Tarini Basireddy
Biomedical Data Science Capstone Project

1. Background and Motivation

Sleep disorders are conditions that adversely impact the quality and duration of an individual's sleep. There are over 80 different sleep disorders, of which the two most prevalent are insomnia and obstructive sleep apnea.[1] Insomnia is characterized by difficulty in both initiating and maintaining sleep. Individuals with insomnia typically awaken multiple times during the night and often wake significantly earlier than desired. Obstructive Sleep Apnea (OSA) is characterized by recurrent episodes of airway obstruction during sleep, causing an individual to wake up several times throughout the night. Both conditions typically result in excessive daytime drowsiness, cognitive and motor impairment, and a reduced quality of life.[1]

Treatment of sleep disorders quickly after their onset significantly reduces the risk of short-term and long-term health issues. Unfortunately, many people are unaware of the symptoms of sleep disorders, making them unlikely to seek early medical care. This challenge is compounded with the large variability in sleep disorder symptoms between individuals. As a result, sleep disorders are underdiagnosed and undertreated.[2] In this project, I develop a PyTorch Neural Network model with two hidden layers capable of performing multiclass classification, with the goal of using sleep and lifestyle metrics to predict whether or not someone has Insomnia or Obstructive Sleep Apnea. This model could integrate as a phone app or as a pre-screening questionnaire before a doctor's appointment, allowing more individuals to receive prompt care for their sleep disorders.

2. Exploratory Data Analysis

In this project, I explore a publicly-available dataset capturing key details about sleep patterns, lifestyle habits, and related health indicators for 374 individuals with no sleep disorders, Sleep Apnea, or Insomnia.[3] The dataset captures patient information across 15 variables, including lifestyle (Physical Activity Level, Stress Level, Daily Steps, Occupation), sleep (Sleep Duration, Quality of Sleep), and demographics (Age, Heart Rate, Blood Pressure). Before I train and test a model, it is imperative to examine key variables of the dataset and compare it to known trends in insomnia and OSA to evaluate the fidelity of the data.

I first pre-processed the dataset by deleting extraneous variables (occupation and patient ID) and dividing the systolic and diastolic BP of each patient into separate columns. I first explored the gender distribution of my overall dataset **(Fig 1A)**, as well as amongst the patients diagnosed with OSA **(1B)** and Insomnia **(1C)**. The data reveals that the overall gender distribution is roughly even (50.5% vs. 49.5%). However, women comprise a significantly greater proportion of the patients diagnosed with OSA than men (85.9% vs. 14.1%) whereas they comprise a slightly lower proportion of patients diagnosed with insomnia (46.8% vs. 53.2%). These findings deviate from global trends, as OSA is typically more prevalent in men and insomnia is generally more common in women.[1]

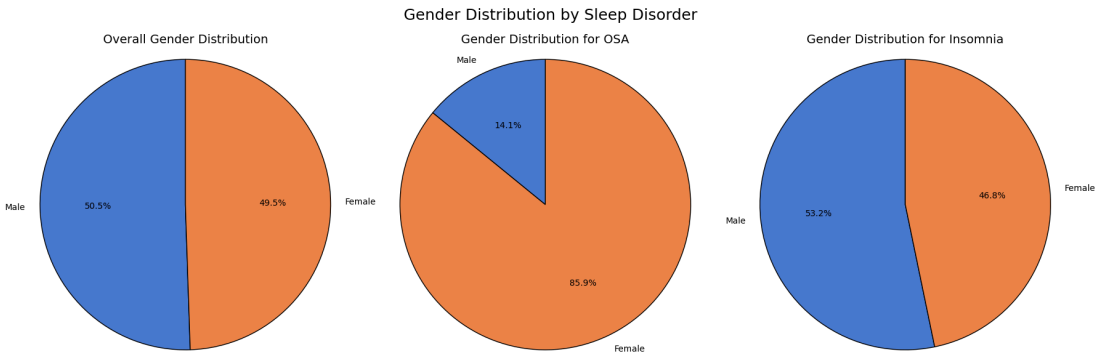

Figure 1. Overall gender distribution (A), Gender distribution of OSA patients (B), and gender distribution of insomnia patients (C). Blue represents males, while orange represents females.

I then evaluated the overall distribution of patient ages **(Fig 2A)** and the age distribution for each specific sleep disorder **(2B)**. Overall, the dataset comprised patients with a broad range of ages between 27-59, with a concentration of individuals between 42-44 years old. On average, individuals diagnosed with OSA tended to be older, between 48-57. In contrast, those with insomnia are generally younger between 43-46. Still, both groups were older than those without a diagnosed sleep disorder, who were typically 32-44 years old. This is expected, as the risk of sleep disorders increases with age.[1]

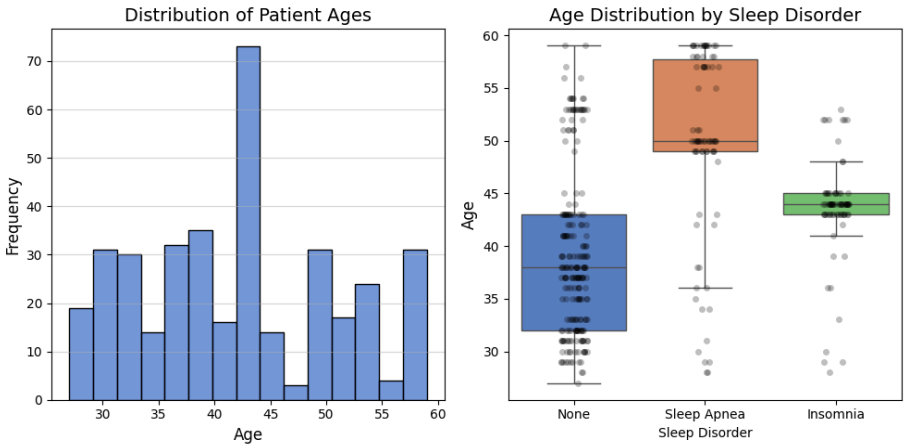

Figure 2. Histogram of overall distribution of patient ages (A) and box plot of age distribution classified by sleep disorder (B). Blue represents no sleep disorder, orange represents sleep apnea, and green represents insomnia.

I also examined the distribution of key metrics, including heart rate **(Fig 3A)** , systolic BP **(3B)**, and diastolic BP **(3C)**, throughout the dataset. Each corresponding plot features a red vertical line to indicate the upper limit of the clinically healthy range for each variable.[4] The analysis reveals that 0% of patients exceed the threshold for heart rate, while 77% of patients exceed that of systolic BP. Additionally, 59% of patients are above the threshold for diastolic BP.

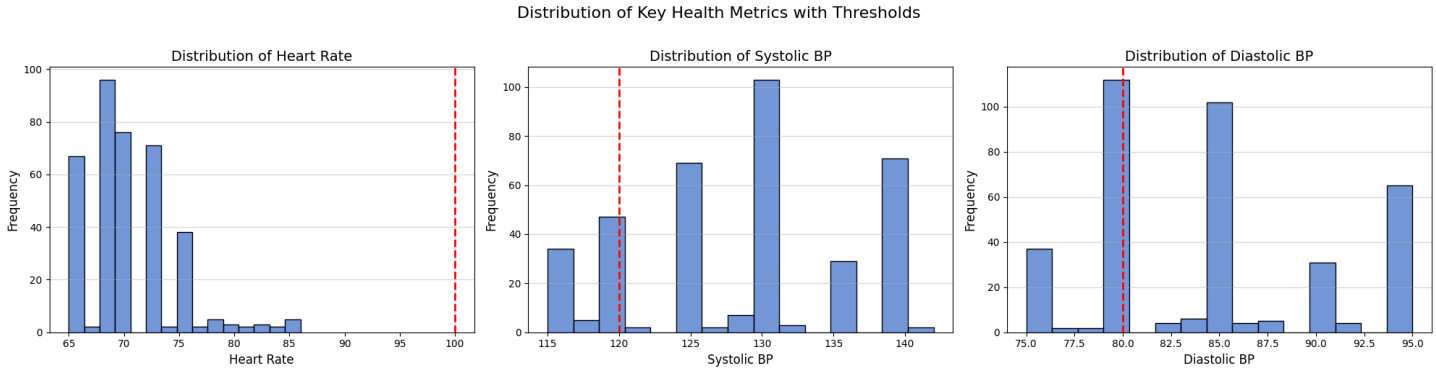

Figure 3. Histograms of the distribution of heart rate, distribution of systolic BP, and the distribution of diastolic BP. The threshold for heart rate is 100 bpm, systolic BP is 120 mmHg, and diastolic BP is 80 mmHg.

I also visualized the relationship between heart rate and diastolic BP. Although these two variables are positively correlated (higher heart rate = higher diastolic BP), they are also independently linked with disordered sleep. For instance, individuals with early-stage OSA often experience a rise in diastolic BP without a corresponding rise in systolic BP or heart rate.[1] When heart rate was plotted against diastolic blood pressure for patients with no sleep disorder (**Fig 4A**), sleep apnea **(4B)**, and Insomnia **(4C)**, I found that patients with no sleep disorders tended to have a lower diastolic BP and heart rate, with heart rates below 76 bpm. However, individuals with sleep apnea and insomnia tended to have patients with moderate-high diastolic pressure and high heart rate up to 86 bpm.

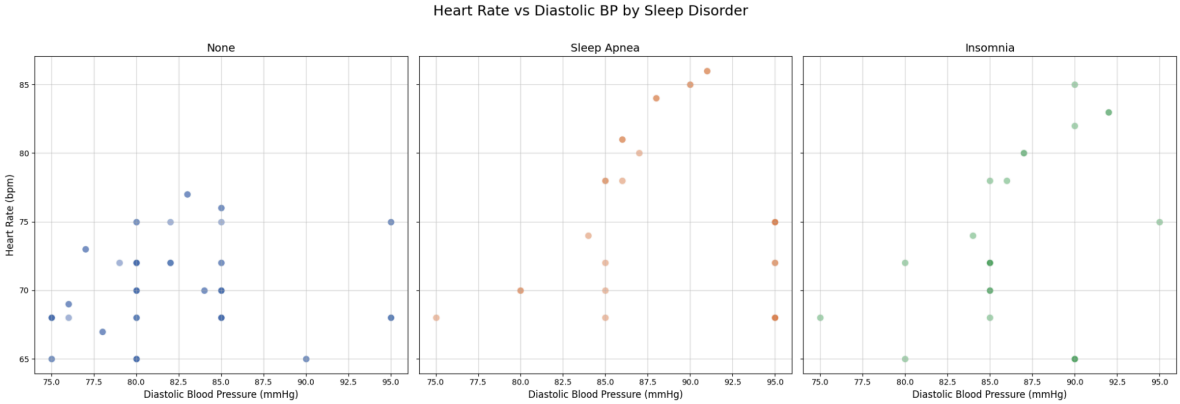

Figure 4. Scatterplot depicting heart rate (bpm) vs diastolic blood pressure (mmHg) for patients with no sleep disorder (**A**; blue), patients with sleep apnea (**B**; orange), and patients with insomnia (**C**, green).

The incidence of sleep disorders varies depending on an individual's body mass index (BMI). Below, I present a plot showing the proportion of patients in the dataset categorized by BMI (normal, overweight, and obese) who have no sleep disorder, sleep apnea, or insomnia **(Fig 5)**. As seen in the plot, a majority of individuals with a normal BMI (~90%) had no sleep disorder, with ~5% having sleep apnea and ~5% having insomnia. Being overweight put individuals at a much higher risk of sleep disorders, as ~45% had insomnia and an equal proportion had sleep apnea. Finally, individuals who are obese are at the highest risk for sleep disorders particularly OSA, as ~40% had insomnia and the remaining 60% had sleep apnea.

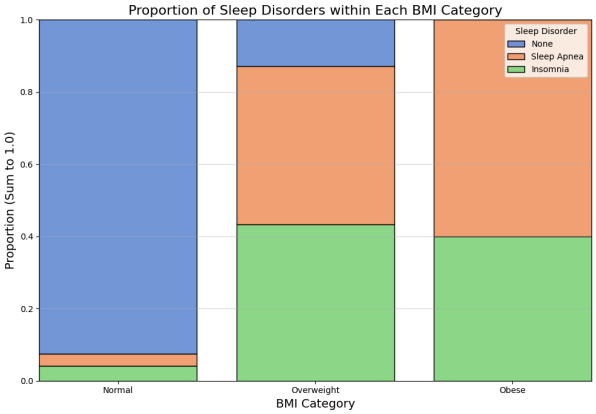

Figure 5. Stacked bar chart depicting the proportion of patients with no sleep disorder (blue), sleep apnea (orange), and insomnia (green), separated by the BMI category (normal, overweight, and obese).

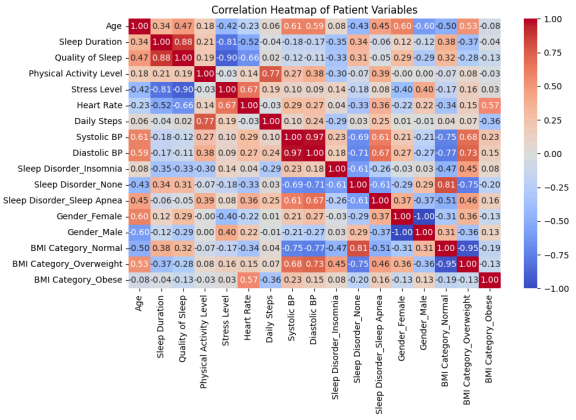

Figure 6. Correlation heatmap of patient variables. Red represents maximum positive correlation, whereas blue represents maximum negative correlation.

Ultimately, I generated a correlation heatmap of each of the patient variables **(Fig 6)**. As expected, the greatest positive correlation was observed between systolic and diastolic BP. Looking specifically at the sleep apnea category, I found that Systolic BP, Diastolic BP, Age, and an Overweight BMI had the greatest positive correlation. These variables were strongly negatively correlated with having no sleep disorder. On the other hand Systolic BP and Diastolic BP were not as strongly correlated with insomnia, which was negatively correlated with sleep duration and quality of sleep as expected.[1] The other variables showed a similar correlation with insomnia compared to sleep apnea, which is slightly concerning as it may be difficult to categorize a patient as likely to have one of the two sleep disorders, compared to whether or not a patient has a sleep disorder at all.

From my exploratory data analysis, I can conclude that my data is fairly representative of my population of interest, and that this dataset may be sufficient to train a neural network to classify individuals who may have OSA and insomnia.

4. Neural Network

I trained a PyTorch neural network model to perform multi-class classification (0 = no sleep disorder, 1 = sleep apnea, 2 = insomnia) using the same dataset as above. I chose to split the data 80% for training, 20% for testing to allow sufficient data for training and testing. I initially trained the model for 500 epochs, however, I noticed a steep spike in the loss of the model following 100 epochs, which is attributed to overfitting.[5] As a result, I lowered the learning rate from 0.1 to 0.05 and only trained the model for 100 epochs, at which point the loss function plateaued **(Fig 7).**

Following training, the model performed fairly well. The model achieved an accuracy of 90.67%, precision of 90.56%, and recall of 90.67%. Furthermore, I calculated the AUC-ROC (Area Under the Curve of a Receiver Operating Characteristic) score using the One-vs-the-Rest multiclass strategy, which allows us to compare the AUC-ROC score of this classifier with the thresholds identified for a binary classifier. The model achieved a AUC-ROC score of 0.8764, which is typically considered excellent for distinguishing between the different classes.[6]

I then plotted the confusion matrix for the prediction, which shows the relative proportions of true and false classifications **(Fig 8)**. As seen in the matrix, the model performed well at predicting all three categories, with only a few off-target predictions. Overall, these results demonstrate the effectiveness of my neural network model in accurately predicting sleep disorders based on the available data.

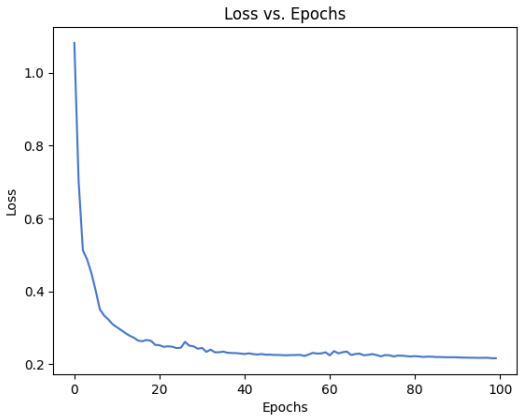

Fig 7. Loss over training epochs.

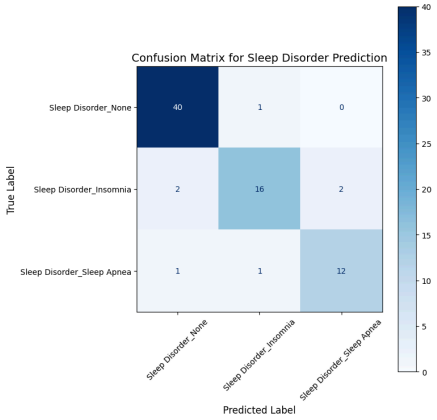

Fig 8. Confusion matrix. White represents 0, navy blue represents the highest value, 40

5. Limitations and Future Analysis

Firstly, the data used in this project is **simulated** for an educational purpose, based on real trends in patient data. As a result, the data may not reflect the true characteristics and variability of actual patient populations. For example, the data used to train the model does not reflect the typical trend of gender and OSA/Insomnia prevalence. However, the data may also be biased towards reflecting certain trends (such as all obese patients having a sleep disorder) during the data synthesis process. In addition, the model was only trained on 300 datapoints. As a result, the model may not function as well when asked to make predictions on real patient data, and will need to be trained further on clinical data before being deployed.

Another limitation of the model is that it can only characterize patients into the categories Insomnia, Sleep Apnea, and no sleep disorder. Although insomnia and sleep apnea reflect a majority of sleep disorders, there are many other sleep disorders including narcolepsy and circadian rhythm disorders, which have some similar symptoms to Insomnia and Sleep Apnea. As a result, the model may inaccurately classify some of these patients as potentially having insomnia or sleep apnea, or worse, may characterize them as having no sleep disorder which may prevent them from seeking prompt medical advice.

A future pathway for analysis may be performing SHAP (Shapley Additive exPlanations) analysis to understand which features are being used, as well as how much each feature positively or negatively impacted the final prediction.[7] It would be interesting to compare this SHAP analysis to the correlation heatmap previously generated to understand whether the model more heavily weights the variables that were more strongly correlated with sleep apnea and insomnia. It would also be interesting to ask clinicians to rank the same variables in order of importance when initially screening an individual for a sleep disorder, and to subsequently compare the variables that the model weighs heavily to this ranking to understand where they deviate.

6. Citations

1. Jha, V. M. (2022). The prevalence of sleep loss and sleep disorders in young and old adults. Aging Brain, 3, 100057. https://doi.org/10.1016/j.nbas.2022.100057
2. Celmer, L. (2023, December 18). Undiagnosed and untreated sleep disorders: Barriers to care. American Academy of Sleep Medicine – Association for Sleep Clinicians and Researchers. https://aasm.org/undiagnosed-and-untreated-sleep-disorders-barriers-to-care/
3. Sleep_Disorder_diagnosis_dataset. (2025, September 17). Kaggle. https://www.kaggle.com/datasets/varishabatool/disorder
4. What is high blood pressure? | NHLBI, NIH. (2024, April 25). NHLBI, NIH. https://www.nhlbi.nih.gov/health/high-blood-pressure
5. What is Overfitting? - Overfitting in Machine Learning Explained - AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/overfitting/
6. Classification: ROC and AUC. (n.d.). Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
7. Lundberg, S. (2018). SHAP documentation. https://shap.readthedocs.io/en/latest/