RESEARCH PAPER

AI Detection of Emotions through EEG

Srinivas Sriram[1] and Nima Leclerc, Mentor[2]

[1]Westford Academy, Westford, 01886, MA, USA
[2] Inspirit AI, USA
Author for correspondence: Srinivas Sriram, Email: srinivassriram06@gmail.com.

Abstract

Electroencephalogram (EEG) brain waves are an extremely important resource for clinical analysis and diagnosis of mental health and other psychiatric diseases. They help current diagnosis of these diseases, as clinics use the external behavior of a patient and the psychiatrist's qualitative reasoning to determine what disease the patient has. However, this approach does not examine internal activity in the brain, and is not uniformly quantifiable. EEG provides a means for psychiatrists to do this, however, this process can be complicated because the variability in EEG signals. Machine Learning (ML) provides a means to analyze these complex data. A computational tool that can recognize patterns in EEG signals would be a valuable starting point in diagnosis psychiatric disorders. We proceeded to develop a recurrent neural network (RNN) model structure that could analyze an individual's brain waves and predict their emotional condition. We used the EEG Brainwave Dataset: Feeling Emotions dataset, which contains various statistical features of the EEG brain waves of two different test subjects as they were experiencing different emotions. We extracted frequency domain features, converted them to the time domain, and filtered them. We trained a gated recurrent unit (GRU) model from these filtered and attained a 96.95% validation accuracy.

Keywords: AI Brainwave Detection, RNN, Emotion

Introduction

There has been skepticism as to the reliability of mental health diagnosis because of how the approach is more qualitative based on the patient's symptoms Novak 2015. To address this, EEG brain waves can be used. Electrical signals are transmitted through different brain regions as a response to sensory stimuli. The EEG signals directly correspond to these signals, and the manner in which the brain communicates different emotions can be extracted and analyzed in order to extract emotional features Li TM. Chao HC. Zhang 2019. However, EEG signals are complex and consequently, require sophisticated models for interpretation. EEG signals have multiple frequency components, hence it is easiest to analyze these signals in the frequency domain. A standard convention is to separate these into different frequency regions: delta, theta, alpha, sigma, and beta waves. For this study, we will focus on alpha waves, which occur at frequencies between 8 and 12 Hz, because these are most prevalent in adults. A core part of our methodology is developing a method to distinguish these types of waves based on their signal properties.

Here we are aiming to build off this principle to discriminate psychiatric states from EEG signals using ML. Previously, other researchers used ML to develop their own solutions to this problem. A commonly used database for EEG signals is the Database for Emotional Analysis using Physiological Signals Dataset (DEAP). Researchers extracted features from raw EEG recordings. These features were fed into machine learning algorithms to classify the emotions associated with these recordings Nandi and Li 2016. This approach worked well due to how we preprocessed the data, using the feature extraction methods. However, the model architecture they used was a basic k-nearest-neighbors model, which is not that useful in real-world scenarios because it does not account for the sequential nature of the signals processed. In another approach, the same DEAP dataset was used, but this time a long-short-term-memory model (LSTM) was used with no feature extraction. This approach led to a high accuracy, but without the use of feature extraction or data preprocessing, there was still room for improvement Alhagry Salma 2017.

To distinguish this research, we are going to build off the work from other researchers to build the best solution. A lightweight recurrent neural network (RNN) based model needs to be used because RNN models tend to be more capable of handling signals and sequential data through their algorithms Vijay 2016. As shown from the approach with the LSTM model, a higher accuracy was achieved that simply a k-nearest neighbors model because of the LSTM's ability to handle more complex data as presented with EEG signals. Additionally, selecting and implementing the correct features is a method that needs to be done effectively. This helps narrow down what the model needs to learn from the complex EEG data.

By developing this algorithm, we could create a starting point to solving the problem of not being able to interpet complex EEG data to diagnose psychiatric diseases. This algorithm could be implemented in a cost-effective EEG headset that could be distributed to clinics, who would then use these headsets to use quantifiable evidence to diagnose these mental health disorders.

Dataset and Feature Extraction

We used the EEG Brainwave Dataset: Feeling Emotions dataset J. J. Bird L. J. Manso and Faria 2018; Buckingham and Faria 2019. Here, two test subjects, male and female, had their EEG brain signals recorded using a Muse EEG headset while watching different movie clips that evoke different emotions, either positive or negative. Neutral resting data was recorded for these test subjects J. J. Bird L. J. Manso and Faria 2018; Buckingham and Faria 2019.

After, the alpha, beta, gamma, theta and delta waves were extracted. Statistical feature extraction was performed, resulting in a total of 2548 features extracted, each of which represents a column of the dataset, and the final column is the label of all the features associated with that row: either positive, negative, or neutral J. J. Bird L. J. Manso and Faria 2018; Buckingham and Faria 2019. Each row represents an EEG signal gathered for both subjects, there are a total of 2132 rows.

We selected the features related to the fast-fourier-transform (FFT) of the brain signals, due to the fact that FFT provides meaningful information of the wave itself and does not summarize the data presented in the wave like the other statistical features. Each feature is a frequency component of the FFT. The 750 features related to FFT were extracted for each subject, resulting in a total of 1,500 features or columns used.

We brought the frequency dependent features to the time domain by performing an inverse FFT. Analyzing the signals in the time domain allowed us to use more sophisticated filtering capabilities and to leverage sequential ML models on the signals. The comparison of Figure 1 and 2 show how one signal looks before and after applying inverse FFT, where Figure 2 resembles a brain signal and its characteristics. A butterworth filter was applied to the inverse FFT signals. The signals experience a lot of high frequency noise, which we were able to filter out with the butterworth filter.

These frequencies needed to be eliminated because they disturb the pattern of the brain signal. The butterworth filter eliminates these high frequencies by ensuring that if the frequency is higher than the cutoff frequency, the signal rolls downward depending on the order of the filter. A first order filter was applied with a cutoff frequency of 0.5 Hertz. Figure 3 displays the signal in Figure 2 after the butterworth filter is applied to the signal, with the result being that the signal is smoother with less high frequency noise. The inverse FFT and butterworth filter were applied to each of the 2132 signals.

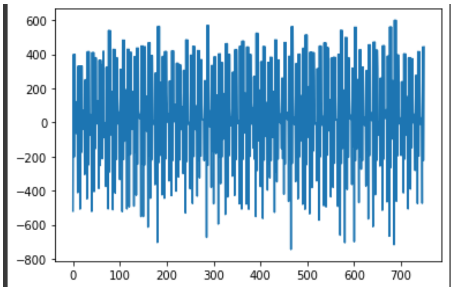


Figure 1. Here, the raw EEG signal of one of the features is selected. There is lots of noise and the pattern of the signal is hard to distinguish.

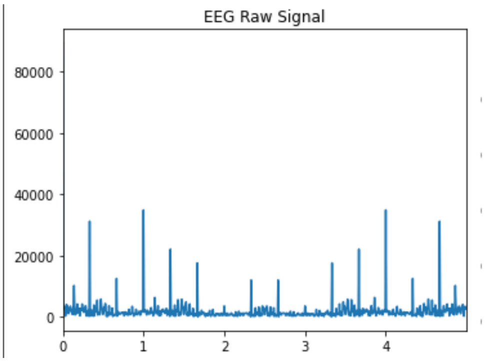


Figure 2. Here, the EEG signal is converted to time domain using IFFT. The signal is easier to distinguish but there is still noise present in the signal.

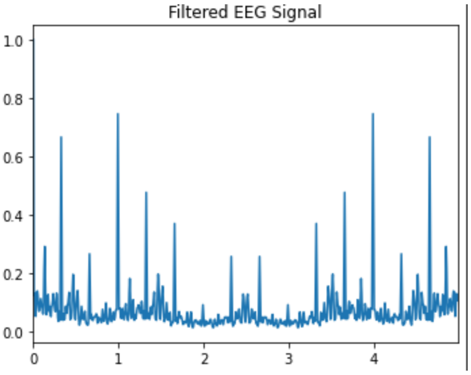


Figure 3. Here, the IFFT signal is presented after applying a butterworth filter, making the signal more curved with a defined peak and pattern.

Following preprocessing, the dataset was split into the following for model training: 80% training data, 10% validation data, and 10% testing data. Most of the dataset was dedicated to training because the model needed to learn as many of the actual signals as possible to get adjusted to the variety of patterns that could exist for a particular type of emotion. However, a sufficient amount of validation and testing data was provided to ensure that the model could evaluate its learning on unseen data during and after training.

To train the model, dummy data had to be applied to the y, or label, data. This was done because the label data was only one dimension, but the model architecture required the output of the algorithm to be two dimensions, like the shape (None,3). Thus, the dummy data applied to the label arrays made it a two-dimensional array and compatible for model training.

Methodology and Models

Modeling could be done on the reformatted data to try and extract meaning. To see how well a simple algorithm could learn from the new dataset, we trained a k-nearest-neighbors (KNN) model on the data. The k-nearest-neighbor model learns the patterns of the dataset by taking a data point, in this case, a filtered signal, and seeing how close that data point resembles other data points or signals whose class is known. The k-nearest-neighbor model was trained with the training data, with a specified 5 neighbors per data point for the model to evaluate patterns from.

After training this model, the test set was used to evaluate the model's performance on new data. The KNN model could predict 98.13% of the test set correctly, which indicates that the model can accurately predict the emotion of new signals that it had not seen before.

However, a more complex model needed to be produced for commercial use. This is because from a commercial standpoint, the brain signals retrieved will not directly align with the patterns seen from this dataset, and these two test subjects. A more complex algorithm would learn from the data and would be better equipped to handle data that differs slightly from the data it was trained from. A recurrent neural network (RNN) was the next choice of model architecture.

RNN models are algorithms that build off previous knowledge as they dynamically learn from a dataset. All neurons and weights are connected to each other by the end of training as it processes a variable length sequence, making it an ideal model to use for sequence prediction. We decided to use a gated-recurrent-unit (GRU) model for training.

A GRU model is similar to a long-short-term-memory model (LSTM) in that it retains important information while not slowing down processing times, however GRU models require less parameters than LSTMs Vijaysinh 2021. GRU and LSTM models have a series of layers and activation functions that make the process run smoothly.

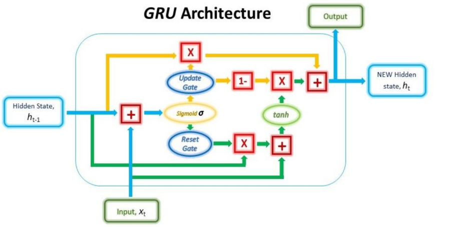


Figure 4. The general overview of the hidden layers and parameters that go into a GRU model Gabriel 2019.

Figure 4 shows a general layout of the structure of a GRU model. We implemented a GRU model over an LSTM model due to previous unsuccessful attempts with using an LSTM. Test accuracies for the LSTM model were close to 30 percent, and we decided to implement a GRU model structure to see if it would show better results.

From there, a simple GRU model was constructed: an input layer, followed by a GRU layer, then a closing Flatten and Dense layer.

Figure 5 depicts the flow of the model developed. Accuracy, Validation Accuracy, and Learning Rate were the metrics used to evaluate the model during training.Additionally, the Early Stopping method was called upon, the method stops training prematurely if the model has reached its maximum validation accuracy, meaning the validation accuracy has not increased past its maximum for 10 epochs. We used the cross entropy loss function to quantify the model's accuracy. For training, we implemented a variant of stochastic gradient descent (SGD) called Adam.

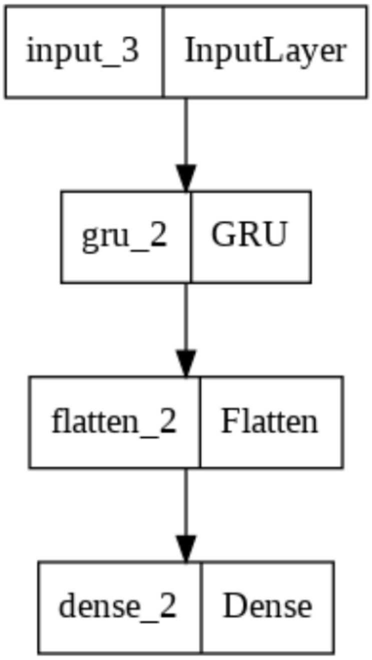
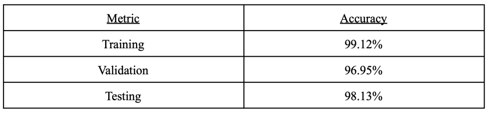


Figure 5. The structure of the GRU model used in this research project.



Metric	Accuracy
Training	99.12%
Validation	96.95%
Testing	98.13%

Figure 6. The training, validation, and testing accuracies after training the model.

Results and Discussion

These results show that the trained GRU model is able to accurately predict not only while training, but on new data the model has not seen in the testing and validation accuracies.

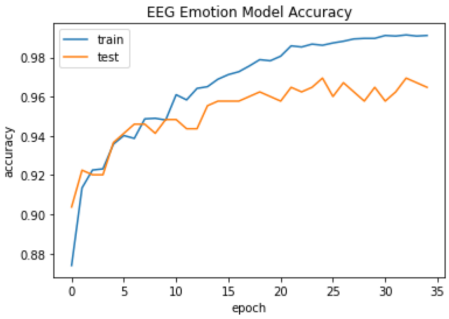


Figure 7. Progression of training and validation accuracy as the number of epochs increases. Accuracy increases as the number of epochs increases for both training and validation, reaches above 95 percent for both.

These plots and graphs show that the model has learned this dataset very well and is able to accurately predict the emotion of a particular brain wave. The GRU model was able to learn from the dataset and then predict on brain signals that it has not seen before. However, although the model is very accurate,

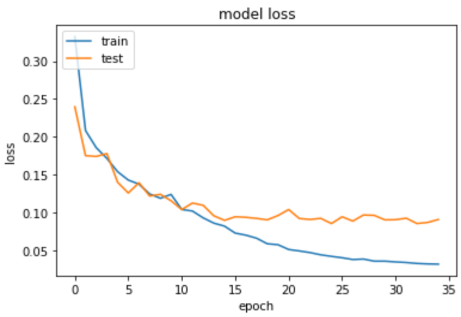


Figure 8. Progression of loss as epochs increases for both validation and training. Loss decreases as the number of epochs increases which is a good sign of accuracy.

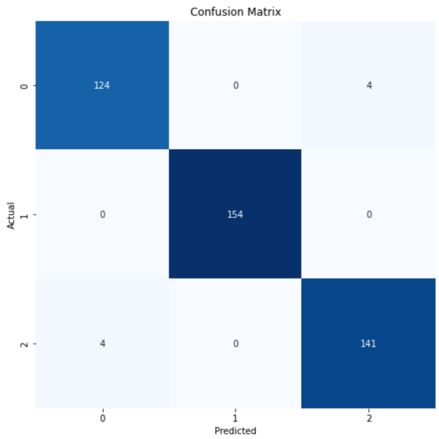


Figure 9. Confusion Matrix of the Test Predictions: 0 is Negative Emotion, 1 is Neutral Emotion, 2 is Positive Emotion. Nearly 100 percent accurate except for some incorrect predictions confusing Negative Emotion with Positive Emotion and vice versa.

it may be biased to the structure of this dataset. For clinical applications, it needs to be seen whether the patterns found in this dataset can be applied to the brain waves of all humans. Additionally, all the preprocessing steps applied to the data before training the model need to be applied as well in real time to ensure that the model can predict these brain waves accurately. However, it can be shown with evidence that if the necessary preprocessing steps are applied to an EEG brain wave, the emotion can be determined from it. A Muse Headset. a commercial headset used to measure and record EEG signals, can be used for a commercial implementation of this research project in the future. It is the headset used for collecting the data used in the dataset in this project, and is very useful for getting EEG signals.

Conclusion

We trained a GRU model from brain wave EEG data: the goal was to determine the emotion that the brain wave represented. A model with 96.95% validation accuracy was achieved. These brain signals needed to be converted to FFT signals, then transformed to time domain signals that were filtered with a butter worth filter. This system produced a model that could have significant real-world applications. The model provides a starting point for clinical diagnosis of psychiatric disorders. By having a system that can accurately detect emotions from EEG waves, a more quantifiable scale of internal diagnosis can be developed. The complex EEG frequencies can be handled with the help of ML. This model can be implemented in commercial EEG headsets and distributed to clinics. From there, they can use the newfound information regarding a patient's emotional state to better diagnose their condition, and help change the industry.

Acknowledgement

Firstly, this project has been implemented based on the EEG Brainwave Dataset: Feeling Emotions dataset, created by Jordan Bird and others (citation below). Additionally, this project would not be possible without the support and guidance of Inspirit AI's AI + X program and my mentor, Nima Leclerc. Their program enables other high schoolers similar to myself to engage and learn in the field of Artificial Intelligence, so check out their program for more opportunities.

References

Alhagry Salma, et al. 2017. Emotion recognition based on eeg using lstm recurrent neural network. 8.

Buckingham, J. J. Bird A. Ekart C. D., and D. R. Faria. 2019. Mental emotional sentiment classification with an eeg-based brain-machine interface.

Gabriel, Loye. 2019. Gated recurrent unit (gru) with pytorch.

J. J. Bird L. J. Manso, E. P. Ribiero A. Ekart, and D. R. Faria. 2018. A study on mental state classification using eeg-based brain-machine interface.

Li TM. Chao HC. Zhang, J. 2019. Emotion classification based on brain wave: a survey.

Nandi, J. Liu H. Meng A., and M. Li. 2016. Emotion detection from eeg recordings.

Novak, Sova. 2015. Diagnosis of mental illness today and tomorrow: a literary review of the current methods, drawbacks, and sociological components of mental health with regard to the diagnosis of mental illness.

Vijay, Choubey. 2016. Understanding recurrent neural network (rnn) and long short term memory(lstm).

Vijaysinh, Lendave. 2021. Lstm vs gru in recurrent neural network: a comparative study.