The University of British Columbia

MANU 465

Effects of Ambient Temperature on Cognitive Function

Santiago Mendoza Reyes	40294852
Arjav Prasad	24948499
Ramiro Bolanos	92234673
Sanskar Soni	24854945
Yuhan Zeng	97324388

Table Of Contents

Abstract

Machine learning is an emerging tool that has plenty of use cases and applications. As time progresses more refined use cases will be developed and used. One application that was tested by the team was the effect on temperature on cognitive function. Using a Muse 2 device (EEG device) and its data collection software, data was captured from various UBC students and subjected to cognitive function tests. This data was then cleaned, preprocessed and fed into multiple machine learning models. Accuracies were calculated and tabulated with the highest being XGBosst with 88%.

Introduction

Motivation

The strive for efficiency is the backbone of engineering. The underlying goal of the profession is to maximize efficiency of systems and reduce cost. However with the introduction of machine learning into the mainstream and the barrier of entry to run experiments and tests becoming lower and lower, more unique and interesting discoveries are being found. What if by changing the thermostat by a few degrees you could unlock more efficient studying time, faster completion of tasks with less mental taxation or less time needed for comprehension, and mastery of a new topic? The motivation behind this project is to reveal the effects that different ambient temperatures have on cognitive performance. This research may be useful to further our understanding of the brain's functions.

Project Goal

The goal of this experiment was to use machine learning models to investigate the effect of ambient temperature on cognitive performance. Machine learning models and classification techniques were used on data collected in different environments from subjects during their completion of the Stroop test [1]. The Stroop test is regarded as a valid way to assess cognitive flexibility, attention capacity and mental processing speed, all of which fall in line with our overarching goal. By analyzing the data captured, the effect of ambient temperature will be captured in the EEG. These data waveforms were then pre- and post-processed and fed into machine learning models with varying success.

Hypothesis

The goal of the project is to confirm the relationship between ambient temperature and cognitive performance. Seppänen, Fisk, and Lei found that 21 to 22 degrees Celsius is the temperature range that corresponds to maximum performance in an office setting, with performance dropping from variance from those temperatures [2]. This study's scope was focused on an office setting only, but this project's goal is to investigate the effect of ambient temperature on cognitive performance in general. Our hypothesis should reflect this. Additionally, the study reported a continuous relationship between the percent change in performance per degree Celsius and the absolute performance both as a function of temperature. This project does not aim to establish the relationship between performance and temperature as a function, but rather confirm its existence and trend. Thus, the hypothesis is that for temperatures under a comfortable and typical indoor ambient temperature, or 21 to 22 degrees Celsius, a lower temperature reduces cognitive performance. This hypothesis is tested with the Stroop test, which can be used as a tool for assessing overall cognitive performance independent of the setting it is used in.

Note that the hypothesis claims that a lower temperature corresponds to a decline in cognition, so for the hypothesis to be logically confirmed, all of the following must be supported by the experiment: the results from the machine learning models are meaningful; a relationship between temperature and performance on the Stroop test exists; lower temperatures correspond to lower performance on the test; the results apply to wider cognitive performance, not only the Stroop test; and there are no other factors that improve or reduce cognition other than the different temperatures. The discussion section will assess whether these requirements are met.

Industry Applications

The innovative application of EEG (electroencephalography) technology to study the effects of temperature on cognitive abilities opens up a multitude of potential uses across various industries. This section delves into the practical implications of these findings, exploring how they can be harnessed to optimize environments in workplace settings, educational institutions and elderly care. Each application is supported by existing research, underscoring the relevance and potential of this study.

Workplace Productivity Enhancement

Utilizing EEG-based temperature impact assessments in workplace settings can substantially augment employee productivity. This approach involves meticulously determining the temperature ranges that optimally enhance cognitive functions such as concentration and problem-solving. This methodology resonates with the findings of Seppänen, Fisk, and Lei (2006) [2], who elucidated the direct correlation between ambient temperature and cognitive proficiency in workplace environments. By implementing these insights, organizations can strategically modulate office climates to foster a more productive and cognitively stimulating work atmosphere.

Educational Settings Improvement

The application of EEG-based thermal impact studies in educational environments presents a unique opportunity to bolster learning efficacy. Leveraging this technology to finely tune classroom temperatures can significantly augment the learning environment. Enhancing focus and memory retention among students through optimal thermal conditions can lead to substantial improvements in academic performance. This premise is supported by Heschong's (2002) comprehensive analysis [3], which underscores the integral role of environmental factors in educational outcomes.

Elderly Care and Cognitive Health

In the realm of geriatric care, the application of this research could be substantial. Tailoring ambient temperatures in elderly care facilities to align with the findings from EEG-based studies can substantially support cognitive health in older populations. Van Hoof and Kazak highlight the sensitivity of the elderly to environmental conditions [4], underlining the potential for improved cognitive functioning and overall well-being through such tailored environmental controls.

Background and Overview

This project explores the complex relationship between environmental temperatures and cognitive performance, utilizing electroencephalography (EEG) to capture the brain's electrical activity in response to thermal stimuli. The investigation specifically employs the Stroop test to evaluate cognitive flexibility and attention under varying thermal conditions. This research is poised to provide insights into optimal environmental conditions for cognitive functioning, with potential applications in numerous fields ranging from workplace ergonomics to educational settings.

Background on temperature's effect on the brain

Temperature regulation is vital for maintaining the homeostatic balance required for optimal neuronal function. Research has consistently shown that both hyperthermia and hypothermia can perturb this balance, resulting in impaired cognitive capabilities [5]. Hyperthermia can lead to a decline in attentional resources and memory retention, while hypothermia may slow neural processing speed and reaction times. These effects are thought to arise from temperature-induced alterations in cerebral blood flow, neurotransmitter synthesis, and synaptic efficacy.

Background on EEG

EEG represents a pivotal technique in the field of cognitive neuroscience, offering the ability to track and record the electrical activity generated by the cerebral cortex with remarkable temporal precision. It is highly sensitive to the sub-second fluctuations in brain dynamics, allowing for a granular analysis of cognitive states and processes. The application of EEG in cognitive studies provides a window into the oscillatory activities that underpin cognition, making it an indispensable tool for exploring the neural correlates of temperature effects on cognitive function [6].

Background on the Stroop test

The Stroop Test is a well-established cognitive challenge that interrogates the cognitive control mechanisms of the human brain. It is an empirical measure of the executive function that assesses the ability to override automatic cognitive processes. The test's design capitalizes on the natural propensity to read words more easily than naming colors, presenting a conflict that requires cognitive control to resolve [7]. The Stroop Test's sensitivity to cognitive load makes it an excellent proxy for studying how environmental factors, such as temperature, modulate cognitive flexibility and processing speed.

Figure 1: A. (left) Congruent Stroop test. B. (right) Incongruent Stroop test

Reasoning for choosing Stroop test

The selection of the Stroop Test [1] for this investigation is predicted on its proven efficacy in measuring the cognitive interface and the executive function of the brain, which are hypothesized to be affected by thermal stress. It is highly effective because it challenges the brain's thinking ability, offering a precise way to assess how different temperatures could have an effect on mental performance.

Data collection procedure

Data collection was done in a "Cold" and "Normal" environment. Normal was assigned to be ambient room temperature (~20° C). Whereas cold was established to be a bench outside during early November (~8° C). For each one of these environments a timed congruent sample was taken as a base line, followed by a timed incongruent sample. To remove as much variation as possible in the collection process, subjects were asked to keep their attire the same between test. If a subject was wearing a T-shirt inside while recording their Normal environment sample, they were asked to not put on their coat while the Cold environment data was recorded. To remove any bias caused by the order in which they do the test, half the volunteers were recorded in "Normal" condition, and half at "Cold". Given that the test only lasts two minutes there was minimal risk of temperature related health complications and there was ample amount of hot cocoa mix and tea on standby if needed.

The Muse 2 headband, an EEG recording device, was wiped clean and dry with an alcohol patch between subjects and sensors were calibrated in the app to the best of the team's ability before the experiments started. The headband was set to a notch frequency of 60 Hz to fall in line with North American standards. Wave types recorded were absolute graphs and the average values of all the sensors were taken and recorded.

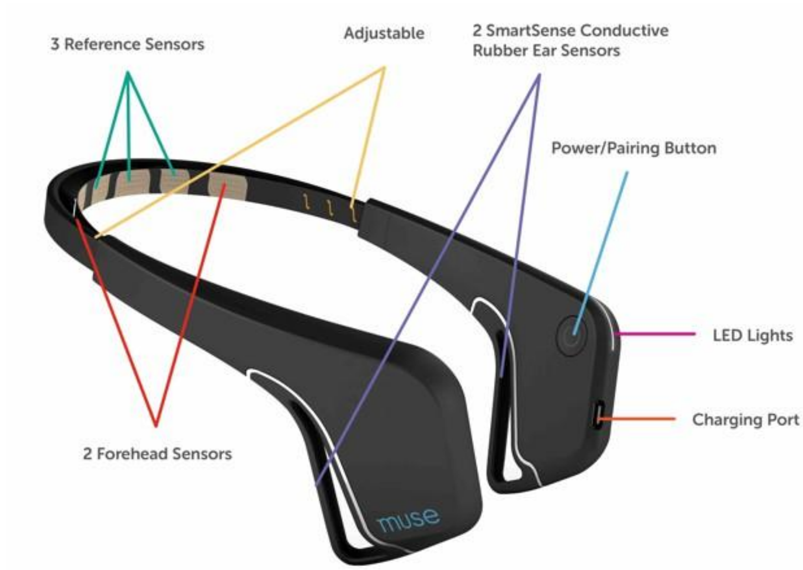

Figure 2: Muse 2 headband used for data recording

Data cleaning

Once all data had been collected 3 recordings from Room and Cold classes each were separated into a test folder. All remaining samples were assigned as train data.

Extracting second test only from recordings

After collection, the data needed to be visualized and inspected, removed of noise and determined the important features which are relevant to our objective. The team acknowledges that only the second half of the stoop test may be beneficial to linking cognitive ability and performance (incongruent test, see Figure 1)

From Figure 1, the first half of the Stroop test serves the purpose of acclimatizing the user to the test. The color of the words and the words itself are the same. It is only in the second test that the words and their spellings and the color they are written in are different. This is what shows the cognitive abilities of subjects and their differences with temperature. It was then decided to only use the second half of all recordings in our tests. While collecting data, the time it took for each of the test subjects to finish both the tests was recorded separately. The column 'Time 2nd Test' was then used to extract only the second test data that the muse headband recorded for each test subject.

	A	B	C	D	E	F
1	**Name**	**Temp (C)**	**Class**	**Date**	**Time 1st Test**	**Time 2nd Test**
2	Test 1	9.5	Cold	Nov 11th	20	34.246
3	Test 2	9.5	Cold	Nov 11th	15	22.739
4	Test 3	9.5	Cold	Nov 11th	15	23.2
5	Test 4	9.5	Cold	Nov 11th	20.174	26.65
6	Test 5	9.5	Cold	Nov 11th	15.14	18.86

Figure 3: Part of the test records with time for each test recorded (refer to appendix for full table)

Filling NaN values

In order to fill NaN values, it would make sense to use either forward fill or backward fill for time series data. The team decided to use forward fill for this project to fill all NaN values. This imputes NaN values with the last valid value of the feature in question.

Cleaning Alpha, Beta, Gamma waves

Upon some research, it was concluded that alpha (8-12 Hz), beta (13 - 30 Hz) and gamma waves (>30 Hz) better captured the brain activity suited to this project. NHAhealth suggests that delta waves (0.1-3.5 Hz) occur in deep sleep, and theta waves (3.6-8 Hz) contain slow activity data related to creativity and daydreaming [8]. Whereas, alpha, beta and gamma waves are more related to fatigue, fast activity like analytical judgment and decision making, and high level information processing respectively.

The muse headband has 4 sensors, and for each sensor, it provides alpha, beta, and gamma waves. These features were visualized in time domain and it revealed that there was lots of noise and unusable data in the features.

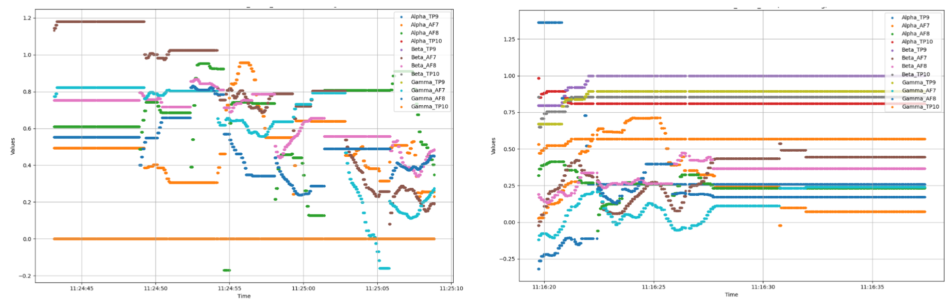

Figure 4: Time domain visualization of some of the recorded data with a lot of noise

Exploratory data analysis showed that while there weren't a lot of outliers, often, some signals either did not change at all throughout a subject's recording, while in some time segments, all of the 12 features visualized were constant. This was considered to be an error in the reading from the device. In trying to clean the data, various methods were tried:

1. If a feature doesn't change at all throughout a test recording, it is set to -1 so that the models can learn that this is not useful data
2. Removing segments of 2 seconds where none of the 12 features change at all (deleting the rows). Replacing the features with the average values of the respective features was considered, however with the sheer volume of segments where all 12 features did not change, this would create too much similarity between the two classes and there would be too little variation between the classes.
3. Setting segments of features that don't change for 10 seconds to -2. A feature not changing values for over 10 seconds was also considered to be an error in reading.

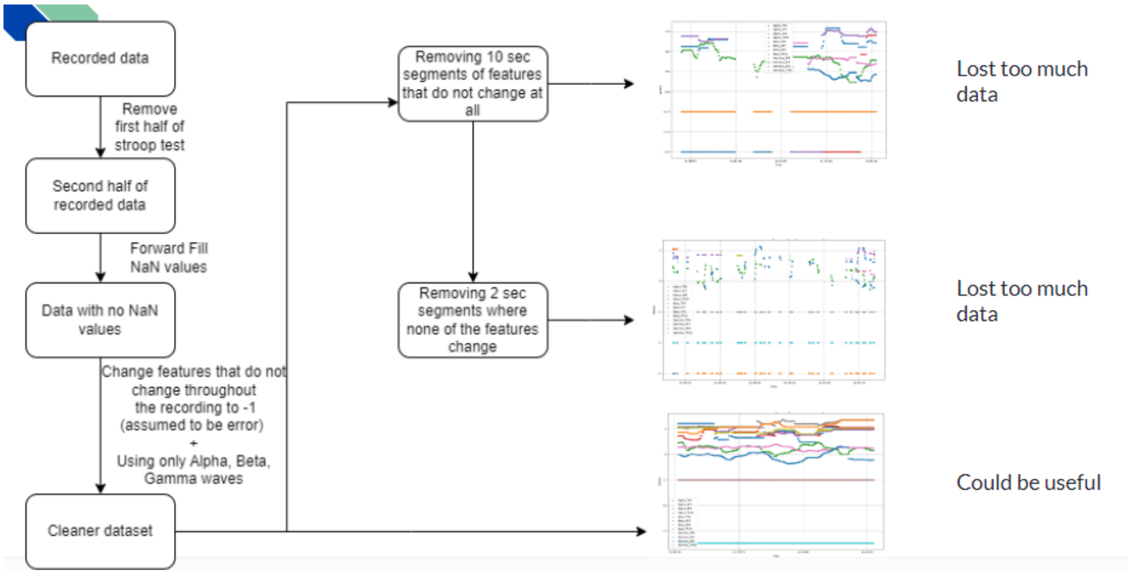

Figure 5: Flowchart of preprocessing and cleaning techniques used on Alpha, Beta, Gamma waves

After trying out various combinations, it was seen that using many cleaning steps led to a lot of data being lost. In the end, the team decided to only use step 1 from the above mentioned actions.

Dataset Creation

After cleaning, it was time to create datasets with features that would help in separating the two classes. Various different methods of dataset creation were tried. To get an initial understanding of their relative performance against each other, they were fitted to a simple ANN that was then used to make predictions on the test set.

Fast Fourier Transformation

Unfortunately, time series data by itself is not very useful. In these cases, time series is often transformed to frequency domain using a process called fast fourier transformation. In theory, any wave can be represented as a sum of many sine waves of different frequencies and magnitude. Finally, these waves are represented in frequency v/s magnitude graphs. This data (the magnitudes at different frequencies) can often turn out to be meaningful.

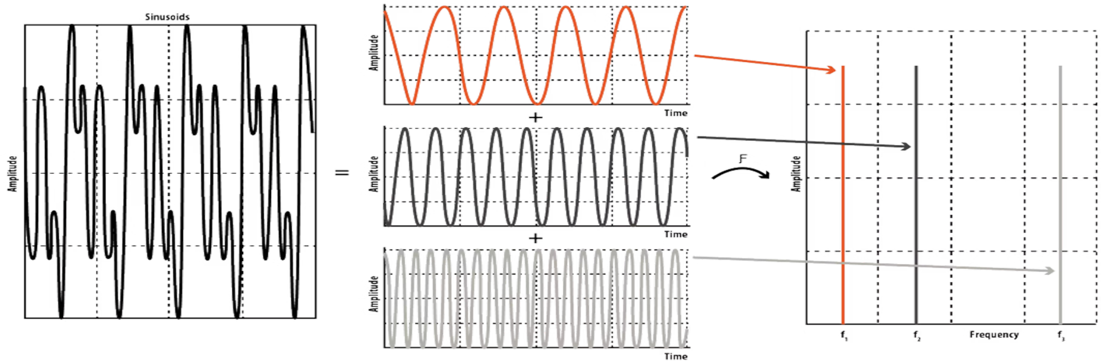

Figure 6: Visualization of how FFT works [9]

For this, the scipy.FFT module's FFT function was used in this project. This module helps generate the magnitudes at different frequencies, where the frequencies are given as an array generated by the following function: `np.FFT.FFTfreq(n_samples, d=1/sampling_rate)`

FFT on entire recordings

Once the data was cleaned, Fast Fourier Transformation (FFT) was run on each recording to generate a row with 60 features. For each of the 12 features, Upon running FFT, the magnitudes were squared to help make big peaks bigger and small peaks smaller.

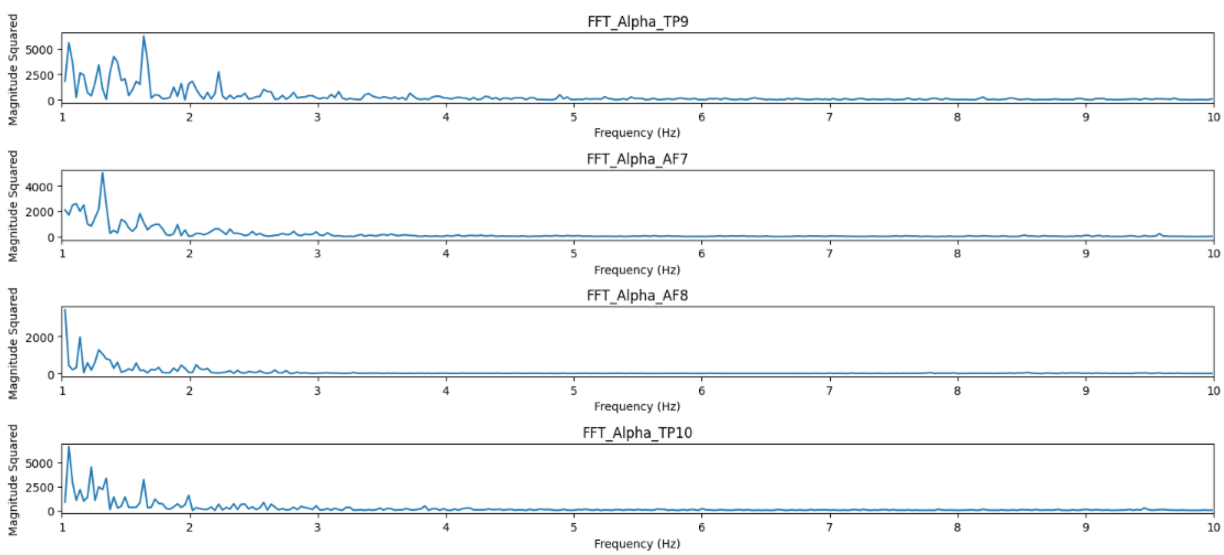

Figure 7: FFT on some of the features for one recording

Surprisingly, all the big peaks for most features were at less than 3 Hz similar to what can be seen in the figure above. The output array from the FFT had squared magnitudes at 8742 different frequencies.

	FFT_Alpha_TP9	FFT_Alpha_AF7	FFT_Alpha_AF8	FFT_Alpha_TP10	FFT_Beta_TP9	FFT_Beta_AF7	FFT_Beta_AF8	FFT_Beta_TP10	FFT_Gamma_TP9	FFT_Gamma_AF7	FFT_Gamma_AF8	FFT_Gamma_TP10
0	3.220015e+07	8.254556e+06	4.021225e+06	2.217615e+07	3.085620e+07	320601.524630	236933.244664	3.755690e+07	1.890081e+07	8.613445e+06	4.886872e+06	2.052234e+07
1	1.057531e+05	2.274592e+04	1.615184e+05	4.228887e+04	2.352866e+04	5943.972037	249083.775247	5.007530e+03	3.061503e+05	2.705469e+05	1.673446e+05	9.497559e+03
2	1.175462e+05	7.449117e+04	3.294887e+04	2.536409e+05	9.019004e+04	53666.159728	16600.106220	3.086102e+03	8.956262e+03	3.032288e+03	2.286502e+04	9.046782e+03
3	1.113597e+05	5.014515e+04	1.655771e+04	3.917185e+05	7.017674e+04	11199.965603	48963.212855	1.026042e+04	4.324970e+03	1.327120e+05	4.478573e+04	1.424844e+04
4	1.052425e+05	8.730798e+04	7.828717e+04	5.730118e+04	1.161038e+05	40059.479225	18256.137869	8.733160e+03	2.400148e+03	5.057377e+03	1.642756e+05	1.572000e+04
...	...	...	...	...	...	...	...	...	...	...	...	...
8737	6.209117e+05	7.801304e+04	4.626813e+03	7.990295e+04	2.311255e+05	6787.206907	16629.736116	2.804574e+03	1.573650e+04	8.359825e+03	8.987426e+04	8.487355e+04
8738	1.052425e+05	8.730798e+04	7.828717e+04	5.730118e+04	1.161038e+05	40059.479225	18256.137869	8.733160e+03	2.400148e+03	5.057377e+03	1.642756e+05	1.572000e+04
8739	1.113597e+05	5.014515e+04	1.655771e+04	3.917185e+05	7.017674e+04	11199.965603	48963.212855	1.026042e+04	4.324970e+03	1.327120e+05	4.478573e+04	1.424844e+04
8740	1.175462e+05	7.449117e+04	3.294887e+04	2.536409e+05	9.019004e+04	53666.159728	16600.106220	3.086102e+03	8.956262e+03	3.032288e+03	2.286502e+04	9.046782e+03
8741	1.057531e+05	2.274592e+04	1.615184e+05	4.228887e+04	2.352866e+04	5943.972037	249083.775247	5.007530e+03	3.061503e+05	2.705469e+05	1.673446e+05	9.497559e+03

8742 rows × 12 columns

Figure 8: Output from FFT

Now since most of the peaks were at very low frequencies, in selecting features for dataset creation, the first 5 magnitudes (squared) values from the FFT on each wave were selected. In total, the dataset had 12 (waves) x 5 (5 smallest magnitudes squared) = 60 features. For each of the 29 total recordings, 1 row with 60 features was added to the dataset. In the end, there were 23 train data and 6 test data.

	Alpha_TP9_0	Alpha_TP9_1	Alpha_TP9_2	Alpha_TP9_3	Alpha_TP9_4	Alpha_AF7_0	Alpha_AF7_1	Alpha_AF7_2	Alpha_AF7_3	Alpha_AF7_4	...
0	3.324147e+07	1.093129e+06	1.452789e+05	4.897626e+04	3.917255e+04	1.003771e+07	7.911451e-28	5.998906e-29	1.255229e-27	2.241625e-27	...
1	3.159305e+07	2.857094e+05	1.602672e+04	4.789029e+04	9.171909e+03	3.705688e+06	7.577700e+04	5.156399e+04	1.368133e+05	2.515583e+04	...
2	1.236385e+07	2.180262e+05	4.958827e+04	7.339902e+04	3.077363e+04	2.531303e+08	5.169442e-24	2.018801e-26	1.261217e-25	2.016753e-26	...

Figure 9: Part of the final dataset created using FFT on entire recordings

Upon using an ANN to train on the train data, it was used to make predictions and correctly predicted 5 / 6 entries. This was a very good result. However, since this model used just 23 train data and 6 test data, it was concluded that this was not enough evidence that the model had good generalization.

FFT on 10 sec windows of recordings

In order to create a more robust and generalized model, the team decided to run FFT on the 12 features but on only 10 second windows. So if a recording was 53 seconds in total, it would create 5 rows in the final dataset (3 seconds at the end not used as a data entry). There were a total of 24480 features generated from FFT, which were then reduced to 38 using PCA. In the end, there were 57 train data points and 17 test data points.

Given the results from the FFT on the entire recordings, there was optimism that this method would also result in good accuracy. However, surprisingly, it only gave a best accuracy of 58%.

Timestep of 20 on raw features

The team also tried creating a data structure with raw features and a TimeGroup of 20 entries. What this means is that, the features for each row were the last 20 entries of the 4 raw features. In total, the final dataset had 80 features. This was done to capture the change of data with respect to time in a very short period (20 rows <0.1 seconds).

	Raw_AF7	Raw_AF8	Raw_TP9	Raw_TP10
1	7_1	8_1	9_1	10_1
2	7_2	8_2	9_2	10_2
3	7_3	8_3	9_3	10_3
⋮				
20	7_20	8_20	9_20	10_20
21	7_21	8_21	9_21	10_21

Figure 10: Example of initial dataset from the EEG headband

	0	1	...	19	20	21	...	39	40	41	...	59	60	61	...	79
1	7_1	7_2	...	7_20	8_1	8_2	...	8_20	9_1	9_2	...	9_20	10_1	10_2	...	10_20
2	7_2	7_3	...	7_21	8_2	8_3	...	8_21	9_2	9_3	...	9_21	10_2	10_3	...	10_21

Figure 11: Example of dataset after creating data structure using TimeGroup

In the end, the train dataset was of size 80x16300 and the test dataset was of size 80x4200. After training it on an ANN, the predictions had an accuracy of 65%.

FFT on 8 second windows of FFT

When it was seen that the raw data had some correlation with the classes, some EDA was done on them. It was seen that most of the Raw features from various recording looked similar to this in the time domain:

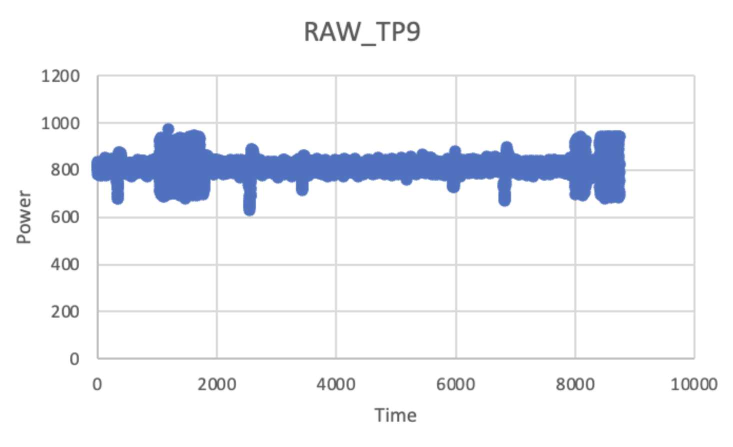

Figure 12: RAW_TP9 in time domain (from one of the train folder recordings

This data by itself does not look very useful, but it could be made useful after running FFT. FFT works very well on data similar to this where there is a lot of noise and variation and the frequency of the wave is very high.

A window size of 8 seconds was used to perform FFT on each of the 4 Raw features received from the headband. This generated a total of 8160 features. There were a total of 57 train data entries and 17 test data.

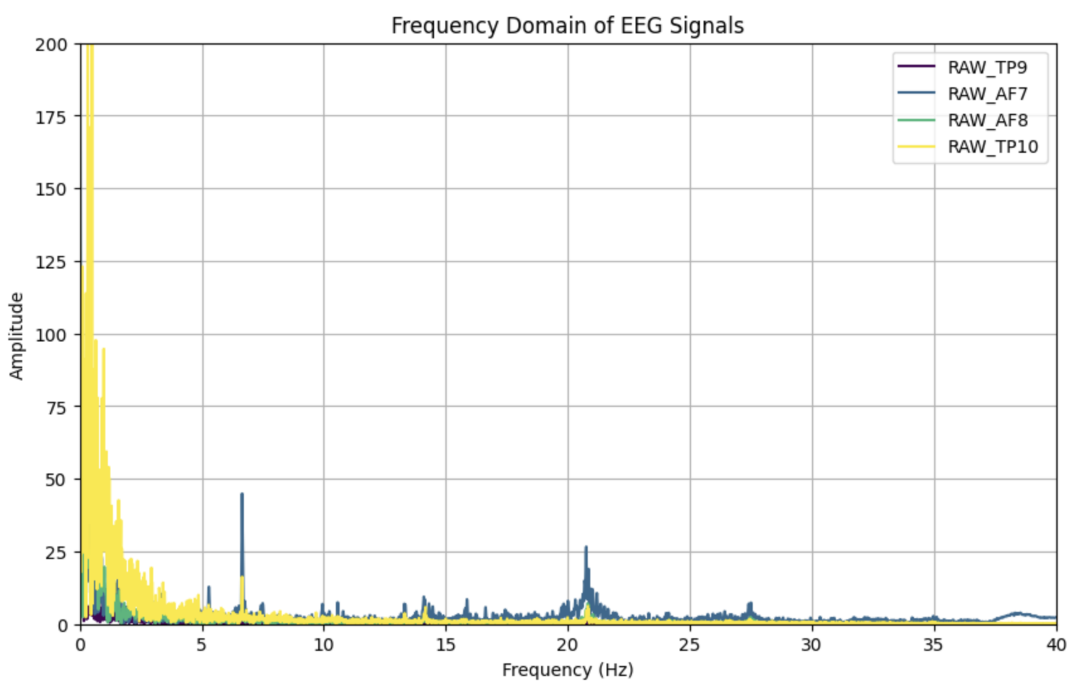

Figure 13: Visualization of raw signals in the frequency domain

After training on ANN and using it to make predictions on the test set, the model gave an accuracy of 70%. Since this gave the best accuracy using the simple ANN, the team tried this dataset on various models, the results of which can be found in the next section.

Dependent variable encoding

The team stored the recorded data in a folder structure like this:

- Train
    - Room
    - Cold
- Test
    - Room
    - Cold

When the various datasets were created, the functions were run on all the different folders and then merged. For example, to create the test set, the same function was run on the Train\Room folder and then on the Train\Cold folder and then merged together. The team made sure to always read the name of the folder the files are in and depending on whether it read ‘Room’ or ‘Cold’, a new feature called ‘condition’ was created which was imputed with ‘0’ or ‘1’ respectively. This feature was then extracted out as the dependent variable at the end before training.

More Data Analysis

Apart from using the models to classify the data, some visual data analysis can also offer insight into the data. The following histogram was created from further processing the data that was fed into the models, with the objective of visually exploring the processed data. The starting datasets used were the 57x8160 training (X_test) and 17x8160 testing (X_test) datasets that contained the squared magnitudes of all the eight-second window FFT. Because this step would not be used to provide data to the models, but rather to provide a holistic view of the cleaned data, the training and testing distinctions were replaced by simply classifying the data as coming from the room or cold temperature tests, and so the two datasets were recombined to make one 74x8160 dataset. From this dataset, the arithmetic average of each row was calculated, which corresponded to averaging the square magnitude found at each frequency within the FFT’s scope. This step reduced the dataset’s size from 74x8160 to 74x1. Using the 57x1 training (y_train) and 17x1 testing (y_test) output vectors, each row’s average was classified as belonging to either the room or cold temperature datasets. This gave two vectors, a 38x1 vector with the averages from the room temperature data and a 34x1 vector with the averages from the cold temperature data. The absolute frequencies of the numbers in these vectors were then plotted in the histogram, with the room and cold temperature data separate from each other.

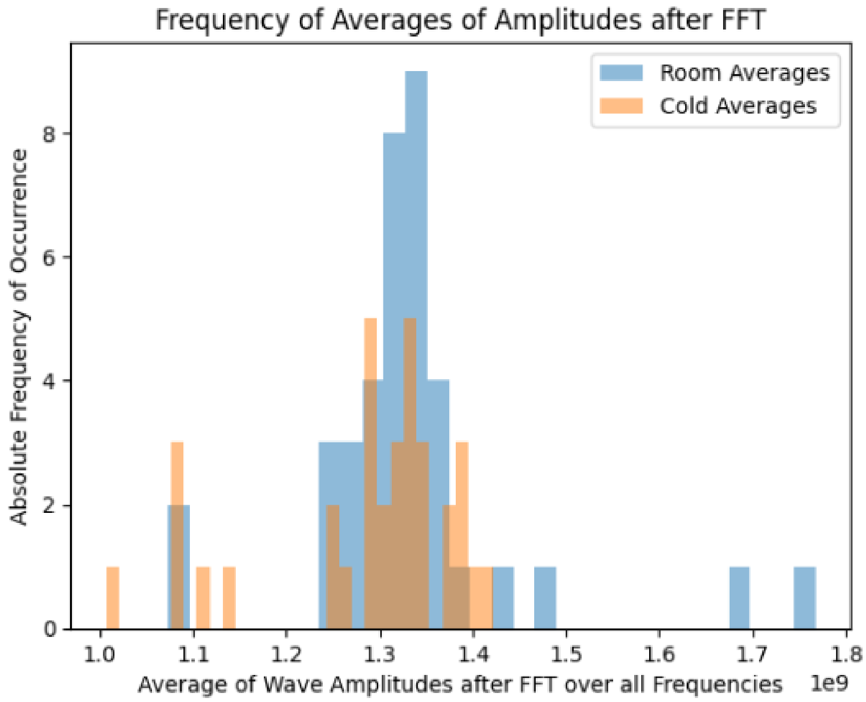

Figure 14: Post-FFT squared magnitude averages classified by room or cold temperatures

The histogram clearly shows that, although there is some overlap between the room and cold averages' magnitudes, there is a greater tendency for the cold averages to have more occurrences with lower magnitudes than with greater magnitudes (less in the y-axis and towards the left in the x-axis), and there is a greater tendency for the room temperature averages to have more occurrences with higher magnitudes than with lower magnitudes (more in the y-axis and towards the right of the x-axis).

Machine Learning Models

A variety of machine learning models on the dataset. The accuracy of each model was measured and compared to each other. To measure accuracy the accuracy_score class from the metrics modulus in the sklearn library was used.

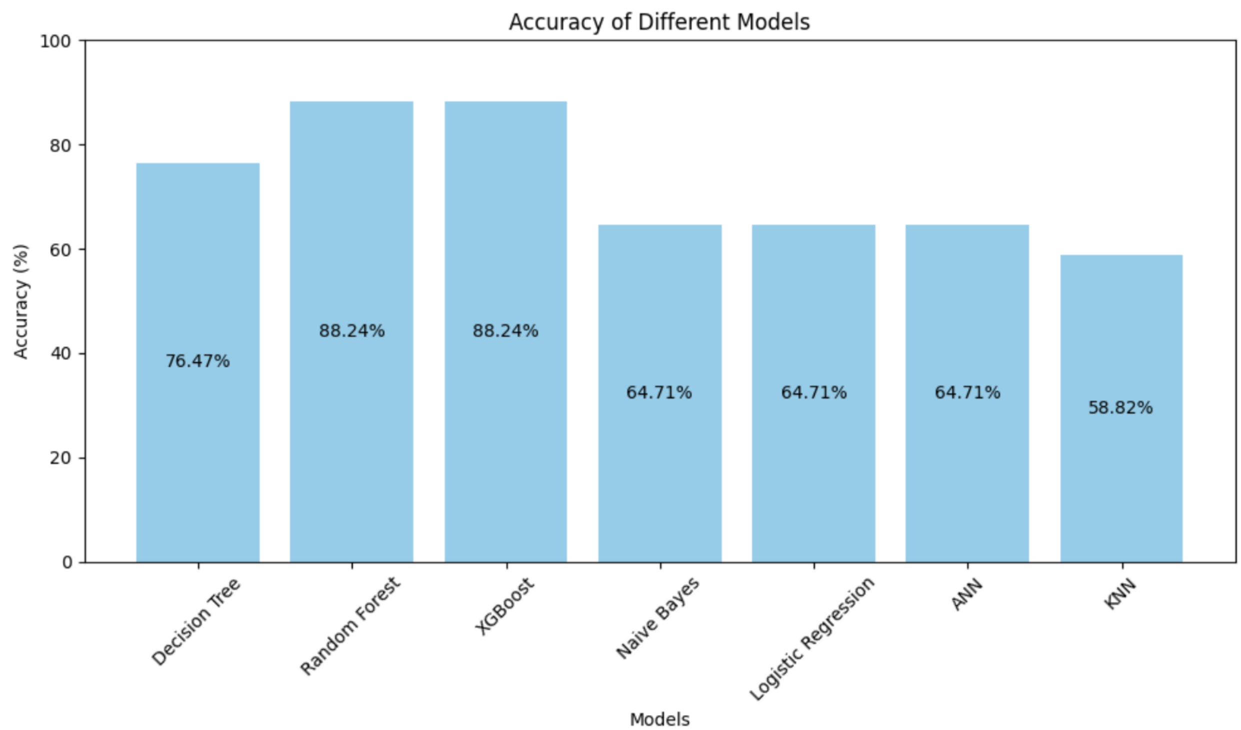

Figure 15: Column Chart showing the accuracies of different machine learning algorithm

Decision Tree

The decision tree is a popular model that is used often in machine learning contexts. A decision tree is a flowchart-like structure where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (a decision taken after computing all attributes). The paths from root to leaf represent classification rules.

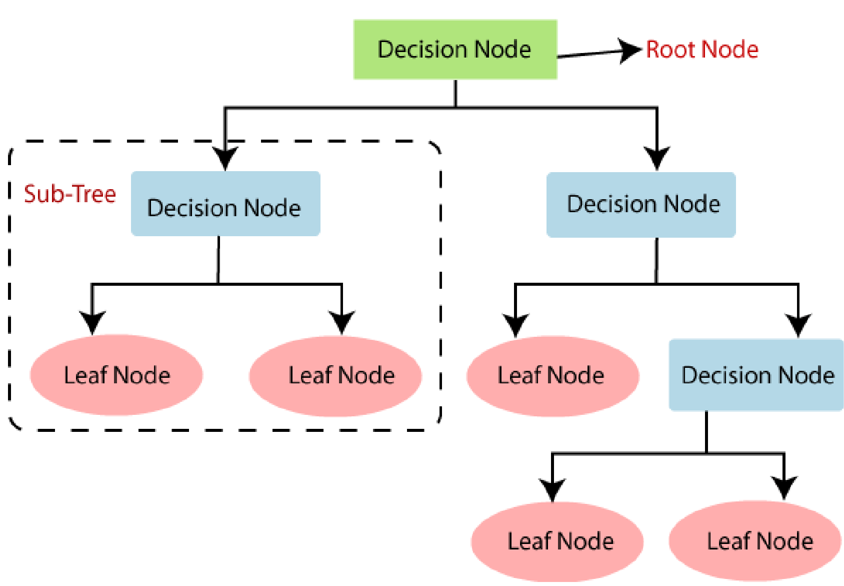

Figure 16: Visual representation of the decision tree structure [10]

The figure below shows the code we used to run the decision tree algorithm to our dataset. This method scored an accuracy of 76.47%.

```
from sklearn.tree import DecisionTreeClassifier

# Creating and training the Decision Tree classifier
dt_classifier = DecisionTreeClassifier(random_state=42, max_depth = 2)
dt_classifier.fit(X_train_scaled, y_train)

# Making predictions
y_pred = dt_classifier.predict(X_test_scaled)

# Evaluating the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Decision Tree Accuracy: {accuracy*100:.2f}%")

Decision Tree Accuracy: 76.47%
```

Figure 17: Code to apply decision tree algorithm

Inspired by our result, we decided to try other models that are built off the decision tree algorithm. These were Random Forest and XGBoost, which are stronger and more complicated versions of the Decision Tree algorithm.

Random Forest

Random forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Essentially, it builds multiple decision trees and merges them together to get a more accurate and stable prediction.

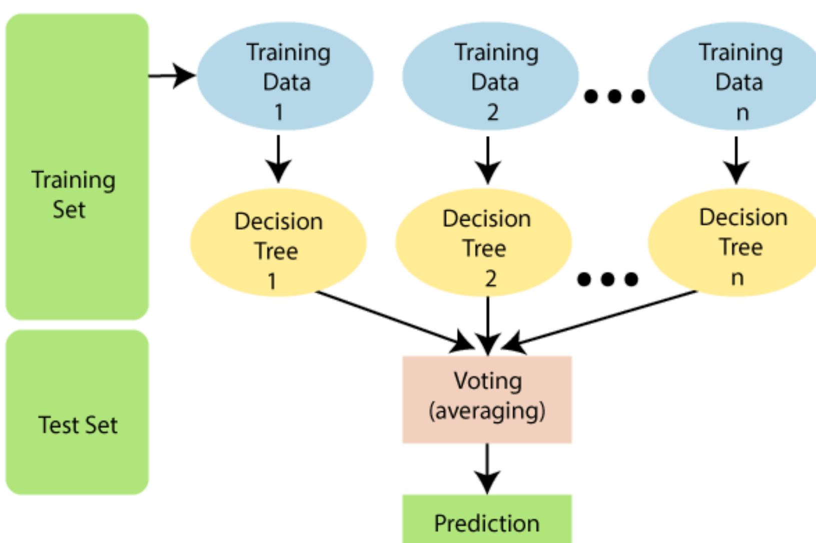

Figure 18: Visual representation of the random forest algorithm [11]

The below figure shows the code we used to run the decision tree algorithm to our dataset. This method got an accuracy of 88.24%.

```
import xgboost as xgb

# Create the XGBoost classifier
xgb_classifier = xgb.XGBClassifier(objective='binary:logistic', random_state=42, max_depth = 2, n_estimators = 100, learning_rate = 0.3)

# Train the classifier
xgb_classifier.fit(X_train_scaled, y_train)

# Make predictions
y_pred = xgb_classifier.predict(X_test_scaled)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"XGBoost Accuracy: {accuracy*100:.2f}%")

XGBoost Accuracy: 88.24%
```

Figure 19: Code to apply random forest algorithm

As expected this accuracy was stronger than the decision tree algorithm because it essentially combines many decision tree algorithms together.

XGBoost

XGBoost (Extreme Gradient Boosting) is another ensemble method, specifically designed to optimize the performance of Decision Trees. It builds an ensemble of Decision Trees sequentially, with each subsequent tree trying to correct the errors made by the previous ones.

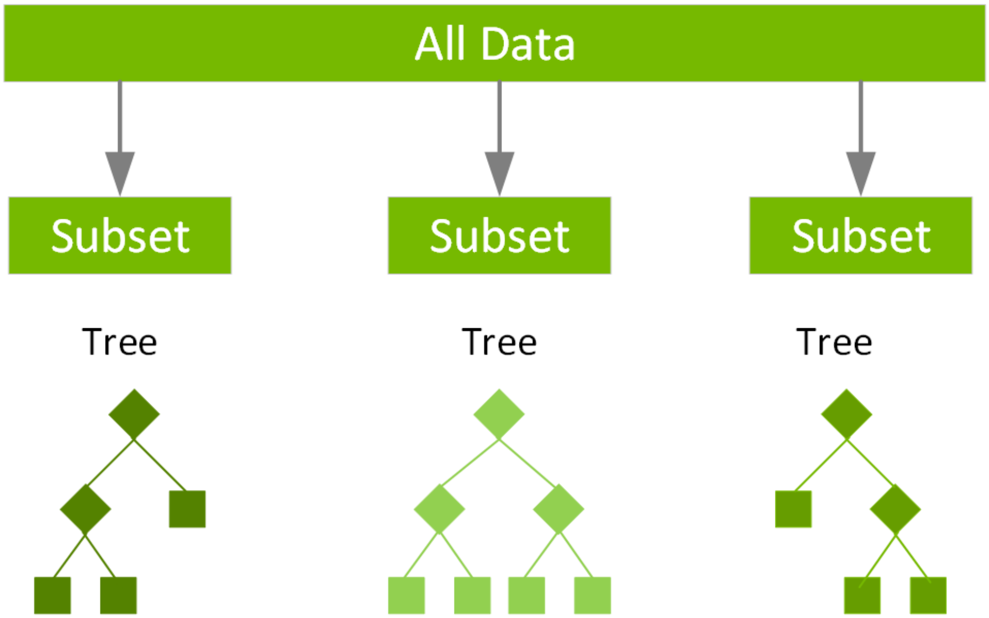

Figure 20: Visual representation of the XGBoost algorithm [12]

The below figure shows the code we used to run the XGBoost algorithm to our dataset. After some testing we realized that the optimal parameters were a max_depth of 2, n_estimators of 100 and a learning_rate of 0.3. This method got an accuracy of 88.24%.

```
XGBoost

[17] import xgboost as xgb

     # Create the XGBoost classifier
     xgb_classifier = xgb.XGBClassifier(objective='binary:logistic', random_state=42, max_depth = 2, n_estimators = 100, learning_rate = 0.3)

     # Train the classifier
     xgb_classifier.fit(X_train_scaled, y_train)

     # Make predictions
     y_pred = xgb_classifier.predict(X_test_scaled)

     # Evaluate the model
     accuracy = accuracy_score(y_test, y_pred)
     print(f"XGBoost Accuracy: {accuracy*100:.2f}%")

     XGBoost Accuracy: 88.24%
```

Figure 21: Code to apply XGBoost algorithm

Naive Bayes

Naive Bayes is a simple yet powerful probabilistic machine learning algorithm that is primarily used for classification tasks. It is based on Bayes' theorem [13] and is considered "naive" because it makes a strong independence assumption between the features used for classification.

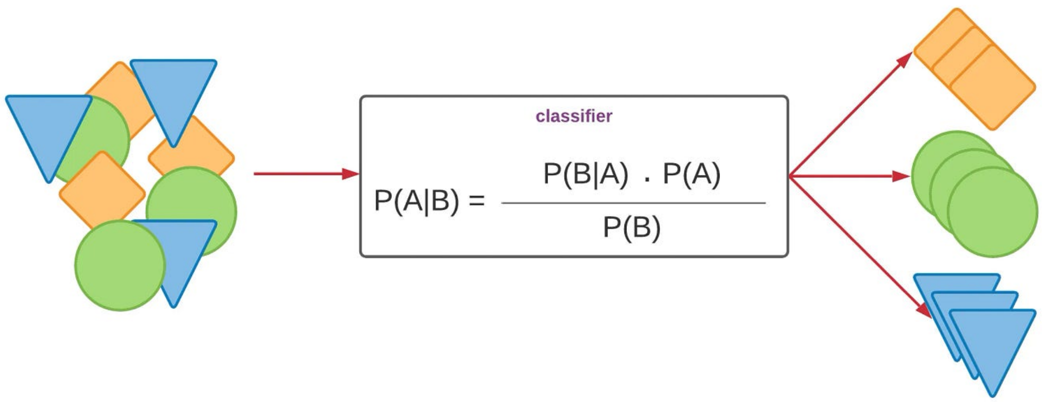

Figure 22: Visual representation of the Naive Bayes algorithm [14]

The below figure shows the code we used to run the Naive Bayes algorithm to our dataset. This method got an accuracy of 64.71%.

```
[24] from sklearn.naive_bayes import GaussianNB

     # Create a Gaussian Naive Bayes classifier
     nb_classifier = GaussianNB()

     # Train the classifier
     nb_classifier.fit(X_train_scaled, y_train)

     # Making predictions
     y_pred = nb_classifier.predict(X_test_scaled)

     # Evaluating the model
     accuracy = accuracy_score(y_test, y_pred)
     print(f"Naive Bayes Accuracy: {accuracy*100:.2f}%")

     Naive Bayes Accuracy: 64.71%
```

Figure 23: Code to apply Naive Bayes algorithm

Logistic Regression

Logistic Regression model used for binary classification tasks, where the goal is to predict one of two possible outcomes (usually denoted as 0 and 1). Despite its name, it is a classification algorithm and not a regression algorithm.

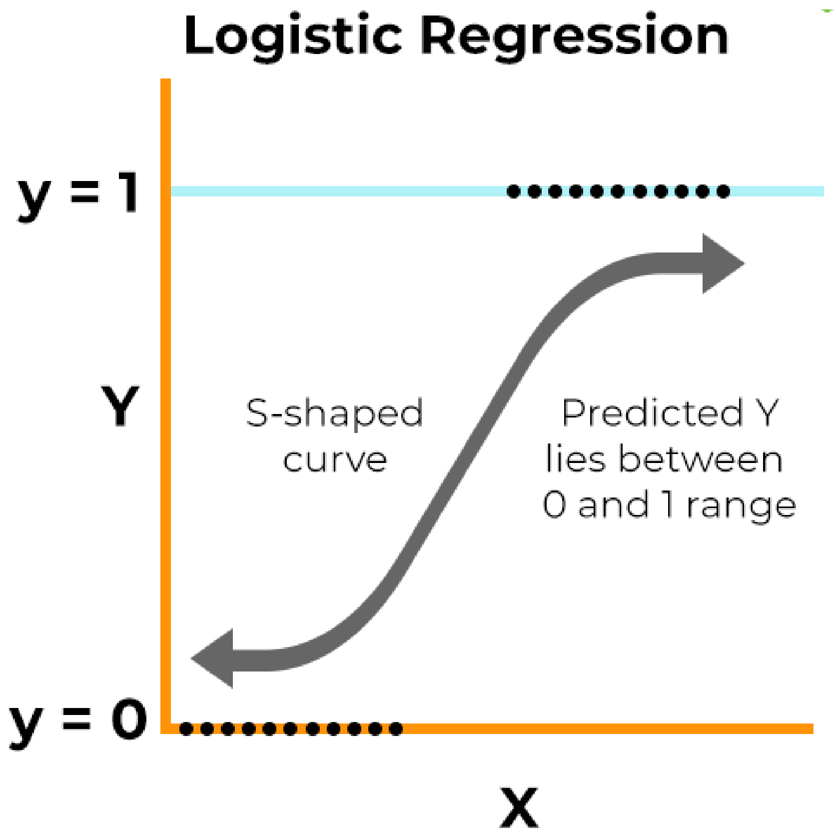

Figure 24: Visual representation of the Logistic Regression algorithm [15]

The below figure shows the code we used to run the Logistic Regression algorithm to our dataset. This method got an accuracy of 64.71%.

```
[19] from sklearn.linear_model import LogisticRegression

     # Creating and training the logistic regression model
     log_reg = LogisticRegression(max_iter=1000)
     log_reg.fit(X_train_scaled, y_train)

     # Making predictions
     y_pred = log_reg.predict(X_test_scaled)

     # Evaluating the model
     accuracy = accuracy_score(y_test, y_pred)
     print(f"Logistic Regression Accuracy: {accuracy*100:.2f}%")

     Logistic Regression Accuracy: 64.71%
```

Figure 25: Code to apply Logistic Regression algorithm

ANN

An Artificial Neural Network (ANN) is a computational model inspired by the structure and functioning of the human brain's biological neural networks. It consists of interconnected nodes (neurons) organized into layers.

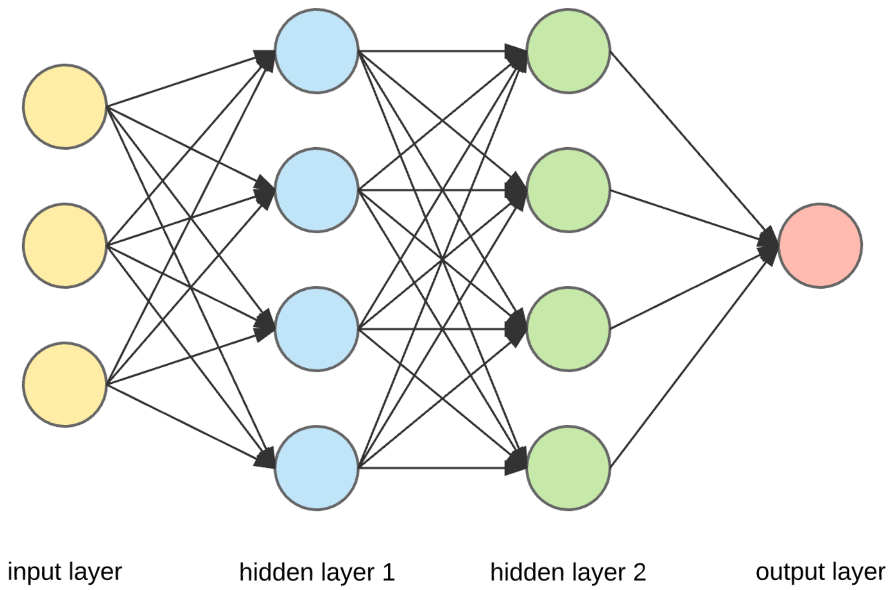

Figure 26: Visual representation of the ANN algorithm [16]

The below figure shows the code we used to run the ANN algorithm to our dataset. This method got an accuracy of 76.47%.

```
[20] from keras.models import Sequential
     from keras.layers import Dense

     # Building the ANN model
     model = Sequential()
     model.add(Dense(128, activation='relu', input_shape=(X_train_scaled.shape[1],)))
     model.add(Dense(64, activation='relu'))
     model.add(Dense(1, activation='sigmoid'))  # Use 'softmax' for multi-class classification

     # Compile the model
     model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

     # Train the model
     history = model.fit(X_train_scaled, y_train, epochs=5, batch_size=32, validation_split=0.2)

     # Evaluate the model
     loss, accuracy = model.evaluate(X_test_scaled, y_test)
     print(f"Test Accuracy: {accuracy*100:.2f}%")

     Epoch 1/5
     2/2 [==============================] - 6s 208ms/step - loss: 0.7421 - accuracy: 0.5556 - val_loss: 1.2731 - val_accuracy: 0.0000e+00
     Epoch 2/5
     2/2 [==============================] - 0s 51ms/step - loss: 0.5178 - accuracy: 0.7333 - val_loss: 0.9079 - val_accuracy: 0.4167
     Epoch 3/5
     2/2 [==============================] - 0s 38ms/step - loss: 0.4370 - accuracy: 0.9778 - val_loss: 1.3574 - val_accuracy: 0.0833
     Epoch 4/5
     2/2 [==============================] - 0s 30ms/step - loss: 0.3572 - accuracy: 0.8889 - val_loss: 1.8687 - val_accuracy: 0.0000e+00
     Epoch 5/5
     2/2 [==============================] - 0s 30ms/step - loss: 0.2957 - accuracy: 0.9333 - val_loss: 1.6705 - val_accuracy: 0.3333
     1/1 [==============================] - 0s 23ms/step - loss: 0.5806 - accuracy: 0.7647
     Test Accuracy: 76.47%
```

Figure 27: Code to apply ANN algorithm

KNN

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for classification and regression tasks. It is a non-parametric and instance-based algorithm, meaning it doesn't make explicit assumptions about the underlying data distribution and instead relies on the data points themselves to make predictions.

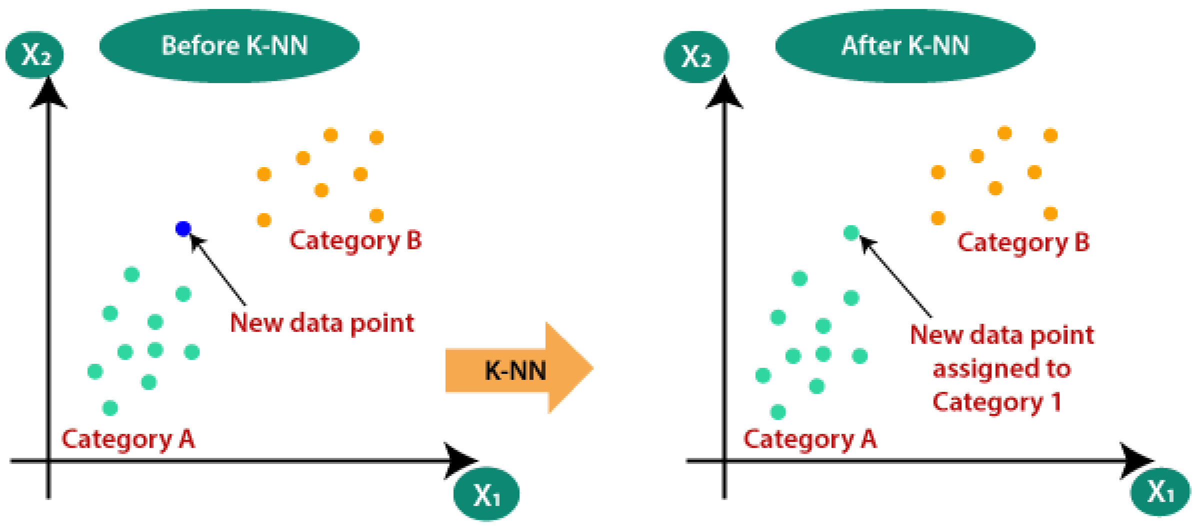

Figure 28: Visual representation of the KNN algorithm [17]

The below figure shows the code we used to run the KNN algorithm to our dataset. This method got an accuracy of 58.82%.

```
[21] from sklearn.neighbors import KNeighborsClassifier

     # Creating and training the KNN classifier
     knn = KNeighborsClassifier(n_neighbors=5)
     knn.fit(X_train_scaled, y_train)

     # Making predictions
     y_pred = knn.predict(X_test_scaled)

     # Evaluating the model
     accuracy = accuracy_score(y_test, y_pred)
     print(f"KNN Accuracy: {accuracy*100:.2f}%")

     KNN Accuracy: 58.82%
```

Figure 29: Code to apply KNN algorithm

PCA

Principal Component Analysis is a dimensionality reduction technique used to transform high dimensional datasets to low dimensional ones. It transforms the data into a new coordinate system, one that allows to capture most of the significant data, while minimizing information loss. This is done by identifying the data's principal components, which are orthogonal axes aligned with the directions of maximum variance within the data.

PCA finds these components by computing the eigenvectors and eigenvalues of the data's covariance matrix. The eigenvectors represent the directions of maximum variance, while the corresponding eigenvalues indicate the magnitude of variance along these directions. By retaining the top principal components that explain the most variance, PCA allows for a lower-dimensional representation of the original data.

The dataset created by running FFT on 10 second windows of raw data had 8160 features. In efforts to reduce the dimensionality of the data, the team tried to run PCA on the dataset using sklearn's PCA function from the decomposition module. If you set the parameter,

$$0 < n\_components \ <= 1 \, ,$$

the library uses the least number of principal components needed to explain that fraction of the total variance.

```
# Applying PCA
pca = PCA(n_components=0.95)  # Retain 95% of the variance, adjust as needed
X_train_pca = pca.fit_transform(X_train_scaled)
X_test_pca = pca.transform(X_test_scaled)

print(f"Original number of features: {X_train.shape[1]}")
print(f"Reduced number of features: {X_train_pca.shape[1]}")
```

```
Original number of features: 8160
Reduced number of features: 40
```

Figure 30: Implementation of PCA

The team set n_components to 0.95. So, the PCA function used the first 40 principal components which in total explain 95% of the total variance in the initial data. However, this does not always result in good datasets. When the team trained and tested all the previously implemented models on this dataset, the accuracies were worse than before.

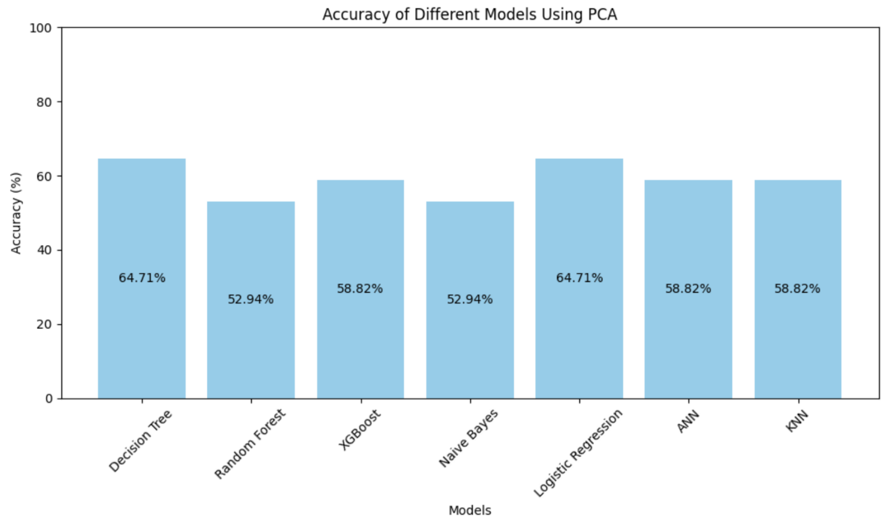

Figure 31: Summary of models

The team ultimately decided to go with a pipeline that did not involve PCA. PCA might not have produced favorable results due to several reasons. While PCA aims to retain the most significant information by capturing high-variance directions in the data, sometimes critical information for the predictive task might not be solely encoded in the directions of highest variance. If the original dataset doesn't exhibit strong linear correlations or if the variance across the dimensions doesn't entirely represent the essential features for prediction, compressing the data onto principal components may lead to loss of discriminative information.

Results

After all the machine learning techniques were tested, the XGBoost technique attained the highest accuracy at 88.24%. (The Random Forest model also had the same accuracy, but for simplicity, XGBoost will be considered the best model.) In industry, an accuracy of 80% is considered good, so the performance of the XGBoost model can be considered more than satisfactory. From this accuracy, it follows that the model can successfully differentiate between test subjects who felt comfortable when doing the Stroop test and those who felt cold, which shows that there is indeed a difference between the brain wave patterns measured by this EEG in the cold and room temperatures.

XGBoost is a technique that improves on the results of the Decision Tree and Random Forest techniques by building upon those results. Based on the explanations provided earlier in the report on each model, the XGBoost model can be considered reliable because, since it is based on the Random Forest model, which is in turn based on the Decision Tree model, it does not require a hyperparameter detailing the data’s distribution, and so the model is free to categorize the data as it sees fit. This is especially relevant due to the large dataset the model had to learn from.

The results from the data analysis, and namely the histogram, reveal the trends of the data. While the results from the models show that they can distinguish between data collected in the cold and data collected at room temperature, in terms of the meaning of the data represented by the histogram, it ultimately shows that, on average, there is more brain activity during the Stroop test when the ambient temperature is room temperature than when it is colder than it. The results from the histogram can be used in conjunction with the results from the models to draw meaningful conclusions about the findings of the project.

Conclusions

The accuracy and trends of the results lead to three conclusions. The first conclusion, which is based on the accuracy of the models, is that the results can be considered meaningful; that is, since the best model's accuracy is almost 90%, the model is better than a random guess on the binary target variable, which would statistically have an accuracy of 50%. Based, then, on the meaningfulness of the best model's accuracy, it can be shown that there is indeed some discernible relationship between the cognitive performance of the test subjects and the ambient temperature.

The above conclusions based solely on accuracy, however, are not enough to confirm the hypothesis, which was that a lower ambient temperature corresponds to a lower cognitive performance when the ambient temperature is under room temperature. This is because the results from the models specifically would be mostly similar if the room temperature and cold temperature labels were interchanged. To help confirm the trends of the hypothesis, the histogram is invoked to show that there is indeed a positive correlation between the temperature and performance - indeed, the lower temperature corresponds with lower overall brain wave activity and the higher temperature corresponds with greater overall brain wave activity.

The article by NeuroHealth Associates is called upon to generalize the positive correlation established through the Stroop test to more general cognitive performance [8]. It explains that Alpha, Beta, and Gamma brain waves most strongly correlate with cognition, which are the brain waves detected by the sensors whose data were used for the models. With this information, since the histogram shows that the average activity of these three waves is more intense in the hotter temperature, the conclusion can be generalized from the Stroop test only to broader cognition, or at the very least cognition that activates mainly Alpha, Beta, and Gamma brain waves.

To establish causality and fully confirm the hypothesis, it must be shown that there were no other factors that contributed to the improvement or decline in cognitive performance. These factors usually come in the form of similarities between the test subjects or the data collection procedure. While the experiment did not account for factors that a test subject could have been preconditioned with, such as time of sleep, neurologically active drugs, or more, the only aspects that were the same for each test subject, apart from the two hotter and colder temperatures, were the order in which the Stroop test was administered and the device used to measure their brain waves. Other factors, such as distraction and noise levels and familiarity with the test, were all different for the test subjects. This is known to be true since, even though all the data was collected in MacMillan building, the test was performed by several subjects during different times and different days in a space (the IGEN Clubroom) that is not private and that generally has fluctuating levels of activity and noise. Additionally, the EEG device used would arbitrarily disconnect and yield worse data, which adds to the randomness of the conditions in which the data was collected.

Though all the data was collected from Integrated Engineering students who are young adults, which is addressed in the following section, causality between temperature can finally be established based on the randomness of the test subjects and of the data collection procedure, as discussed above. Therefore, all the components of the hypothesis are satisfied: the results are significant, meaningful, and indicative of a relationship between temperature and performance, as discussed in the first paragraph; the lower temperature corresponds to lower performance and vice versa, as discussed in the second paragraph; the results can be generalized from the Stroop test to other forms of cognition, as discussed in the third paragraph; and there is enough evidence that no other factors affected performance other than temperature, as discussed in the last paragraph. The conclusion of this report is, thus, that the results of the project confirm the hypothesis with confidence.

Limitations

This project has several limitations, including in its scope and the validity of the data collected. In terms of data collection, this experiment was limited because the data was only electrical potential recorded by a commercial and portable EEG. It is possible that different relationships between cognition and temperature could be discovered using other metrics and techniques, such as magnetoencephalography instead or electroencephalography, or if the device was of medical grade. Also, the researchers were limited because they do not study biology. Another limitation is that, because of the device's sensitivity, the subjects were forced to stay very still to maximize the quality of the collected data, so movement and activity level were not considered. In terms of the data processing, the project is limited because it only explored the data in a few ways and selected the best results, which could have introduced bias. The data used in the end was the squared magnitude of the FFT performed on the raw data after it had been cleaned. It is possible, for example, for a different conclusion to be reached if the magnitude was not squared or if different features or techniques, such as the average or the area-under-the-curve, were used in the models' learning data.

The project is also limited in its scope, since it cannot establish with empirical evidence that the results are generalizable from the Stroop test to other cognitive functions. Rather, this project uses others' well-established knowledge on the general attributes of the five brain waves to come to its specific conclusion through the Alpha, Beta, and Gamma waves' relationship to cognition. Additionally, the experiment could not be feasibly conducted in a hotter ambient temperature than room temperature or in temperatures close to 0 degrees Celsius, and so the results are only valid for temperatures between 21 degrees Celsius and freezing temperatures. The results are also not valid for dangerously low or high temperatures, though this was initially outside of the proposed scope of the project.

Finally, there could be some implicit biases in the data that come from the people who were studied. Though the profiles of the subjects were different and varied, they were few in number and all UBC Integrated Engineering students. Although not very likely, this could potentially affect the data collected in unexpected ways.

Potential Sources of Error

There are several potential sources of error, but their effects are difficult to establish due to the limited time and scope of the experiment. One major source of error was the EEG device. Even though the EEG device is not cheap, it is also not of medical grade, and it still suffered from faulty connections during the data collection process. The following extensive data cleaning process, which involved changing values when there were flat lines where there should have been waves, could have therefore affected the magnitude of the frequencies identified by the FFT and created false relationships where none existed.

Other potential sources of error are the ways the data was processed. Because the project had a limited timeline, it was not feasible to try out more than a few combinations of cleaning and preprocessing techniques. This means that there could be hidden relationships that were not accounted for, or that the dataset that resulted in the 88.24% accuracy was an outlier when compared to other datasets created from different combinations of cleaning and preprocessing techniques. It should be noted, however, that all models resulted in an accuracy higher than 50%, which would be the accuracy expected from random guesses of a binary target variable.

Finally, it must be acknowledged that all of the data collected came from a few Integrated Engineering students at UBC who are young adults. This could potentially mean that the conclusion, which is somewhat generalized, might not apply to people of different programs (in UBC), backgrounds, age groups, cultures, or other dimensions. While it is possible that this might have skewed the data, it is unlikely that this would greatly affect the results for three reasons. One reason is that all the subjects had different lifestyles in dimensions such as sleep, exercise, diet, intelligence, and more despite being in the same program, and so it is unlikely that they all had similarities in these dimensions that introduced false relationships into the data, especially given the high accuracy. More importantly, however, is the reason that the hypothesis aims to establish not absolute cognitive performance as a function of temperature, but rather cognitive performance when a *change* in temperature is involved. Even if the factors mentioned above skew the data, it is more likely that the similarities would affect absolute cognitive performance rather than the differential cognitive performance. The last reason is that the magnitude of the change in cognitive performance is not being accounted for in the hypothesis, and so variations in the magnitude of the difference in performance with respect to temperature should not affect the validity of the results. Despite these reasons, it is still possible that the sample data could introduce biases that might change the results and thus the conclusion.

Further Research

More research should be done to establish stronger relationships between ambient temperature and cognitive performance. This applies to the researchers in this project as well as to future research done by others. There are several areas where the project could be improved in the future, which are based on the limitations and potential sources of error of this project. The quality of the data collected may be improved by reviewing the data collection procedure. For example, data should be free of errors from the sensors to ensure consistency, especially since changing incorrect values may affect the FFT procedure and results. This may be achieved by spending more time collecting the data or by using a device of higher quality. Also, more samples should be taken, and they should be from a different pool of participants, preferably of varying ages and backgrounds. Additionally, more methods of exploring and preparing the data should be employed so that it can be understood in greater depth. Two methods recommended to explore the data are the averaging and area-under-the-curve, which can be employed at different stages of the preprocessing step.

Care should also be taken to properly address incorrect or missing values in the data. In this project, incorrect values were given invalid inputs for the FFT, which could have led to information loss. Techniques for dealing with faulty data in waves should be researched. Additionally, to find more general relationships and to help confirm that room temperature represents a maximum in the cognitive performance versus temperature curve, the effects of temperature should be explored above and below room temperature, with more cognitive tests to generalize the relationships and with a greater degree of detail. Though the findings of this project show that a lower temperature reduces cognitive performance when compared to room temperature, it does not explore the magnitude of absolute or differential performance. Finally, to consolidate the causality between temperature and performance, the biological mechanisms by which temperature affects performance should be explored. This area of study, however, falls under a different umbrella than the researchers' expertise.

Summary

The motivation behind this project was to explore the effects that ambient temperature may have on cognitive performance. Depending on the project's findings, it was theorized that they could have impacts in workplaces, educational settings, and geriatric care, since cognitive performance is crucial for maintaining productivity, learning, and promoting health in these three settings, respectively. The working hypothesis was that for temperatures under a comfortable indoor ambient temperature, or 21 to 22 degrees Celsius, a lower temperature reduces cognitive performance. The relationship between ambient temperature and cognitive performance was explored by measuring the brain's electric response to thermal stimuli using a commercial-grade electroencephalogram (EEG), the Muse 2.

The literature review revealed that the Stroop test can reliably measure overall cognitive performance, and so the Stroop test was selected as a representative measure of performance for the purposes of the project. The Stroop test is a timed activity in which the subject must, as quickly as possible, first read names of colors that match the color of their own font (e.g., Red), called the congruent stimuli, and then read names of colors that do not match their own font (e.g., Red), called the incongruent stimuli. The congruent stimuli are easier and quicker to read than the incongruent stimuli and do not require much processing, so the data collected during the congruent test did not form part of the data explored. The relationship between temperature and performance was instead explored through the data collected during the incongruent test, since it is a more cognitively taxing task. The literature consulted also explained that out of the five types of brainwaves, those most associated with cognition were the Alpha, Beta, and Gamma frequency ranges, and so data in the Delta and Theta ranges was ignored.

After the data collection process, the data was cleaned from incorrect information caused by faulty connections between the subjects' skin and the sensors. The different datasets created by cleaning the data differently were tested in a simple ANN model to find the best results. After the best cleaned dataset was selected, several ways to conduct the FFT were tested using a similar ANN model. First, the FFT was performed on entire test recordings, which could last over 20 seconds. Then, the FFT was tested on 10- and 20-second timesteps. Finally, the FFT was performed on raw data with eight-second windows, which yielded the best results. Several datasets based on the best cleaned dataset were tested in this process.

Through more testing, it was established that squaring the results of the eight-second window FFT gave the best visually meaningful results. With this data, averaging was performed to find relationships between temperature and performance, and the positive correlation between the two was observed, as hypothesized. This data was fed into the models - the Decision Tree, Random Forest, XGBoost, Naive Bayes, Logistic Regression, ANN, and KNN models - and the XGBoost and Random Forest models gave the best accuracy of 88.24%. Principal Component Analysis was used to reduce the dimension of the rather large dataset, but it ultimately yielded worse results than the large dataset.

The last section discussed the significance of the results and the conclusions, and it showed how the hypothesis was confirmed: the results were meaningful; a relationship between temperature and performance in the Stroop test was established; the results were broadened from the Stroop test to a wider range of cognitive performance; a positive correlation between temperature and cognitive performance was shown; and causality was established by showing that there was little chance of similarities between test cases to be present. The limitations of the project and potential sources of error were also discussed, among which were the small and rather uniform sample size. Finally, some suggestions were made for future research that could improve and validate the results and conclusions of this project.

Appendix

Copy of Experiment Data

References

[1] Eric H. Chudler, "Stroop Effect," Washington.edu, 2019.
https://faculty.washington.edu/chudler/java/ready.html

[2] O. Seppänen, W. J. Fisk, and Q. Lei, "Effect of Temperature on Task Performance in Office Environment," indoor.lbl.gov, 2006.
https://indoor.lbl.gov/publications/effect-temperature-task-performance

[2] Effect of temperature on task performance in office environment | Indoor Environment. (n.d.). https://indoor.lbl.gov/publications/effect-temperature-task-performance

[3] Heschong, L. (2002). Daylighting and human performance. ASHRAE Journal, 44(6), 65.

[4] J. Van Hoof and J. K. Kazak, "Thermal comfort, Perceived Air quality, and Cognitive Performance When Personally Controlled Air Movement Is Used by Tropically Acclimatized Persons," Indoor Air, vol. 28, no. 1, pp. 318–329, 2018.

[5] Wright, H. E., McLellan, T. M., & Larose, J. (2017). The role of human thermoregulation in regulating cognitive performance. Journal of Thermal Biology, 69, 55-63.

[6] Niedermeyer, E., & da Silva, F. L. (2005). Electroencephalography: Basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins.

[7] Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643.

[8] "The Science of Brainwaves - the language of the brain," NeuroHealth Associates, https://nhahealth.com/brainwaves-the-language/#:~:text=The%20raw%20EEG%20has%20usually (accessed Nov. 18, 2023).

[9] S. L. Gjelstrup, "FFT Analysis (Fast Fourier Transform): The Ultimate Guide to Frequency Analysis," dewewsoft, https://dewesoft.com/blog/guide-to-fft-analysis (accessed Dec. 3, 2023).

[10] "Decision tree algorithm in Machine Learning - Javatpoint," www.javatpoint.com, https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed Dec. 1, 2023).

[11] "Machine learning random forest algorithm - javatpoint," www.javatpoint.com, https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed Dec. 1, 2023).

[12] "What is XGBoost?," NVIDIA Data Science Glossary, https://nvidia.com/en-us/glossary/data-science/xgboost/ (accessed Dec. 1, 2023).

[13] "Bayes' theorem," Wikipedia, https://en.wikipedia.org/wiki/Bayes%27_theorem (accessed Dec. 1, 2023).

[14] Ajaymehta, "'demystifying Naïve Bayes: Simple yet powerful for text classification,'" Medium, https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7 (accessed Dec. 1, 2023).

[15] V. K. A. Researcher et al., "Logistic regression: Equation, assumptions, types, and best practices," Spiceworks, https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/ (accessed Dec. 1, 2023).

[16] A. Dertat, "Applied deep learning - part 1: Artificial Neural Networks," Medium, https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (accessed Dec. 1, 2023).

[17] "K-Nearest Neighbor(KNN) algorithm for Machine Learning - Javatpoint," www.javatpoint.com, https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Dec. 1, 2023).