on the final dataset and target prediction tasks, we engaged in an exploratory task to develop a highly generalizable framework. We first worked to recreate the cohort used in the Maksabedian et al. paper and conducted some analysis on it as a toy example. We delineate the steps of the cohort creation process below and summarize our findings.

First, after analyzing the cohort description in the Maksabedian et al. paper, a set of inclusion and exclusion criteria were outlined. The inclusion criteria required us to account for all patients who had a documented claim for Etanercept, which is a tumor necrosis factor inhibiting drug that is commonly prescribed to treat a variety of autoimmune conditions such as rheumatoid arthritis, plaque psoriasis, and ankylosing spondylitis. Further inclusion criteria included patients being between the ages of 18 and 64, and having claims/being enrolled during the time period starting 1/1/2010 to 12/31/2019.

The cohort build specifications included utilizing a patients first claim of Etanercept date as their index date, in order to only include those who were enrolled in plans within the marketscan database network 6 months prior to their index date, and continued to be enrolled for at least another 12 months after their index date. This would allow us to develop a follow-up period that captured the 12 months after a patients index date.

From there, we highlighted demographics and outcomes of interest in order to compare the makeup of the cohort extracted via our python framework from the marketscan database and that of the cohort built via an external SAS procedure to act as a “sanity check” and allow us to know how to navigate cohort design and extraction for our actual problems and cohorts of interest.

The specific demographics and outcomes we measured and compared included 1) the general number of patients within the cohort 2) the ratio of men to women in the cohort 3) the mean age and standard deviation of the cohort and 4) counts of patients with inpatient admissions of heart failure, outpatient claims of hypertension, and inpatient claims of infusion devices into their superior vena cava.

This outcomes of this endeavor (toy problem cohort design and validation) proved helpful in a couple of ways. For one, it allowed for the development of a streamlined cohort building framework that could be re-used in part for the development of the sleep-wake disorder cohort of interest. Second, and perhaps more importantly, it revealed limitations in our cohort extraction from the Marketscan database at an early stage in our research.

IBM describes MarketScan as containing "de-identified records for more than 273 million patients since 1995." It included multiple data tables such as summaries of inpatient admissions, inpatient services, outpatient pharmaceutical claims, and facility header information (Segal, 2016). We focused, for our initial prediction tasks, on using data from patients enrollment files and commercial pharmaceutical claims, and we also used data from inpatient admission records to conduct sanity checks and compare outcomes.

Because the database is exceedingly large, consisting of billions of patient records, we faced an early roadblock when some of the extraction and data cleaning methods we initially employed led to losing critical pieces of data. For example, we utilized the DuckDB python package (https://duckdb.org/why_duckdb) to have access to a tool that could assist in analytical querying and allow for more efficient cohort extraction. However, it began to fail on larger files and would abridge. After this we attempted to move the MarketScan data to local PostgreSQL databases, but because the Marketscan files included irregular data representations and could not be manually corrected due to data access limitations such as downloading the data, we temporarily paused this method as well. After this, we attempted to divide the larger marketscan files into smaller CSV files to make them a suitable size for data loading processes down the line. However, to divide them into files of our target size, we realized it would require so many files that it would overwhelm the linux directories. In order to address this issue, we initiate a process such that we manually re-organize our directories and migrate our data to a different database platform. The solution we concluded with included utilizing SQL to efficiently pull large portions of the cohorts defined in the

Optum datasets into CSV files, and combine them to produce sufficiently large cohort sample sizes of 10,000+ patients for training and testing exercises. We start by building an initial overall cohort of 100,000+ sleep-wake disorder diagnosed patients. After developing smaller testing experimental suites of approximately 10,000 patients, we can see bodies of diagnostic histories of 5+ million claim records, before we filter through these to de-bias the dataset and throw away data from after a patients target prediction time.

3.3 Comparing ML base models

After building the cohort, we engaged in replicating an exploratory classification task as described in the Maksabedian et al paper to see what methods and tools would be of the greatest use when applied to our actual cohort of interest and the corresponding prediction tasks. We focused on trying to predict whether or not the cohort of TNF inhibitor treated patients did or did not have a rheumatoid arthritis diagnosis. As previously described, we created a feature matrix that consisted of a variety of pharmaceutical claims information about patients, but which highlighted whether or not patients had taken any of the top 500 most commonly prescribed drugs within the cohort via one hot encoded vectors. The paper summarized the outcomes of utilizing different modeling methods and found that logistic regression alone was ideal for their application, but that decision trees and basic gradient boosting algorithms provided high accuracy outcomes as well. After experimenting with similar models on our replicated problem example problem, we developed a basic cohort extraction process, feature matrix building framework, and an understanding of what classification model options existed and might be of the best performance for our later work [9].

Chapter 4

Predictive Tasks

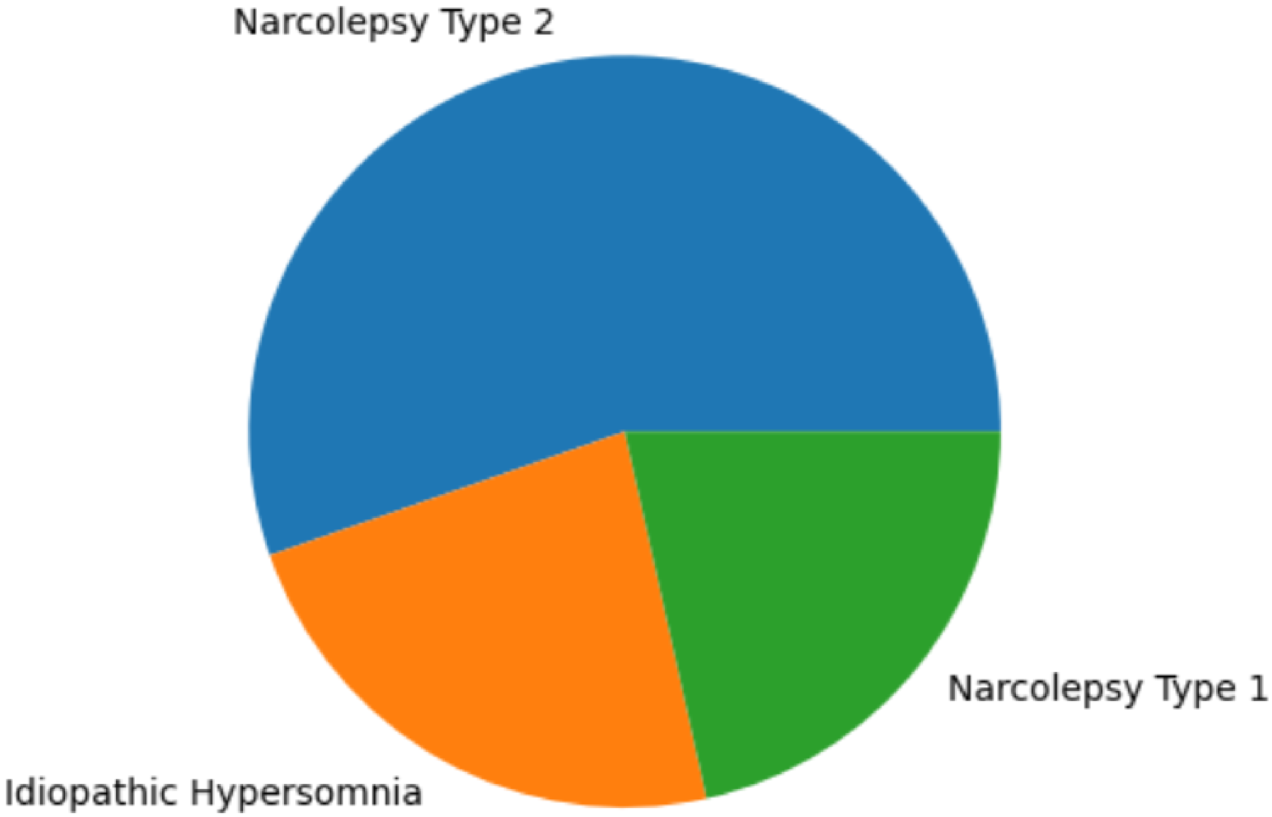
In this work, we focus on two different prediction tasks. We describe the goals of these prediction tasks, their necessity and inherency in real world use of our machine learning framework, and why they are well suited to be the research problems of choice.

4.1 Final Diagnosis Prediction

The first predictive task we engage in is predicting the final (most recent) sleep disorder diagnosis that a patient is given within the time frame of the medical records database that we have. We define a three class classification task wherein a patient can either be classified as receiving a final diagnosis of narcolepsy type 1, narcolepsy type 2, or idiopathic hypersomnia. From there, we experiment with a variety of predictive frameworks to produce a highly fair, highly confident, and highly accurate estimate of what diagnosis or diagnoses a patient may have. This is a very pertinent problem in real world clinical settings for a variety of reasons. For one, sleep disorders can pose a serious challenge for doctors to attempt to accurately diagnose, and patients can often deal with decades of symptoms without the right diagnosis. Moreover, narcolepsy type 1, despite being the most common form of narcolepsy and oftentimes being accompanied by more severe symptoms, is especially difficult to diagnose and patients suffering from this illness may take even longer to diagnose.

There also exist representational issues in American claims data regarding narcolepsy type 1, and we elaborate on these more later.

Figure 4.1: Proportion of Patients in a Generated Cohort with Each Final Sleep-Wake Disorder Diagnosis



This can lead to patients experiencing unnecessary anxiety and also to unnecessary strain in both their social and professional relationships [16]. Furthermore, certain demographic groups may experience even greater diagnostic delays. For example, despite the fact that men and women with narcolepsy present with virtually the same symptoms, men have a 53% higher chance of being accurately diagnosed than women at any time after the start of their symptoms. In a study by Won et al, they found that it took about 16 years after the start of their symptoms for 85% of men with narcolepsy to be accurately diagnosed, whereas it would take about 28 years for women to reach that collective level of accurate diagnosis [17]. It's clear then that there exists a need for a framework that can predict a patients possible diagnosis with a high level of confidence, and that may be of help in hastening accurate, confident, and fair diagnoses for patients, especially those who are afflicted with diagnostically challenging conditions such as narcolepsy type 1. We further present ways in which these issues can be mitigated via the application of a conformal predictive framework with capabilities such as calibrated and fair selective classification.

4.2 Treatment Plan Prediction

The second predictive task we focus on is predicting what treatments might be prescribed to patients given their diagnostic history. More specifically, we attempt to predict what classes of medication or specific medications are prescribed to patients using a conformal predictive algorithm. In order to define and solve this problem, we reduce it to a multiclass classification task. Predicting a set of possible medications or medication types that could be prescribed to a patient with a high level of confidence is helpful for a variety of reasons. For one, prescribing patients a treatment, even with a definite diagnosis, is a high risk clinical task. As such, predictive models that produce highly confident results with bounded levels of uncertainty can be of great help. Furthermore, in the context of sleep disorders such as narcolepsy, there are a variety of issues that can arise as a result of the pharmacological interventions used to treat narcolepsy patients. For one, certain medications can lack efficiency and cause patients to still suffer from symptoms of their sleep disorder. Additional concerns may include debilitating side effects or a lack of responsiveness to the treatment due to a new level of tolerance. Finally, in recent years there has been a marked increase in the level of understanding of the actual biological mechanisms of some wake-sleep disorders such as narcolepsy type 1. Due to this, there is an increase in candidate medications that are biologically implicated in treating narcolepsy but perhaps are less common or standardly prescribed. Given all of these clinical limitations, a precise list of candidate medications or treatments for patients who suffer from sleep disorders that is generated via conformal prediction may be of great help as it can mitigate less commonly prescribed medications from being overlooked when more commonly prescribed medications are associated with a lower level of certainty. Additionally, a precise set of predictions may allow for the suggestion of other treatment options that also yield high levels of confidence when patients may fail an initial treatment due to issues such as side effects, lack of effectiveness, etc [18].

Chapter 5

Conformal Prediction

Here we discuss in detail the mathematical basis of and rationale as to how conformal prediction is set up and works. We will then lead you through the process of conformal prediction using an example from a predictive problem addressed in this thesis, and summarize the outcome of our findings.

5.1 Basis of Conformal Prediction

The initial step is formulating a question or predictive problem in which conformal prediction is possible, and to ideally formulate a question within a context that would be particularly well addressed by conformal prediction. For example, a simple binary classification task alone would not be well suited to conformal prediction. This is because conformal prediction widens the possible prediction set when statistically necessary. That is to say, when the initial prediction for a given instance does not meet the given confidence or certainty threshold, the prediction set will expand to include other probable classification labels, until the certainty threshold is met, in order to maximize predictive precision while maintaining a specific level of predictive confidence. In the case of a task such as binary classification, however, expansion of the prediction set would lead to an imprecise prediction – the entire possible classification set would be the output, and very little meaningful information would be relayed unless other aspects of the model or question were addressed. However, in the case of a multi-class classification task, presenting a few labels as part of the prediction set can still allow for a precise and meaningful predictive outcome, while maintaining predictive confidence, and can allow for a narrowed down "checklist" of possibilities provided by the algorithm to allow further investigation whether it be for model interpretability purposes or to thoroughly investigate all likely possible cases for a high risk issue such as clinical diagnoses [1].

The next step is to then develop a trained model for our given dataset and well-chosen/crafted predictive problem. We describe this process at length and how we did this for our data and a couple of predictive problems that we work through in chapter 3 section 3. From here on out, we can refer to this model as h, or, more specifically in our case, as our base classifier or model.

After this point, we begin the calibration step. In the calibration step, we utilize data that we partition away from our general training and testing set as our "calibration data". This data allows us to develop uncertainty and confidence values through the calculation of error probabilities. In doing so, we are able to utilize this information to develop prediction sets that can be output by the model. These prediction sets are sets of predictive labels that are possible predictions and collectively are able to cover enough confidence that the model is highly certain that an accurate prediction exists among these values [1].

Let's illustrate this concept in further detail and with an example: Our model input is the vectorized claims data that we extracted and processed for each patient, and each of these patients is a member of one of some number, N, of classes that we have divided them into. In the work we do, we partition patients into groups based on their final sleep-wake disorder diagnosis in order to predict patients final diagnoses, and also into groups based on the treatments they are prescribed [1].

From there, we have our base model, h, which would calculate an approximate probability value for each of the given classes. In the work of Angelopoulos and Bates, we can see that this is described as:

$$h(x) \in [0,1]\text{\^{}N}$$

[1]

After this, we can take our specially partitioned calibration data (we save 1000 patients for this), which consists of both patient input vectors and accurate class labels (dependent on the problem, either their final diagnosis or treatment class) and was previously “unseen” by the model, i.e. our hold out data, and we can utilize this body of data to calibrate the model as our calibration set. This calibration set is useful in that we can take our base model, h, and the calibration data together to produce prediction set outputs that meet the given condition, as also defined in the work of Angelopoulos and Bates:

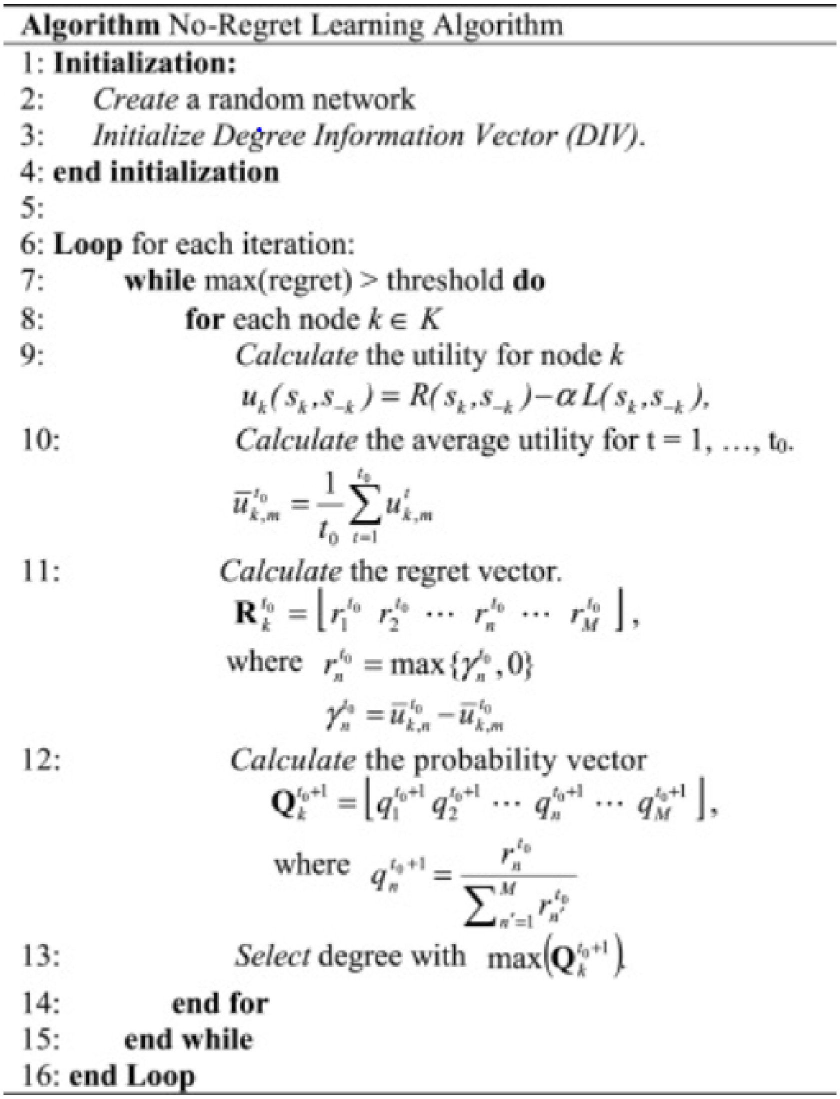
$$1 - \alpha \leq \Pr[\mathrm{Y}_{\mathrm{val}} \in \mathrm{C}(\mathrm{X}_{\mathrm{val}}))\ ] \leq 1 - \alpha + 1/(\mathrm{i}+1)$$

[1]

To clarify, here we use α, our alpha value, to represent the user defined error rate, which can be a decimal anywhere between 0 and 1. i is the size of our set. This value tells us how lenient or stringent our conformal algorithm will be by defining what level of uncertainty we will “accept”, and is very similar to the use and function of values in conventional hypothesis testing. In practice, this means that, 1 - α is approximately equivalent to the odds that the correct prediction is contained within our prediction set. When we apply this concept/method to our examples, we experiment with values of .1, .05, and .10, in order to produce prediction sets that include accurate labels 99, 95, and 90% of the time. The conventional term for this phenomena is marginal coverage, since this is the level of accuracy covered by our marginal probability [1].

5.2 No-Regret Learning

In game theory, there exists this concept of No-regret. To understand the concept of no-regret, and all resulting ideas and methods such as no-regret learning as well as more generally no-regret learning algorithms, we must first define regret. If you pause a game and look at a given time point, the “regret” accumulated by an actor is the difference between the benefit of the method they employ and an alternative method that they could have employed, with the assumption that surrounding actors also employ the alternative method. It makes sense, then, that a successful algorithmic strategy could be to minimize regret. When an algorithm is able to successfully minimize regret to the point of all alternative strategies producing more regret than those elected by the algorithm, then the model is referred to as a no-regret algorithm. One benefit of no-regret learning is that it needs very little information relative to other similar learning algorithms, and we will illustrate below the basis of no-regret learning via an algorithm defined by Sohn et al [12][14].



Algorithm No-Regret Learning Algorithm

```
Initialization:
    Create a random network
    Initialize Degree Information Vector (DIV).
end initialization

Loop for each iteration:
    while max(regret) > threshold do
        for each node $k \in K$
            Calculate the utility for node $k$
                $u_k(s_k, s_{-k}) = R(s_k, s_{-k}) - \alpha L(s_k, s_{-k})$,
            Calculate the average utility for t = 1, …, $t_0$.
                $\bar{u}^{t_0}_{k,m} = \frac{1}{t_0}\sum_{t=1}^{t_0} u^{t}_{k,m}$
            Calculate the regret vector.
                $\mathbf{R}^{t_0}_{k} = \left[ r_1^{t_0}\ r_2^{t_0}\ \cdots\ r_n^{t_0}\ \cdots\ r_M^{t_0} \right]$,
                where $r_n^{t_0} = \max\{\gamma_n^{t_0}, 0\}$
                      $\gamma_n^{t_0} = \bar{u}^{t_0}_{k,n} - \bar{u}^{t_0}_{k,m}$
            Calculate the probability vector
                $\mathbf{Q}^{t_0+1}_{k} = \left[ q_1^{t_0+1}\ q_2^{t_0+1}\ \cdots\ q_n^{t_0+1}\ \cdots\ q_M^{t_0+1} \right]$,
                where $q_n^{t_0+1} = \frac{r_n^{t_0}}{\sum_{n'=1}^{M} r_{n'}^{t_0}}$
            Select degree with $\max(\mathbf{Q}^{t_0+1}_{k})$
        end for
    end while
end Loop
```

[14] We integrated an aspect of no-regret learning into our initial model architectures to better optimize the classification abilities of model. For example, we are able to preliminarily decide what k value to choose, among other parameters, when we utilize a K-nearest neighbors (KNN) classifier via minimizing the regret of other possible k-value options. Experimentally, this yielded for us better algorithmic outcomes than rationally deciding on parameter values. No-regret learning is especially beneficial in high risk clinical settings and for predictive tasks similar to the ones that we address as it allows for a regulated decision making paradigm that weighs alternative routes and tries to minimize algorithmic regret, thereby increasing the likelihood of satisfactory clinical outcomes. For our final algorithmic

design, we retired the use of KNN classifiers as our base model as there existed better base model candidates after iterating through larger cohorts of patients, more diverse feature sets, and different feature representations, but we still recognize the benefit of coupling this no-regret learning method with an overall conformal algorithm.

5.3 Base Models

Below, we describe the structure and basis of the base classifiers that we used primarily for our final outcomes.

5.3.1 K-Nearest Neighbors Classifier

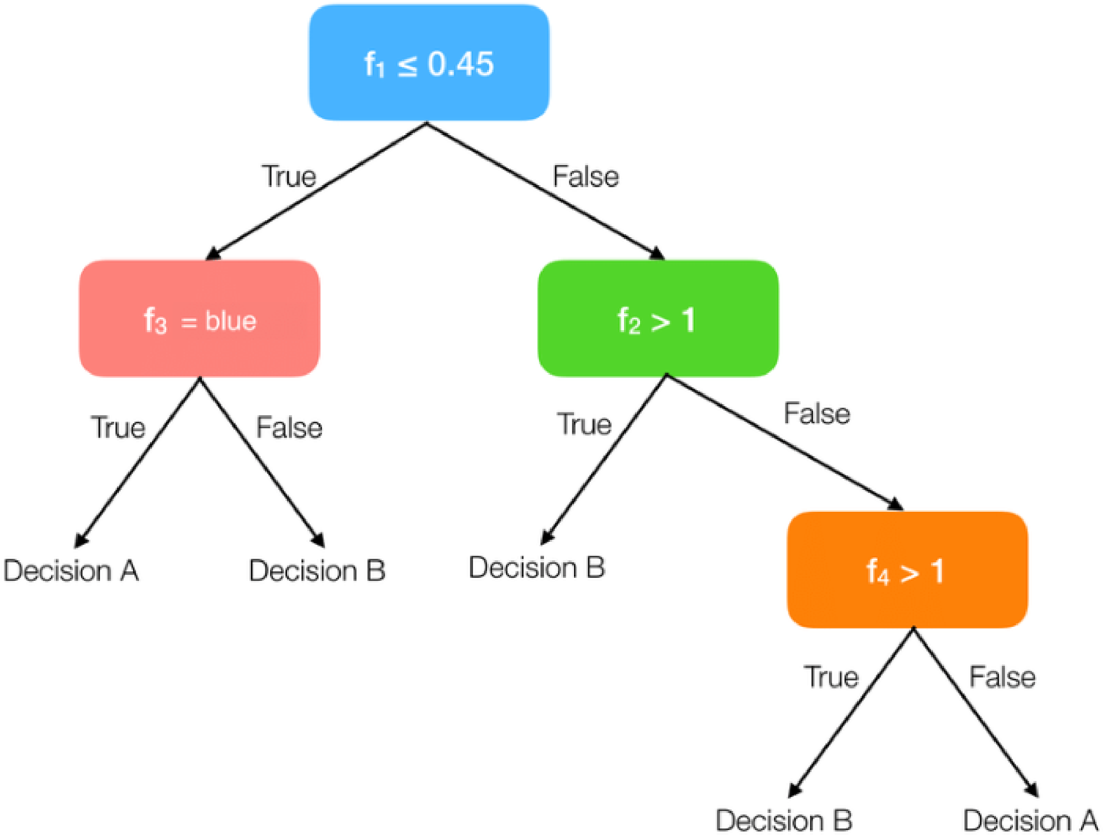
One of the base fitted models we experimented with was Sci-Kit Learn's K-Nearest Neighbors (KNN) classifier. In order to simply illustrate the way a binary classification task might be conducted using a KNN classifier, we can envision a a two dimensional grid with two distinct clusters, and with each cluster containing points that belong to one of the two classes due to their positioning in either cluster one or two, and each axis represents a given feature used to classify each point into one of the two groups. The KNN algorithm, for which the KNN classifier is named and based upon, can help us continually classify points in this system by answering the question: If we have an unlabelled point that lies between the clusters, how are we to label it? The first step and basis of a KNN classifier is calculating the Euclidean distance (although some KNN applications utilize Manhattan distance) between the sample that is to be classified and the other points, which have pre-existing class labels [19].

From there, once the distances between the unclassified sample and the other labeled points are calculated, the most common label of the "nearest neighbor(s)" is assigned to the unclassified sample. The number of nearest neighbors considered is equal to a usually pre-defined "K" value, hence the name of the algorithm: "K-nearest neighbors". If K is equal to one, then the label of the sample to be classified will be the same as and rely only on the label of the single nearest neighbor. If K is equal to three, then the unclassified sample will then be classified as being of the same class as the majority of it's three neighbors [19].

5.3.2 Decision Tree Classifier

A further classifier that we utilized as a base classifier for the majority of the work was Sci-Kit Learn's Decision Tree Classifier. The way a decision tree classifier is trained is that, once presented with a some labeled training data, a "decision tree" is developed.

Figure 5.1: An Example Decision Tree to Illustrate the Structure of a Decision Tree Classifier



[10]

As can be seen in the graphic by Mollas et al., there are root and leaf nodes. The decision tree is constructed as the algorithm passes through labeled data with various conditions regarding the feature to split remaining groups of the labeled data until each leaf node contains training samples that belong to the same class. Once the training data has been partitioned by way of the conditions in each root node and fully split, the unlabeled data is passed through the root node conditions until they reach a leaf node, and the class labels of that leaf node are the final predicted class label for the data [15].

5.3.3 Gradient Boosting Classifiers

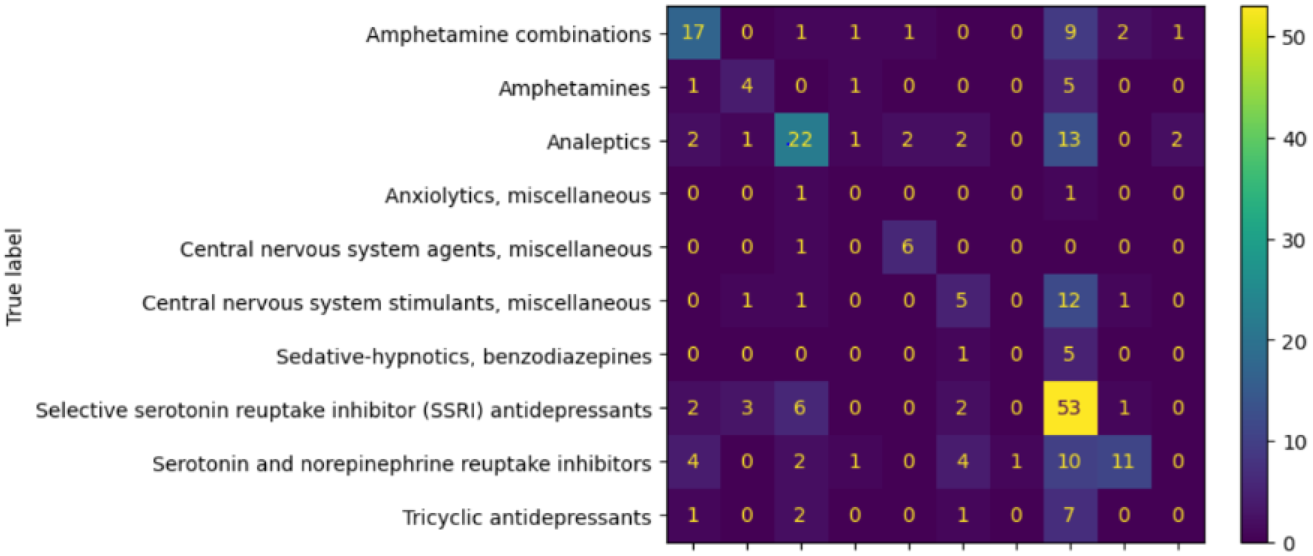
One of the initial forms of classifiers we applied and tested were a few variants of Gradient Boosting Classifiers. These types of classifiers have recently become and are currently very commonly utilized in contexts with similar predictive problems (i.e. large scale diagnostic prediction).

As previously described, there are a number of reasons as to why a gradient boosting classifier might be the best choice for a number of these predictive tasks. The first gradient boosting classifier that we tested, XGBoost Classifier, a classification algorithm backed by the XGBoost library, was significantly more efficient than the other classification algorithms it was tested alongside due to not only an in-built threading functionality that allowed for a quick and relatively optimal processing of such large amounts of claims data, but also due to the general logic and implementation of gradient boosting classification architectures [2]. Our first application of the XGBoost classifier was during the initial exploratory prediction work we conducted on the very large marketscan dataset. Further gradient boosting classifiers we tested, in particular in the context of predicting future sleep-disorder diagnoses and treatments, included Sci-Kit Learn's gradient boosting classifier and AdaBoost classifier.

5.3.4 Preliminary Accuracy Results

When we look at the outcomes of the accuracy of just the base classifiers, we can see that without using uncertainty quantification methods, our outcomes are somewhat useful, but are significantly less certain and accurate than the methods that we later employ.

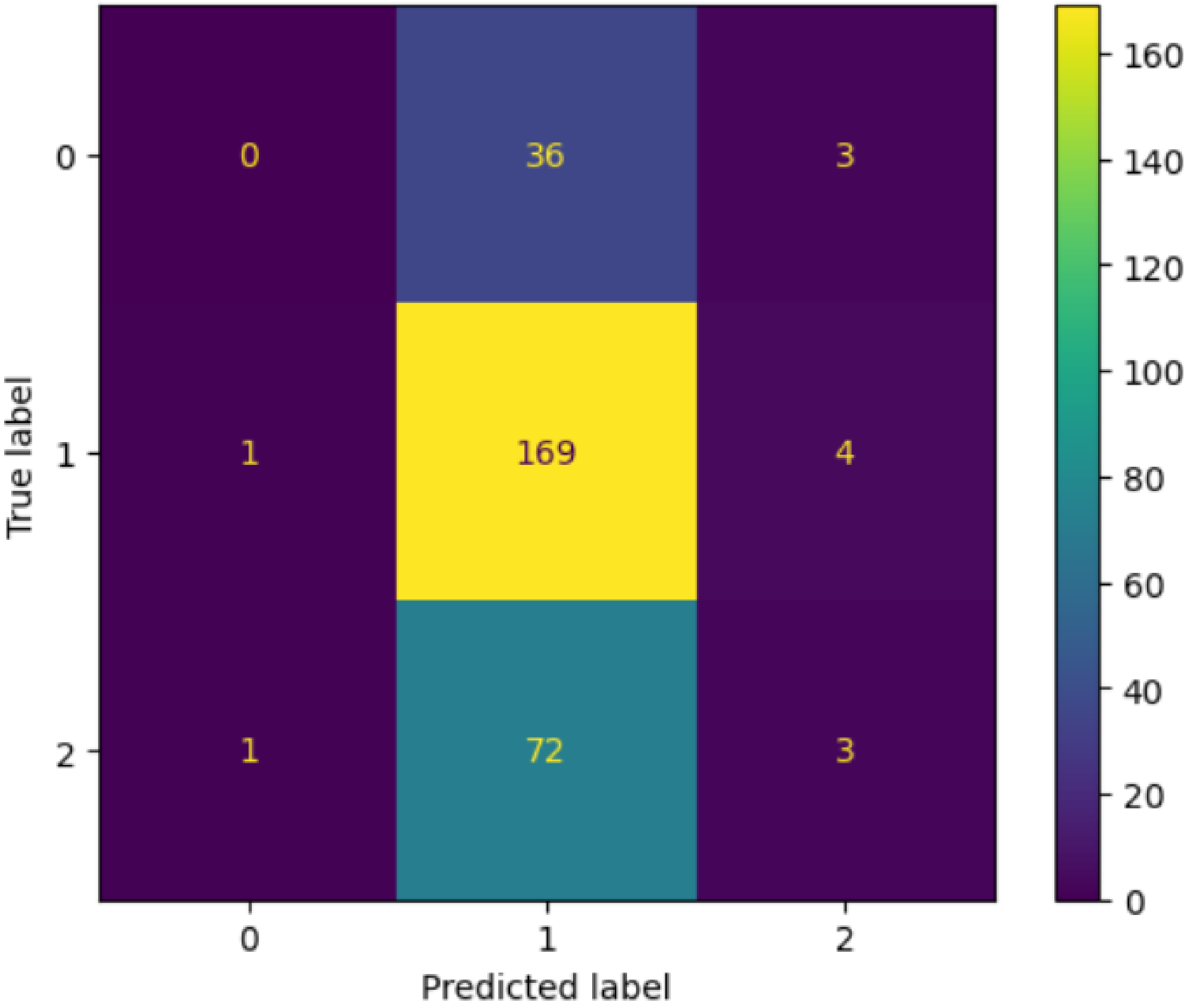
Figure 5.2: Sample Confusion Matrix of Final Chronological Drug Prescription of Patients within the Cohort



[The horizontal axis of predicted labels corresponds to and directly follows the pattern and order of the vertical axis.]

In this figure, we see a significantly improved set up in our base classification model, as we can see that there is a well populated diagonal, with much fewer off-diagonal predictions causing noise, except in the case of SSRIs prescriptions due to how relatively common and ubiquitous they are to prescribe.

Figure 5.3: Sample Confusion Matrix of Final Specific Sleep-Wake Disorder Diagnosis of Patients within the Cohort

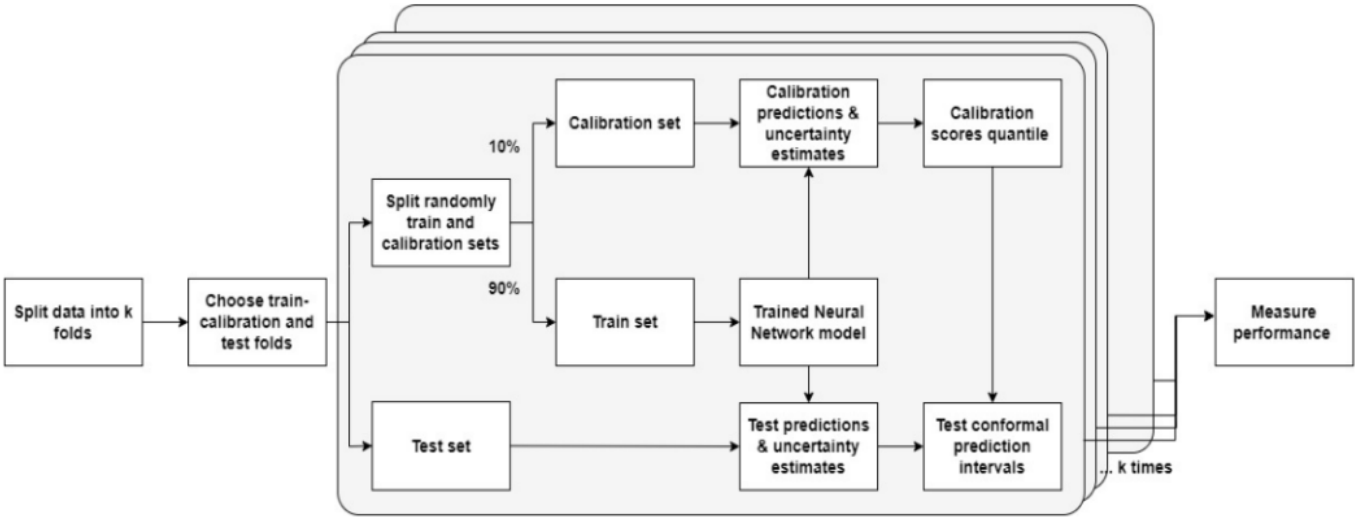


[A label of 0 corresponds to Narcolepsy type 1, 1 to Narcolepsy type 2, and 2 to idiopathic hypersomnia.]

We see in this figure of an initial confusion matrix for our base classifier, that classical approaches to classifying these sorts of sleep-wake disorders over time tend to produce outputs that echo the clinical struggle of “To lump or not to lump” (i.e. whether or not clinicians are to distinguish between such similar disorders, particularly narcolepsy type 2 and idiopathic hypersomnia).

5.4 Model Architecture

Figure 5.4: Diagram of a Conformal Prediction Algorithm


Split data into k folds
Choose train-calibration and test folds
Split randomly train and calibration sets
10%
90%
Calibration set
Train set
Test set
Calibration predictions & uncertainty estimates
Trained Neural Network model
Test predictions & uncertainty estimates
Calibration scores quantile
Test conformal prediction intervals
... k times
Measure performance

[5]

In the figure above, we see an illustration of the general set up of a conformal prediction algorithm. In the sections below, we elaborate on our specific framework and application.

5.4.1 Base Model

We trained, fit, and assessed numerous forms of classifiers via the processes described previously in order to arrive upon an optimal base classifier to incorporate into the architecture of our conformal prediction architecture. Some examples of classifiers that we experimented with included KNN models, a variety of ensemble methods such as gradient boosting classifiers and random forests, support vector classifiers, etc. The base model takes in the medical claims data of patients in order to produce several things of value for the rest of our pipeline and model architecture. It produces, for one, a set of possible final classifications or classes which we can later utilize to produce a final conformal prediction set for each patient. The base model also provides us with, when applied to the calibration data, a "conformal score", which produces prediction probability distribution outputs of the base model for each input sample (and we can get this by just returning the softmax function applied to our base classifier fitted to the calibration data). We can then take these calculated conformal scores and use them to calculate adjusted quantile values. From here, we can take these quantile values, and project them onto our prediction class probability scores, and use this to retrieve the set of probable classifications within a the user-defined confidence range [1].

5.4.2 Calibration Step

In the "Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification" presented by Angelopoulos and Bates, they find that, in most cases, a calibration set size of about 1000 is good for the purpose of generating conformal predictions, and this is the calibration set size we employ in our work. They elaborate on this and say "the coverage of conformal prediction conditionally on the calibration set is a random quantity.“ This is because, regardless of the size of the calibration set, the "coverage guarantee" of a conformal prediction algorithm persists. Essentially, the calibration set size does not impact the fact that coverage provided by the prediction sets that we generate will always be, at a minimum, 1 - α [1].

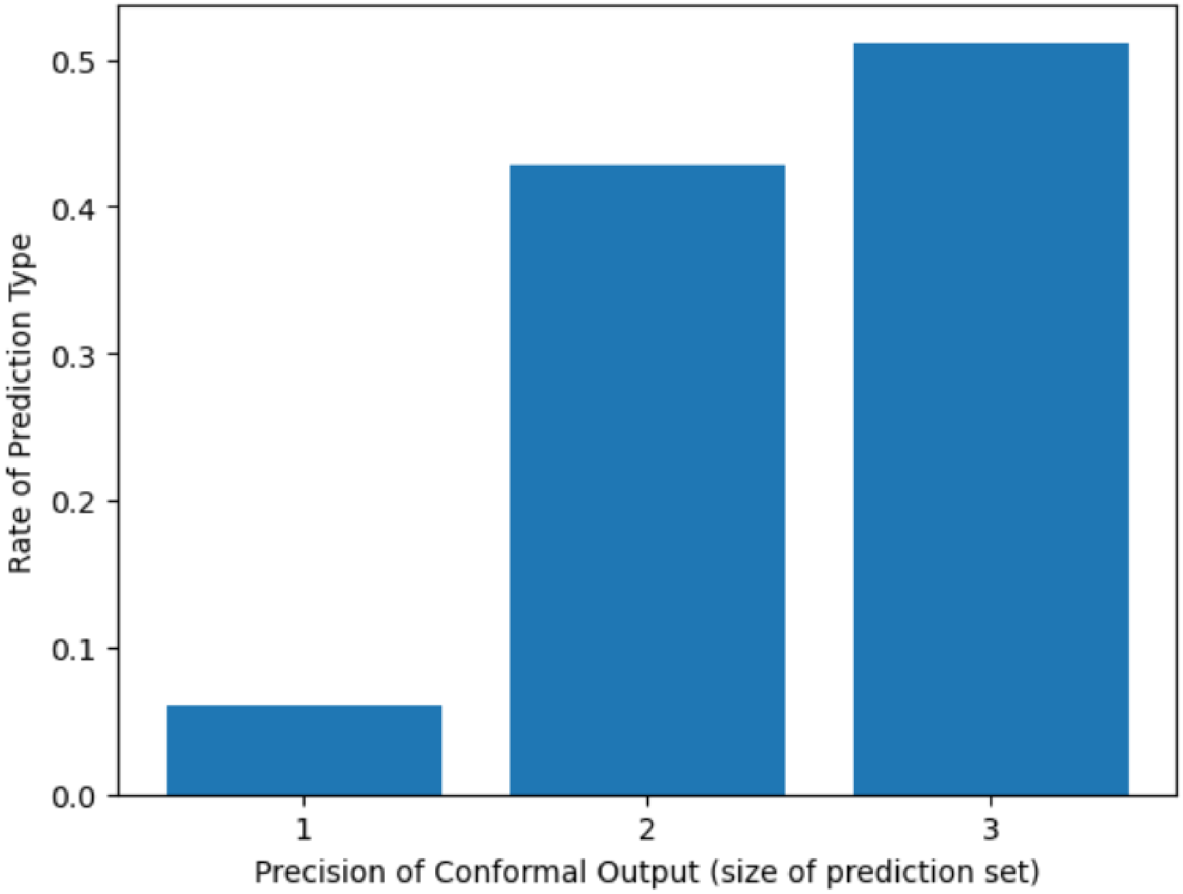
5.5 Model Features

Features we included were things such as patients diagnostic history prior to their final sleep-wake disorder diagnosis, as well as their previous prescription history. As previously mentioned, we included patients lab histories as well, but found that these records generally had no effect on the accuracy of our architecture. We created time ordered diagnosis histories, and combined these vectors of record histories with a vector that indicated whether or not a patient had been prescribed on of the top 500 drugs of interest without already having had the diagnosis that we intended to predict. Finally, we included one-hot encoded vectors of the race and sex demographic information of patients.

5.6 Results

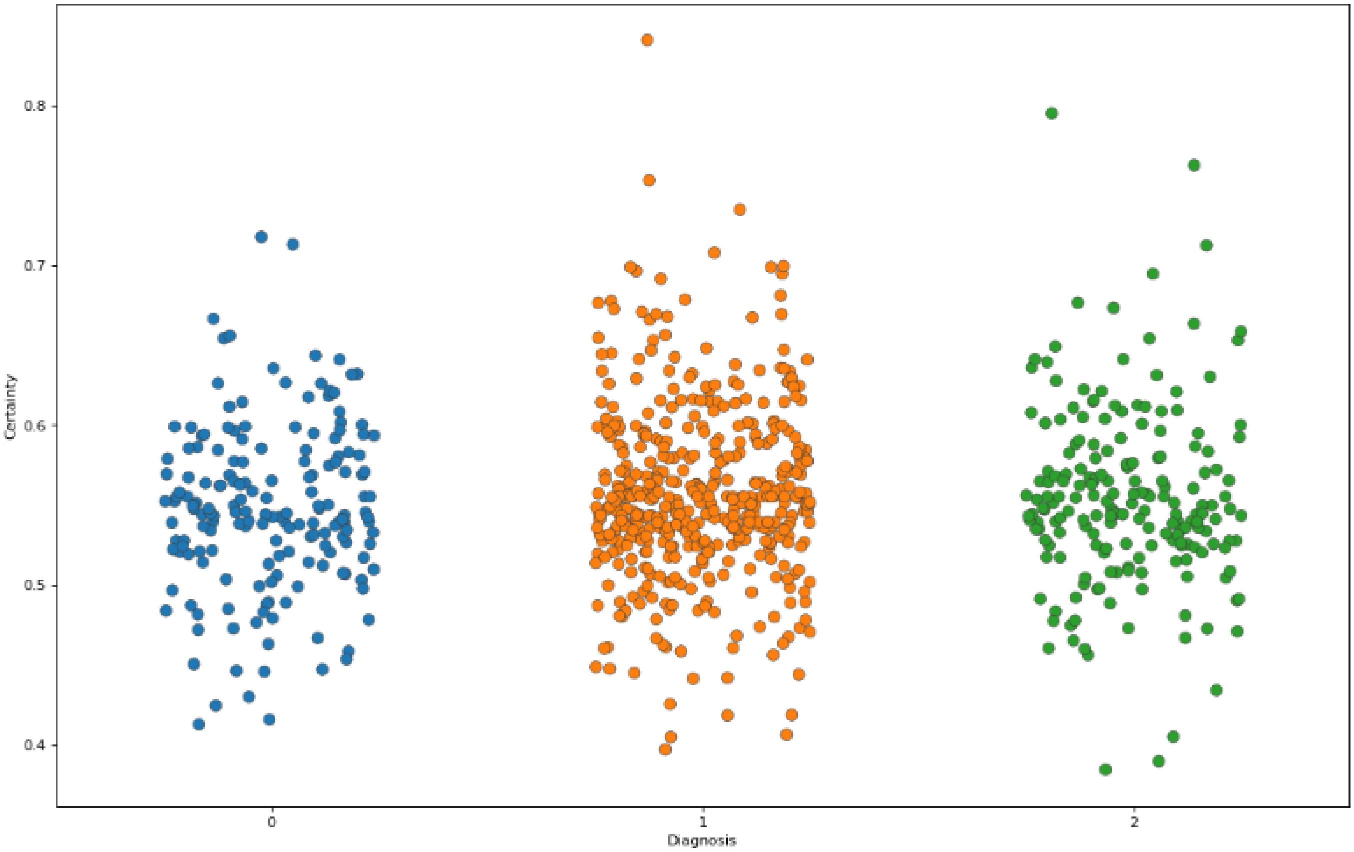
In this section, we go over the outcomes of our conformal prediction methods. Preliminarily: we find that we get an accuracy outcome that is lower bounded by 1 - α when we apply our conformal prediction paradigm, which is exactly in line with what would be expected of a conformal prediction algorithm, as described in the work of Angelopoulos and Bates, among other pieces of literature describing conformal prediction.

Figure 5.5: Precision Plot of Conformal Prediction Outcomes



In this plot, we graph the breakdown of the conformal prediction output set sizes to gain better insight into the ability of the conformal prediction algorithm. The graph of these outcomes tells us a few things: for one, there definitely exists a trade-off – for a highly confident prediction, some amount of precision must be lost, especially if it is difficult to classically solve this sort of prediction question. However, we also see that more confident prediction necessitates further exploring other less certain predictions, which in a clinical setting could translate to patients and doctors considering critical diagnoses that may have been skipped over otherwise, especially when they are simply just less likely, but still necessary to consider in order to reach a certain level of predictive confidence. We also see that, for about half the data set, it is possible to eliminate one of the three very similar sleep disorders, with information from their healthcare claims records alone. In a clinical setting, it can be of great benefit to both doctors and patients to be able to confidently rule out specific diagnoses just via the use of some claims data, as opposed to waiting possibly years for a symptom to arise or struggle to find a label while waiting to see if they can make a diagnosis by exclusion such as idiopathic hypersomnia.

Figure 5.6: Graphical Representation of Sample Algorithmic Confidence of Samples Organized by Ground Truth Final Sleep Wake Disorder Diagnosis



[Narcolepsy type 1 is the blue leftmost cluster, Narcolepsy type 2 is labeled is the orange middle cluster, and Idiopathic Hypersomnia is the green rightmost.]

We can see in this figure, that there exists a great variation in the level of certainty or confidence that the model expresses in predicting a singular "most likely" classification for each of the three given classes. It's also worth noting the differences in distribution of the certainty of the model for each class of diagnosis. As expected, the model was more confident when predicting a ground truth narcolepsy type 2 final diagnoses and less so when predicting a ground truth narcolepsy type 1 final diagnoses. Unexpectedly, and fortunately as a result of efforts to produce better predictive outcomes such as more specific calibration calculations, noise reduction of the initial feature set, etc, we were able to see somewhat similar levels of confidence between the different classes as a result of these efforts, as opposed to more dramatically unequal predictive outcomes.

Chapter 6

Selective Classification

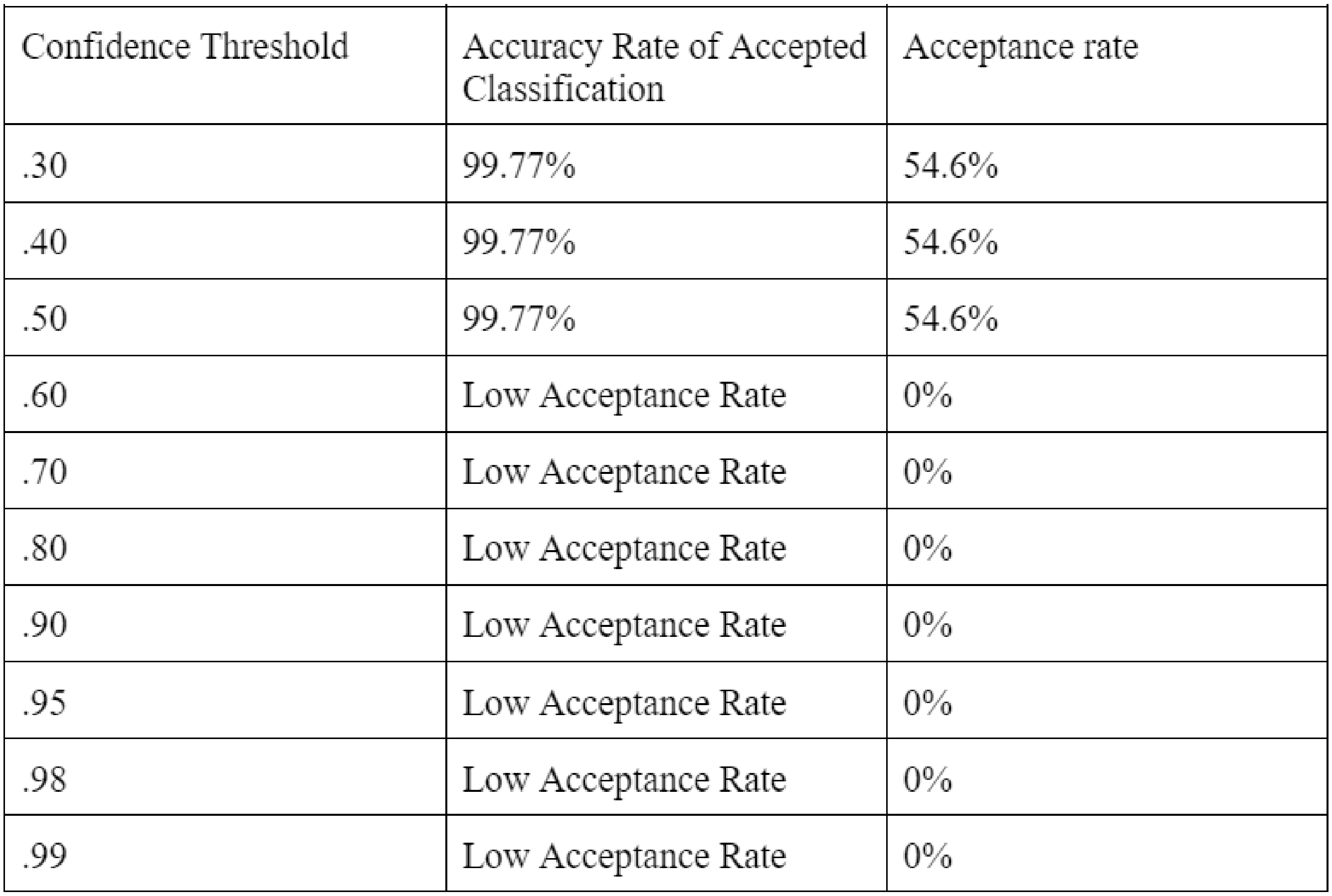
In the context of selective classification, we conducted numerous experiments in order to gauge a variety of the models capabilities and better optimize our model architecture. In this section, we describe the architecture and design of our selective classification models, optimal confidence thresholds for a variety of cases, and resulting useful outcomes and findings - as well as possible applications of these findings.

The basic selective classification architecture involved taking our fitted classification model h(x) (which in our final experiments was set to be a fitted Decision Tree Classifier) and utilize the predict_proba() function, which the SciKit Learn documentation describes as being able to “predict class probabilities of the input samples X” to compare the outcome against a predefined threshold. Should the predicted class probability of the input sample be lower than the given threshold, the model rejects classification of the sample, which we will refer to as “abstaining”. If the predicted class probability of the prediction generated by h(x) is greater than the threshold value, then the predicted classification is “accepted” and produced. This method is helpful in that we are able to at least pinpoint predictions that are likely to be accurate, abstain from those that are not, and yield a high level of accuracy in the final set of generated outcomes. We will later discuss possible clinical uses of this type of algorithm and our numeric results.

6.1 Threshold Selection and Differential outcomes

We experimented with a number of threshold values in order to yield the maximum level of accurate prediction classifications. We tested thresholds values of .30, .40, .50, .60, .70, .80, .90, .95, .98, and .99. Below is a table that illustrates values such as the number of abstentions, number of accepted classifications, and the average percent accurate of the accepted classifications.

Table 6.1: Confidence Threshold and Selective Classification Outcomes



Confidence Threshold	Accuracy Rate of Accepted Classification	Acceptance rate
.30	99.77%	54.6%
.40	99.77%	54.6%
.50	99.77%	54.6%
.60	Low Acceptance Rate	0%
.70	Low Acceptance Rate	0%
.80	Low Acceptance Rate	0%
.90	Low Acceptance Rate	0%
.95	Low Acceptance Rate	0%
.98	Low Acceptance Rate	0%
.99	Low Acceptance Rate	0%

We can see in the table above that at a critical confidence value between .55 and .6, the model begins to routinely abstain from making decisions due to too high of a confidence threshold. However, if we decrease our confidence threshold to well below .30, our model will virtually always provide a prediction, but the prediction will often be inaccurate.

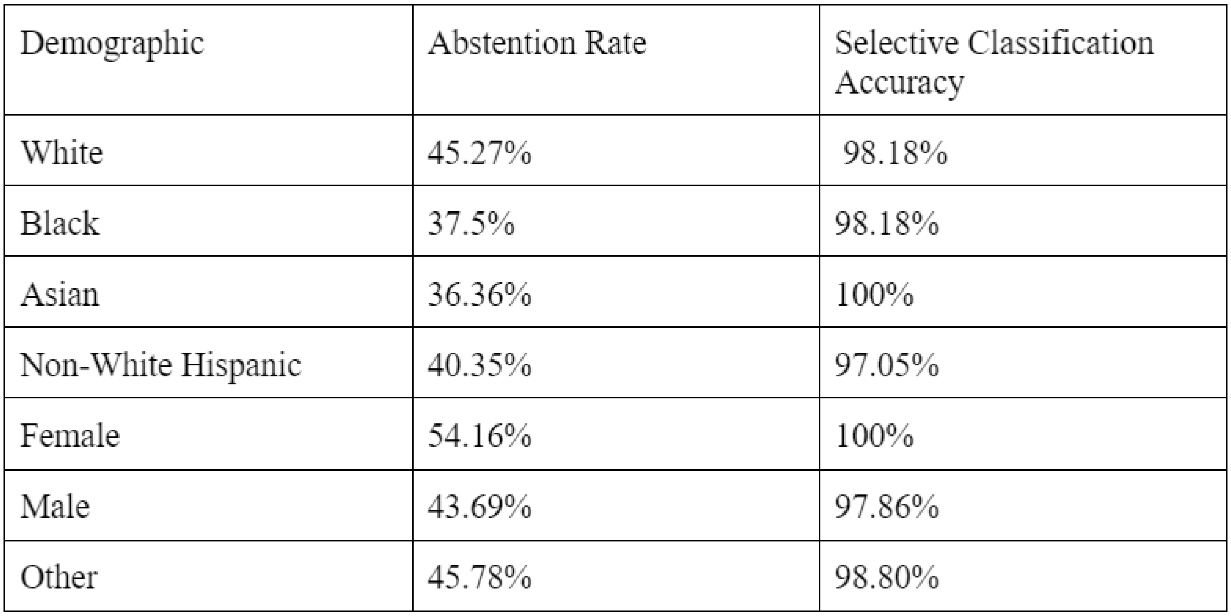
An interesting finding we observed was that as our optimal threshold was reached as we approached .60, we saw an average level of accuracy of, on average, 98% to 100% for validation sets of about 50 samples. However, a limitation presented by that level of confidence is that we were only able to "accept" classifications of narcolepsy type 2. We believe that this results from the fact that the original base classification task is largely intractable due to the limitations we describe around the data and real-world challenges regarding sleep-wake disorder diagnosis processes, identification, and classification. However, the 99.5+% accuracy outcome for even one class of the illness is helpful in that it means that there exists a significant number of patients (as we can see in the percent classifications accepted at near the .60 threshold level) who we are able to predict a diagnostic sleep-wake disorder classification for with near certain levels of accuracy.

6.2 Algorithmic Fairness and Demographic Differences in Abstain Rates

6.2.1 Demographic Differences in Abstain Rates

A further experiment we conducted was measuring and comparing the levels of abstain rates between differing demographic groups, as well as differing measures of the accuracy of our selective classification model. We present showing our abstain rates, as well as our post abstention levels of accuracy, for differing sexes and races. This was done in order to better assess and understand the level of "fairness" of our selective classification paradigm, and to attempt to later address issues in this area as well as pose solutions and implement solutions that we discuss later.

Table 6.2: Abstention Rates and Selective Classification Accuracy Across Demographic Groups



Demographic	Abstention Rate	Selective Classification Accuracy
White	45.27%	98.18%
Black	37.5%	98.18%
Asian	36.36%	100%
Non-White Hispanic	40.35%	97.05%
Female	54.16%	100%
Male	43.69%	97.86%
Other	45.78%	98.80%

Interestingly enough, we see both populations that present with higher (e.g. lower abstention values) and lower levels of algorithmic confidence are able to have high levels of accuracy with respect to the predictions accepted by the selective classification framework. We see that female patients have lower levels of algorithmic confidence, yet higher rates of algorithmic accuracy in the outputs produced for them by the proposed framework. We hypothesize that this may be a result of there being significantly more women in our experimental dataset than men, and there consequently being a more complex feature space for them.

6.2.2 Algorithmic Fairness Considerations

As described in the work done by Lee et al., sometimes selective classification can widen the gap between the performance outcomes of a machine learning framework for different groups. In order to address this issue of algorithmic fairness, we first assessed differential prediction outcomes with respect to different demographic groups in our dataset. We investigate whether or not the useful selective classification outcomes apply to different racial demographics, each gender, and across age groups. In order to do this, we took the various records of patient demographic information that was extracted to help aid prediction and re-used this data to separate the cohort into subgroups for each of the aforementioned demographics. From there, we measure various aspects of the frameworks performance on each of these groups, and conduct statistical analysis to see whether or not significant differences in algorithmic performance arise. Specifically, we compare the accuracy of selective classification methods across groups. We also measure and compare rates of abstention versus rates of prediction acceptance as this also provides valuable information as to how our framework performs across these demographic groups. Earlier, we provided tables of these values for different racial demographic groups, as well as across different sexes [7].

Based on our findings above, we propose a few solutions to increase the fairness of machine learning algorithms such as our application of selective classification methods. For one, we suggest that groups with high prediction acceptance rates/low abstention rates be preliminarily evaluated using our given methods, so that they may have a better grasp of their condition earlier on (for example, Black and Asian patients may benefit more from selective classification, and as racial minorities in America they may even further benefit from this framework). Groups with high abstention rates may benefit from coupling these methods with other predictive methods such as initially utilizing conformal prediction, or from lowered confidence thresholds so as not to leave the majority of the demographic's population with no predictive outcome regarding their condition.

Chapter 7

Conclusions

7.1 Limitations

There were a number of limitations regarding our research. Some of these initial limitations posed interesting questions that we were able to make headway on solving, while others continue to impact our ability to achieve optimal findings. To begin, several limitations were caused by the very nature of American medical claims data. For example, one issue with American medical claims data is that, due to the nature of how insurance coverage is approved, doctors may need to intentionally "misdiagnose" patients in order to allow them to have access to medication that may serve them better, which may then result in noise within the model and lead to less accuracy in predicting what sleep-wake disorder label is appropriate for a patient. Moreover, medical claims data is especially limited when sourced from the United States due to the fact that the United States is the only developed nation that lacks universal healthcare. This can lead to a number of issues. For example, it is possible that the United State's lack of universal healthcare access may be why American medical claims data is not reflective of the natural frequency of certain conditions in nature, such as the fact that Narcolepsy type 1 is significantly more common (and accounting for 64% of all narcolepsy patients), whereas Narcolepsy type 2 only is measured to have a prevalence of about 36% in nature. There are further issues as to why this also may be the case such as a volunteer effect, certain demographics receiving care that happen to be affected by narcolepsy type 1 more, and further reasons worthy of investigation. However, in the medical claims data utilized, these ratios were beyond inversion. The data approximately comprised of greater than 2/3rds narcolepsy type 2 patients and significantly less than 1/3rd of narcolepsy type 1 patients. Interestingly enough, this issue either does not exist or is of a significantly smaller magnitude when observing European data sets. A further limitation exists when attempting to extract physician specialties that are less well documented in claims data. For example, it could prove to be helpful to include information such as when patients see sleep specialist providers and what the outcomes following those visits are, but this isn't always possible because very often they sleep specialist may be labeled as an internist or nonspecifically labeled as a neurologist. A further limitation is the actual measurable level of differences in patients. There are a variety of standing issues with respect to the manual diagnosis of and differentiation between different sleep-wake disorders. For example, the level of similarity between narcolepsy type 2 and idiopathic hypersomnia has to led to debate regarding whether or not they should lumped into one acknowledged condition. Additionally, it is exceedingly difficult to accurately differentiate between narcolepsy type 1 and type 2 initially because a formal narcolepsy type 1 diagnosis can not be made until a bout of cataplexy is observed, which further illustrates the level of similarity between the conditions. Limitations like these render it difficult to utilize classical classification methods alone to produce useful differential diagnoses between different sleep-wake disorders, and can shed light onto the predictive outcomes of the stand-alone base model.

7.2 Findings

In sum, we develop and test a framework and streamlined method such that we are able to address high risk classification problems that appear to be intractable due to issues with noise within the data set, unclear boundaries between classes, and more. We are able to, with a high level of confidence, predict what class(es) the final label of interest will be a member of. Due to the difficult nature of solving these sorts of classification tasks with classical methods, we do lose out on the level of precision of some predictive outcomes, but we are also able to, with a smaller subset of the data and via the use of selective classification methods such as including abstaining as a possible prediction, single out members of the smaller subset such that we can provide them with highly accurate diagnosis predictions of, on average, between 98 to 100%, which is promising for clinical settings wherein it is hard to differentiate between for various reasons (e.g. highly similar diagnoses that are subject to change due to a lack of evidence of symptoms that may develop later on in a patients recorded medical history). We also engage in work to investigate the underlying mechanisms of the model and more general model interpretability work, and also investigate the level of fairness of the model, in order to propose a framework that benefits all demographics represented in our dataset as opposed to only a select few.

7.3 Future Work

We propose a variety of directions for future work. For one, it may be interesting to replicate our work and apply our framework to claims data from a region/regions where insurance coverage is less of a limitation, where the population's genetic breakdowns are different and may lead to different manifestations and rates of sleep-wake disorder diagnosis, and where treatment practices may vary, among other medical paradigms. A further direction for future work may include utilizing patient visit reports and the application of natural language processing methods such as transformers in tandem with out uncertainty quantifying approaching in order to produce even more robust results. Finally, we intentionally designed this framework to be highly generalizable, and hope that it may be of use to any parties interested in applying our framework to other families of closely related illnesses that may be difficult to discern between.

Chapter 8

Bibliography

[1] Angelopoulos, A. N., Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

[2] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

[3] Butler, A.M., Nickel, K.B., Overman, R.A., Brookhart, M.A. (2021). IBM MarketScan Research Databases.

[4] Dauvilliers, Y., Bogan, R. K., Arnulf, I., Scammell, T. E., St Louis, E. K., Thorpy, M. J. (2022). Clinical considerations for the diagnosis of idiopathic hypersomnia. Sleep Medicine Reviews, 101709.

[5] El Mekkaoui, S., Ferreira, C. J., G'omez, J. C. G., Agrell, C., Vaughan, N. J., Heggen, H. O. (2023, August). Neural Networks based Conformal Prediction for Pipeline Structural Response. In Conformal and Probabilistic Prediction with Applications (pp. 134-146). PMLR.

[6] Fisch, A., Jaakkola, T., Barzilay, R. (2022). Calibrated selective classification. arXiv preprint arXiv:2208.12084.

[7] Lee, J. K., Bu, Y., Rajan, D., Sattigeri, P., Panda, R., Das, S., Wornell, G. W. (2021, July). Fair selective classification via sufficiency. In International conference on machine

learning (pp. 6076-6086). PMLR.

[8] Linusson, H., Johansson, U., Boström, H., Löfström, T. (2018). Classification with reject option using conformal prediction. In Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part I 22 (pp. 94-105). Springer International Publishing.

[9] Maksabedian Hernandez EJ, Tingzon I, Ampil L, Tiu J. Identifying chronic disease patients using predictive algorithms in pharmacy administrative claims: an application in rheumatoid arthritis. J Med Econ. 2021 Jan-Dec;24(1):1272-1279. doi: 10.1080/13696998.2021.1999132. PMID: 34704871.

[10] Mollas, I., Bassiliades, N., Tsoumakas, G. (2022). Conclusive local interpretation rules for random forests. Data Mining and Knowledge Discovery, 36(4), 1521-1574.

[11] Rasmy, L., Xiang, Y., Xie, Z. et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. 4, 86 (2021). https://doi.org/10.1038/s41746-021-00455-y

[12] Ross, S., Gordon, G., Bagnell, D. (2011, June). A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 627-635). JMLR Workshop and Conference Proceedings.

[13] Segal, J. (2016, June). Truven Health Analytics MarketScan [Slide show; PowerPoint]. https://ictr.johnshopkins.edu/wp-content/.
https://ictr.johnshopkins.edu/wp-content/uploads/2016/06/jodiMS-slide-deck.pdf

[14] Sohn, I. (2019). Robustness enhancement of complex networks via No-Regret learning. ICT Express, 5(3), 163-166.

[15] Suthaharan, S., Suthaharan, S. (2016). Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269.

[16] Taddei, R. N., Werth, E., Poryazova, R., Baumann, C. R., Valko, P. O. (2016).

Diagnostic delay in narcolepsy type 1: combining the patients' and the doctors' perspectives. Journal of sleep research, 25(6), 709-715.

[17] Won, C., Mahmoudi, M., Qin, L., Purvis, T., Mathur, A., Mohsenin, V. (2014). The impact of gender on timeliness of narcolepsy diagnosis. Journal of Clinical Sleep Medicine, 10(1), 89-95.

[18] Wozniak, D. R., Quinnell, T. G. (2015). Unmet needs of patients with narcolepsy: perspectives on emerging treatment options. Nature and science of sleep, 51-61.

[19] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).