**IBM-312**

**Data Mining For**

**Business Intelligence**

Group Assignment

Classification of Confused Student

Using Electroencephalogram (EEG)

Brain Waves



Ananya 20115013

Arjun Angirishi 20112023

Keshatwar Pratham Naresh 20115052

Prafful Varshney 20115088

Pranav Baraskar 20115092

**Table of Contents**

**Motivation 3**

**Data Specifications 3**

**Methodology 4**

**Results 8**

**Motivation of the problem**

In recent years, advancements in neuroscience and machine learning have paved the way for new approaches to understanding the human brain. With the rise of big data and the proliferation of neuroimaging technologies, it is now possible to study brain function and behavior in unprecedented detail. One area of particular interest is the classification of confused students based on EEG brainwave data. By analyzing EEG signals, researchers hope to gain insight into the cognitive processes underlying confusion and develop new tools for identifying and supporting struggling learners. In this report, we explore the use of four different machine learning algorithms - XGBoost, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Long Short-Term Memory (LSTM) networks - for classifying confused students using EEG data. Our findings suggest that these algorithms hold promise for accurately identifying and supporting students in need of additional assistance, opening up new avenues for research and innovation in education and neuroscience.Concentrating efforts in this field can prove to be fruitful in the field of psychology for providing better assistance to the students in need and diving deeper into the common patterns in those students struggling so that better therapeutic and neurological treatments can be given.

**Dataset Specifications**

The dataset consists of two different files, one containing the EEG data recorded from 10 students and the other file containing the demographic information for each student. We found out that even the demographic information was impacting the output labels and thus we merged both the datasets before training the models on it.





We used different methods like correlation matrix and scatter plots to find if any of the features were redundant. 

None of the features was correlated enough to become redundant. Also no feature was having a high correlation with the output label.

**Methodology**

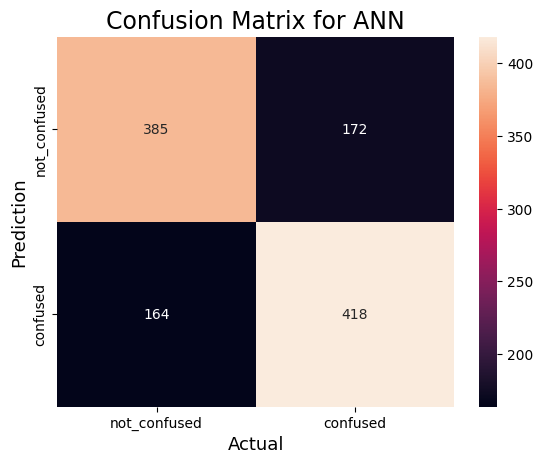
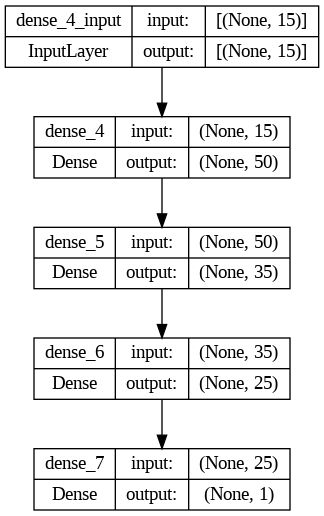
We used the following for comparative study of the performance of different machine learning models on classification of the confused students based on the EEG data:

* Artificial Neural Networks (ANN)
* SVM (using polynomial and gaussian kernel)
* Xgboost algorithm
* LSTM

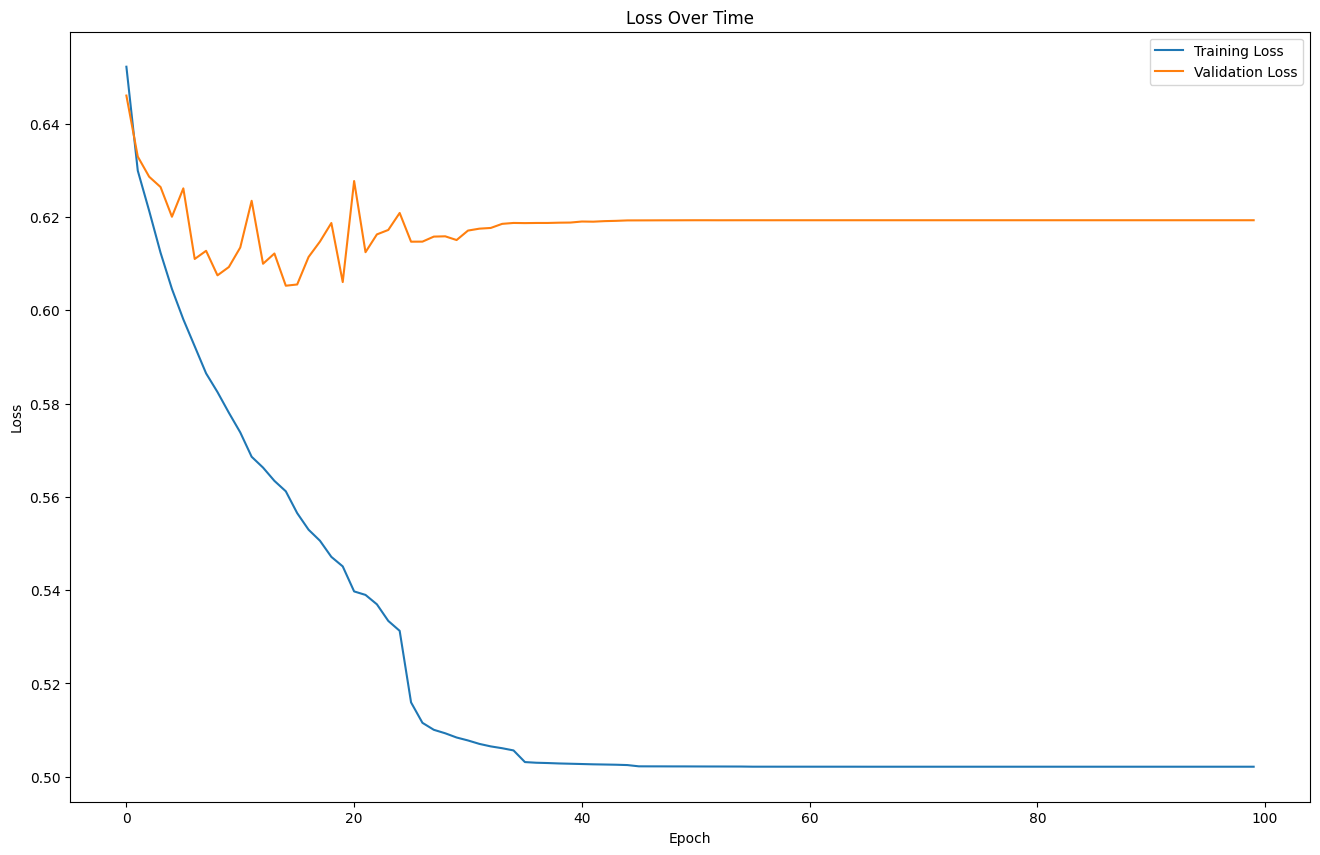
**Artificial neural network (ANN)**

The following ANN architecture was used which also included batch normalization layers in between to improve the performance of gradient descent. Relu activation function was used in the hidden layers and sigmoid for the output layer. The model was trained using binary cross entropy loss and callbacks such as early stopping and model checkpointing were utilized to reduce overfitting and obtain optimal model weights. The ‘ReduceLROnPlateau’ was also employed with an initial learning rate of 0.001 that decreased at a rate of 0.1 everytime the validation accuracy failed to increase for more than 20 epochs.

Backpropagation algorithm was utilized to train the model with Adam optimizer instead of the stochastic gradient descent optimizer for 100 epochs. The model has 4 layers with 3511 trainable parameters.

**Accuracy for ANN: 70.50043898156277%**

****

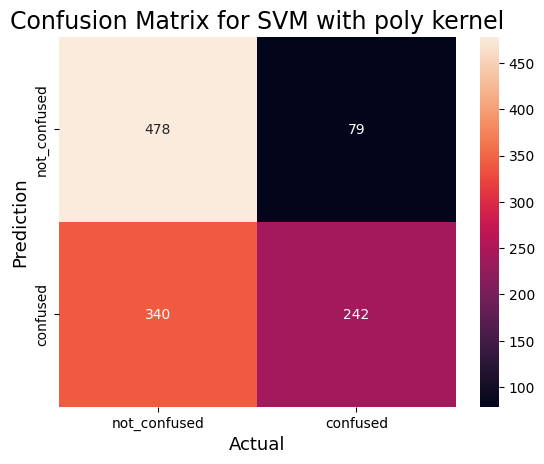
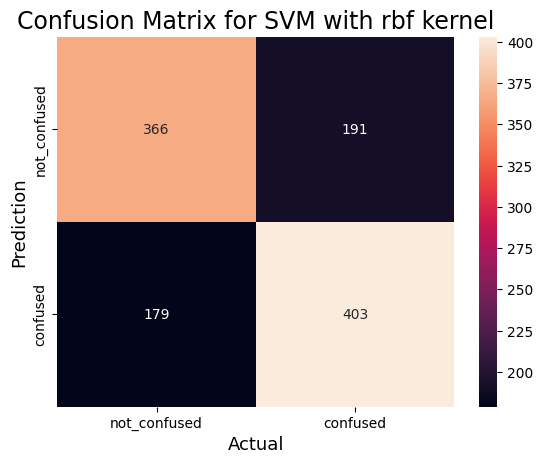
**SVM**

It is a popular supervised machine learning algorithm that is primarily used for classification tasks but can also be used for regression tasks. SVMs handle non-linearly separable data by transforming the data points to a higher-dimensional space where the data becomes linearly separable, using a technique called kernel trick. This is achieved by defining a kernel function that takes as input a pair of data points and outputs their inner product in the higher-dimensional space. There are various types of kernel functions, such as linear kernel, polynomial kernel, Gaussian (or radial basis function) kernel, and sigmoid kernel.

The RBF kernel is defined as:

Where, ‘σ’ is variance and ||*X₁ - X₂||*  is euclidean distance between two points X₁ and X₂.

The polynomial kernel is defined as: 

where, a is constant term and b is degree of kernelWe tried SVM with Gaussian (rbf) kernel and polynomial kernel with degree = 5, and obtained following results.

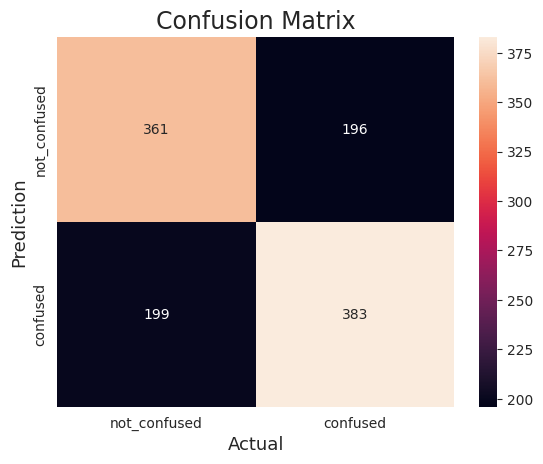
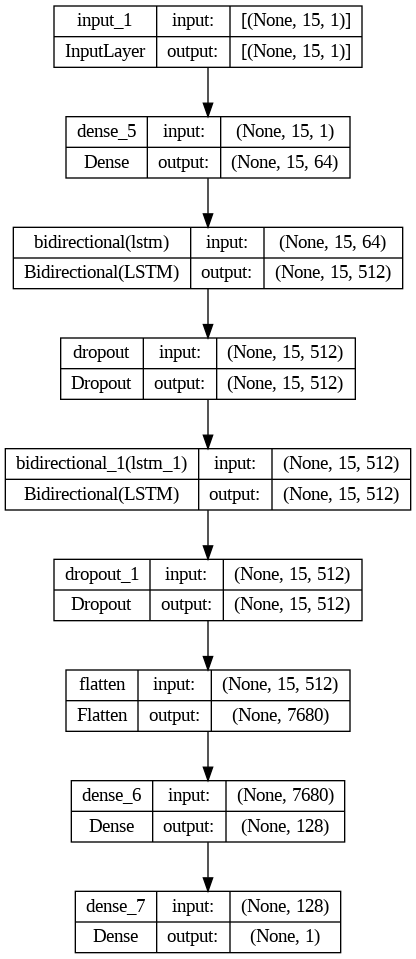
**Accuracy for SVM with rbf kernel: 67.5153643546971%**

**Accuracy for SVM with polynomial kernel: 63.21334503950834%**

**LSTM**

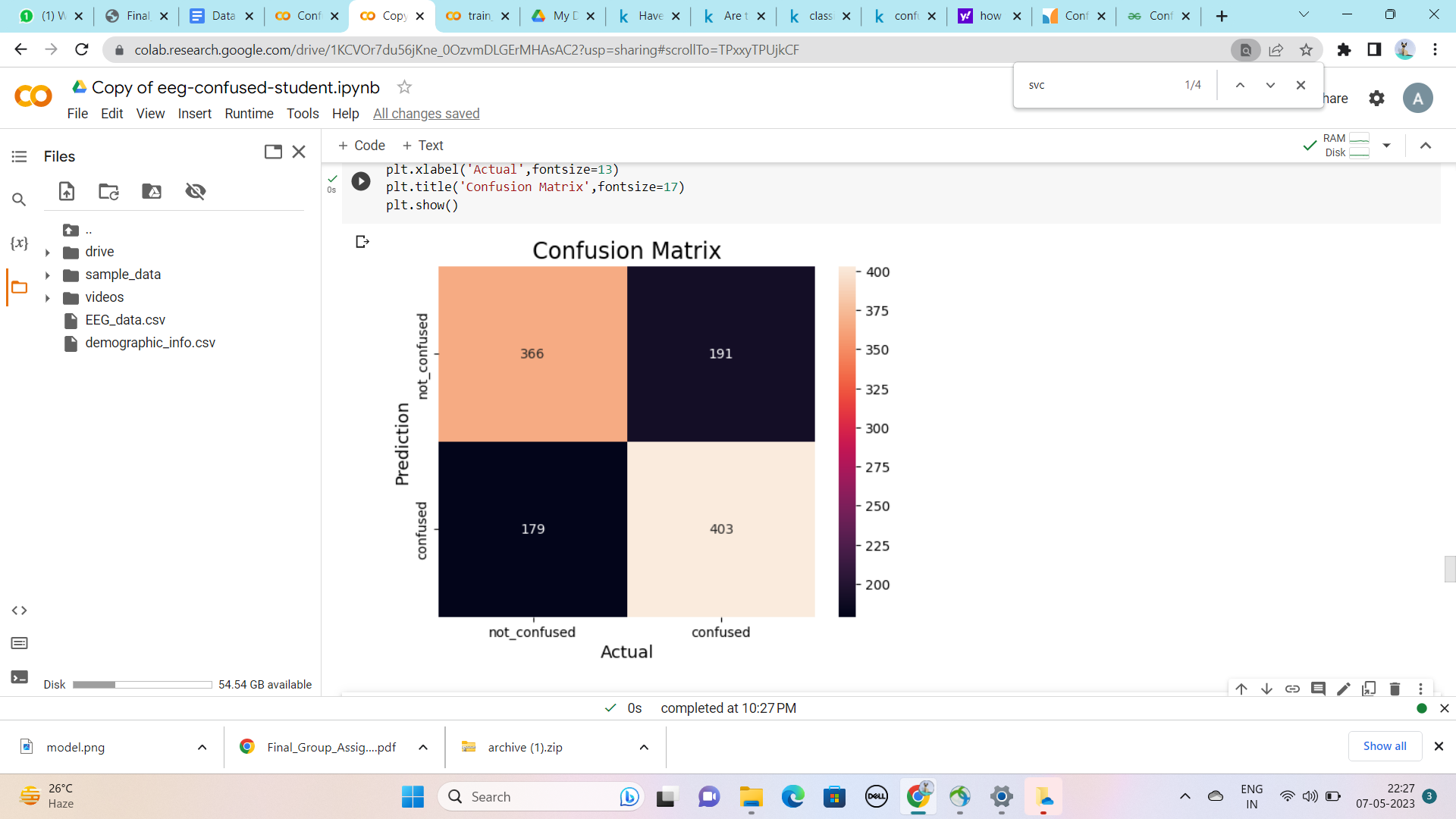
LSTM (Long Short-Term Memory) networks are well-suited for classifying sequential data due to their ability to capture long-term dependencies and handle variable-length sequences. Whether the data is in the form of text, time series, or any other sequential format, LSTMs can effectively learn patterns and make predictions based on the sequential nature of the data. Bidirectional LSTM (BiLSTM) is an extension of the traditional LSTM architecture that incorporates information from both past and future contexts when processing sequential data. While a standard LSTM processes the sequence in a forward direction, a BiLSTM processes the sequence in both forward and backward directions simultaneously.

We trained a LSTM model with 16 batch size with 100 epochs on an adam optimizer ,we got an accuracy of 65.10% on testing data with a peak accuracy of 71% during training on validation dataset.Following is the structure of our mode.



**Accuracy for LSTM model is 65.32045654082529%**

**XGBOOST**

Xgboost is one of the best performing ensemble machine learning techniques and hence we tried using it on our dataset. We used a lot of decision tree stumps (5000) as our model was not overfitting and the validation error reduced further.

Confusion Matrix

**Accuracy for Xgboost model: 69.53467954345918%**

**RESULTS**

All the models performed almost equally well with ANN achieving the highest accuracy of 70.50%. Even after using complex architectures, the models did not overfit which shows that the data is very diverse and it is difficult to find direct relationships between the features and the output.