BIG DATA HEALTH SCIENCE CASE COMPETITION

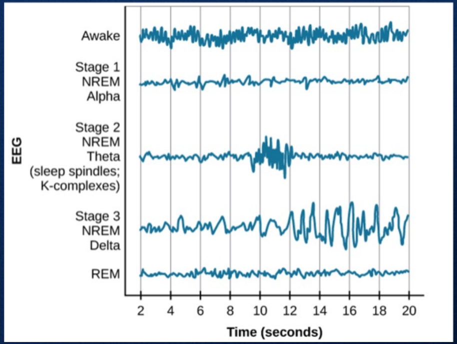
STANDARD DEVIANTS

- Sleep disorders affect 70 million Americans
- Undiagnosed cases lead to higher healthcare costs and reduced productivity
- Translational research with rodents informs our understanding of human biology

BACKGROUND

PROBLEM STATEMENT

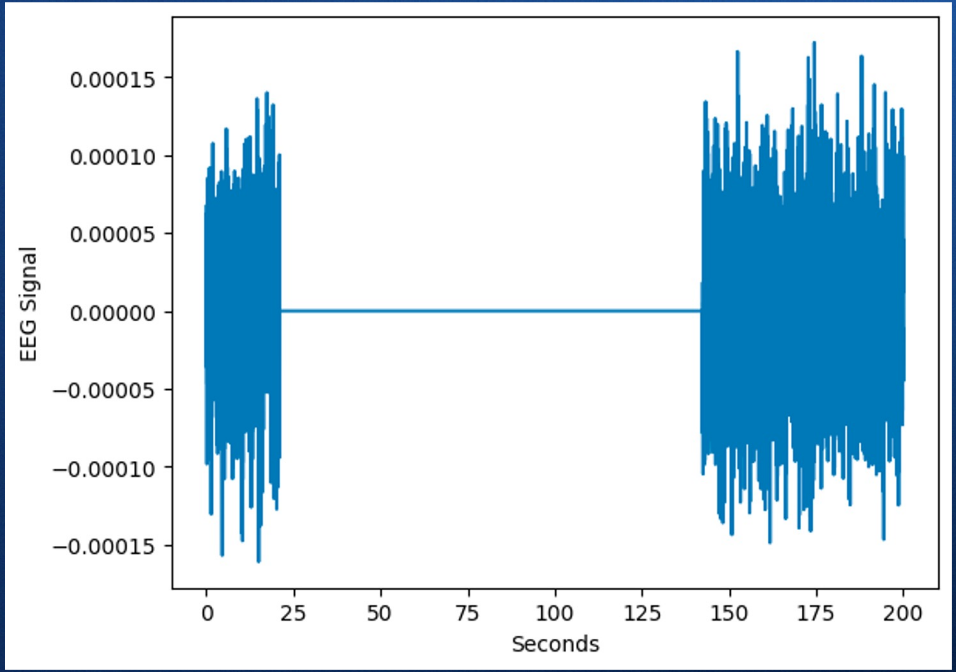
- EEG is a technique for monitoring electrical brain activity in real-time and is effective in helping identify the distinct stages of sleep cycles
- As it currently stands, we have to employ a dedicated expert to manually annotate an EEG chart
- This is both resource-intensive and time-consuming
- With an obvious need for a more scalable and cost-effective solution, we aim to develop an automated system for efficient EEG data processing



THE DATA

Single-Channel EEG

- Uncertainty in reading
- Almost no spatial resolution



LIMITATIONS OF THE DATA

Lack of Features

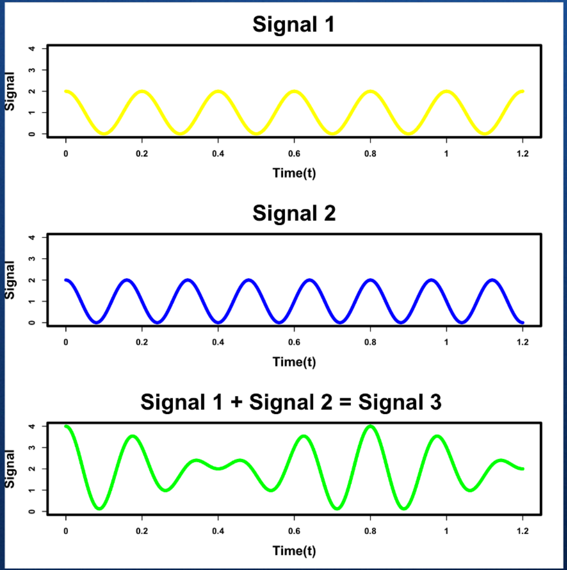
- EEG Readings are the only available information
- No information relating to the rats' activity
- No information relating to the time of day

LIMITATIONS OF THE DATA

DATA PROCESSING

EEG data records the electrical signal from *all* brain activity, even that which is insignificant, making it very noisy

Here, we have the combination of just two waves (yellow and blue), which would be read as a single output (the green wave) since the electrical outputs of each are indistinguishable

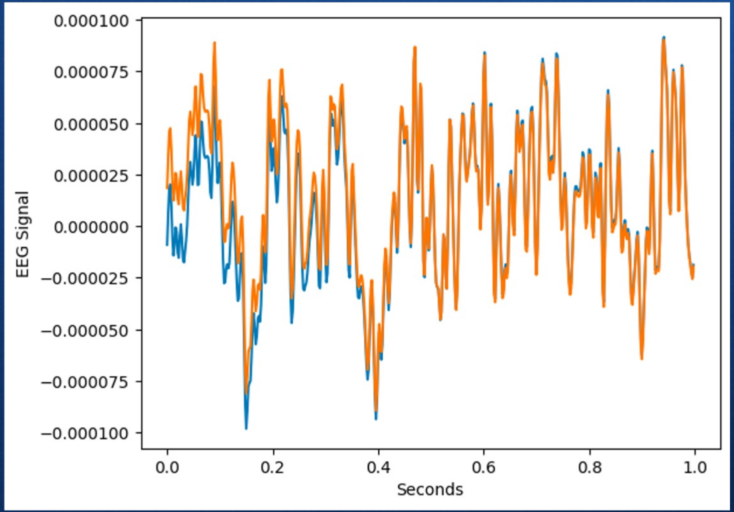


DATA PROCESSING

Our first step in the solution to dirty data is applying a smoothing function, in this case the Butterworth filter because of its common use in EEG data

This filter splits our wave into its different frequency components and then blunts some of those frequencies based on specified criteria

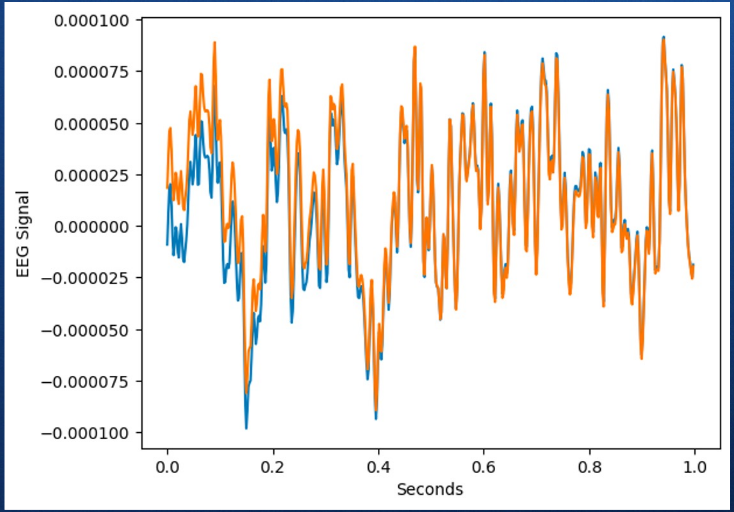
It performs these modifications in chunks (different time intervals)



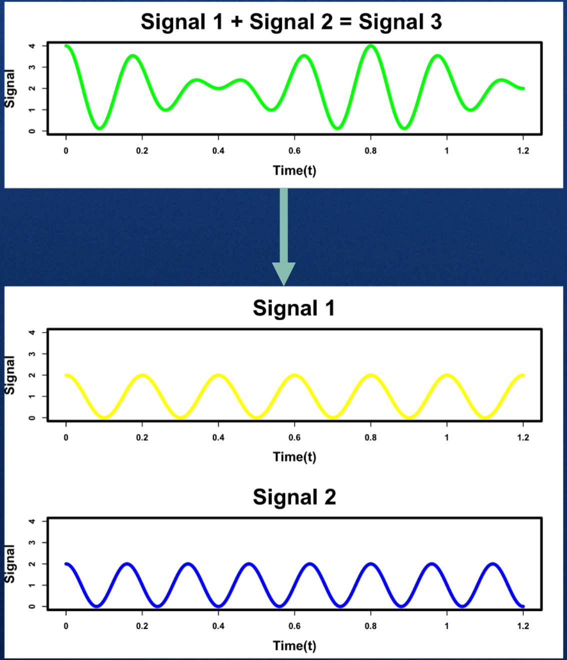
DATA PROCESSING

We specifically constrained the Butterworth filter to isolate waves with frequencies between 0.5 Hz and 200 Hz

This is the range of brain wave frequencies observed for rats



DATA PROCESSING

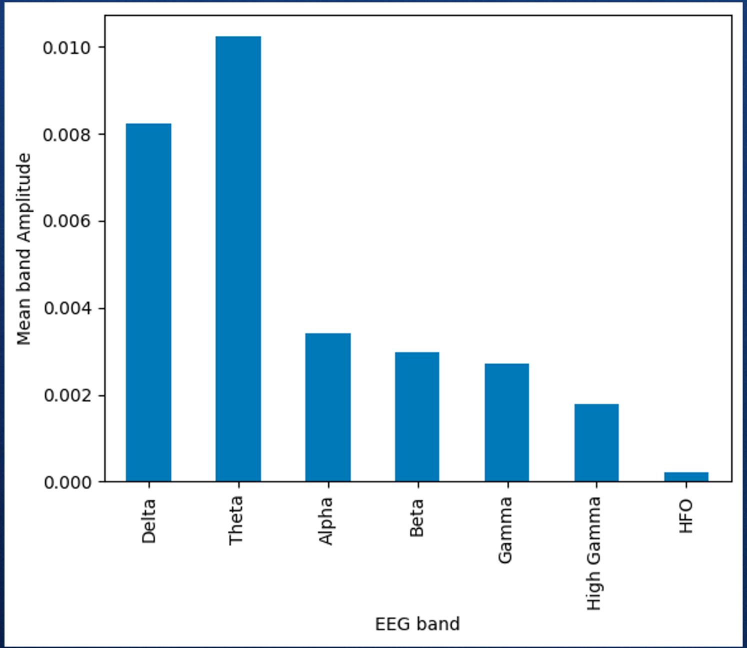


To obtain our final dataset, we performed a Fourier transform to split the smoothed wave into its individual frequencies.

We selected the 5 greatest amplitudes and their corresponding frequencies, recording this information.

This process was performed in chunks of one epoch of data (10s) at a time

DATA PROCESSING



We then grouped the individual frequencies produced by the Fourier transform based on their classifications according to literature

The value reported is the mean amplitude for each group

MODEL FEATURES

	amplitude_1	...	amplitude_5	frequency_1	...	frequency_5	average_delta	average_theta	...
epoch_1									
epoch_2									
epoch_3									

Choosing a Model

Tested different multi-class regression models, ultimately selecting Random Forest because of its low computational demands during training, lesser sensitivity to hyper parameters, and its robustness to overfitting

Decision Tree

Easily interpretable, but did not address the unbalanced class issue and is prone to over-fitting

XGBoost

Less computationally intensive than SVM and Neural Networks but not a good performance for 3 class classification problem compared to random forest

Random Forest

More consistent and better at avoiding overfitting compared to the decision trees, randomizes outcome from decision trees.

Neural Network

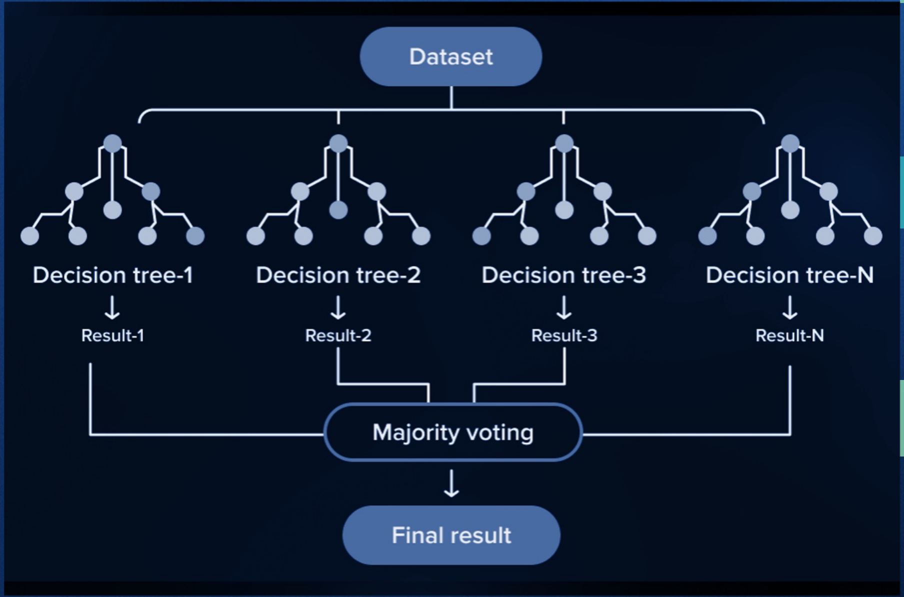
Required excessive computational resources to train, could not run to completion

SVM

Did not notably improve from decision tree but required significantly more computational resources and training time

OUR MODEL

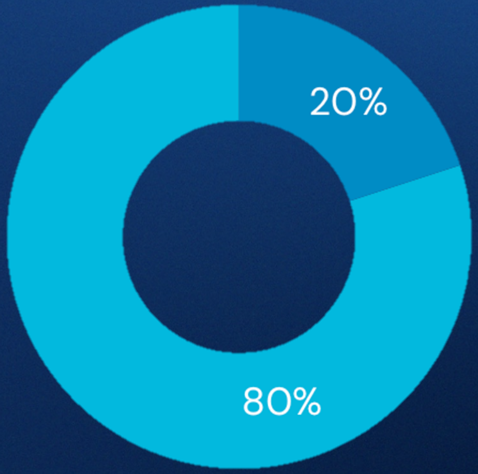
- Once we had our features, we used the Random Forest model to predict the state of sleep a rat was in at a given time
- Random Forest makes a prediction by generating many random decision trees, allowing them to make their own predictions, and then considering the majority consensus



EVALUATION

We partitioned our data into an 80% training set and a 20% validation set, used to test our model. We then evaluated the model's performance by calculating the F1 Score for each sleep stage.

The F1 score provides a more balanced measure of accuracy when the predicted variable is imbalanced, as it is in our case.



Data Analysis Workflow

FINAL MODEL DESIGN AND PERFORMANCE

04 **Model Fitting**
Random Forest Algorithm using 80% training and 20% validation

03 **Feature Extraction**
Selected 5 frequencies with greatest amplitudes and calculated mean amplitudes of frequency bands for each epoch

02 **Data Transformation**
Isolated frequencies of component waves using Fourier transform

01 **Data Filtering**
Smoothed curve to only allow frequencies 0.5 to 200 Hz

F1 SCORES:

0.72	0.94	0.93
Phase REM	Slow-wave Sleep	Wakefulness

RESULTS

CONFUSION MATRIX

Actual Labels \ Predicted Labels	Phase REM	Slow-wave	Wakefulness
Phase REM	1006	41	459
Slow-wave	29	11292	604
Wakefulness	243	744	13230

CHALLENGES WE FACED

- Computationally intensive
- Difficult to visualize the decomposition of the EEG wave into its individual frequencies
- Exploration of niche techniques and packages

POTENTIAL IMPROVEMENTS

- Direct research towards task-specific areas first
- Run scripts on small subset of data to ensure functionality

REAL-WORLD APPLICATIONS

PHARMACEUTICAL INDUSTRY APPLICATIONS & REVENUE IMPACT

- Our research offers an understanding of how medications impact various sleep stages
- Pharmaceutical companies can use this insight to optimize drug formulations for efficacy
- Tailoring drug development strategies can lead to cost reductions by streamlining the research & development process
- Medications with proven efficacy in treating sleep disorders can experience increased market success

HOSPITAL APPLICATIONS & FINANCIAL IMPACT

- Precision in diagnosis allows doctors to tailor treatments to address specific stages of sleep, contributing to better patient outcomes
- Targeted treatments based on accurate diagnoses lead to a reduction in inpatient hospitalizations
- Hospitals can allocate resources more efficiently by addressing the specific needs of patients at different sleep stages
- The reduction in hospital stays not only improves patient quality of life but also results in significant cost savings for healthcare systems

ACADEMIC APPLICATIONS

Give rise to new data collection techniques and deeper insights into animal behavior and its modification by environmental and genetic factors

Potential for further insight and research into epileptic disorders in model animals and by extension humans

HOW CAN WE TAKE THIS RESEARCH FURTHER?

- Expand data collection by using additional physiological measures, such as heart rate and movement data
- Increase complexity of the model by considering a Convolution Neural Network (CNN) approach
- Investigate if the recorded markers show patterns consistent with sleep disorders
- Extend the research's applicability by translating findings into human clinical trials
- Establish a baseline expectation of concordance between annotators

Q&A