Harmful Brain Activity Classification Using Machine Learning

Kshitij
IIIT Delhi
kshitij22256@iiitd.ac.in

Mridul Goel
IIIT Delhi
mridul22303@iiitd.ac.in

Sarthak Singh
IIIT Delhi
sarthak22457@iiitd.ac.in

Roshan Kumar Mahto
IIIT Delhi
roshan22418@iiitd.ac.in

Abstract

Motivation: *The classification of harmful activities in the brain like seizures, GPD, GRDA, and other such abnormal patterns can be clinically diagnosed earlier through neurocritical care for immediate intervention. Electroencephalography (EEG) is helpful in the clinical diagnosis and evaluation of abnormalities in neurological activity. EEG signal processing has become very technical and, therefore, highly time-consuming, especially during manual procedures, which often makes the overall process arduous in any clinical application. These constraints are removed by the development of automated approaches based on sophisticated computational methods. This makes it possible to perform a precise and efficient analysis of EEG data. As a result, this project works to implement machine learning models to automatically classify EEG patterns, resulting in a considerable reduction in diagnosis time through improved clinical decisions.*

Approach: *This project aims to consider harmful EEG pattern classification and the detection of neurological diseases such as seizure, GPD, GRDA, and LRDA etc. using machine learning models such as MLP, SVM, and RF are used in this project. Several preprocessing techniques and feature extraction methods are used to increase the performance of the model. Different evaluation metrics such as Test accuracy and F1-score are applied to show the efficiency of the models.*

Git-Hub:*Harmful Brain Activity Classification Repository*

1. Introduction

Problem Statement: The manual interpretation of EEG signals is time-intensive and susceptible to human error which can potentially delay in critical diagnoses in cases such as seizure detection. This project aims to automate the classification of harmful brain activity patterns by leveraging machine learning models, including Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP), selected for their robustness and versatility.

Significance of Brain Pattern Classification: EEG-based classification can greatly improve the neurocritical care by identifying harmful patterns early. Automating this process not only speeds up diagnosis but also improves accuracy by detecting subtle abnormalities. For conditions like epilepsy, stroke, or brain trauma, timely detection of irregular brain activity can lead to more effective treatment and better outcomes for patients.

2. Literature Survey

Machine learning techniques have been widely used in EEG analysis for detecting conditions like epilepsy, Alzheimer's disease, and other neurological disorders. CNNs have been effective for feature extraction in EEG data, while Random Forests provide a balance of accuracy and simplicity for moderate datasets. Techniques such as oversampling and augmentation have been employed in previous studies to tackle the issue of class imbalance[1].

Support Vector Machines (SVM) are powerful classifiers that achieve nonlinearity by mapping input data to a high-dimensional feature space using kernel functions. Training involves solving a quadratic optimization problem to construct a hyperplane that maximizes the margin between data points, enabling high generalization. However, selecting the right kernel function remains a challenge. Originally a binary classifier, SVMs are extended to multiclass problems using methods like Error-Correcting Output Codes (ECOC). This approach trains multiple SVMs, each distinguishing specific class combinations, effectively enabling multiclass classification[2].

A significant challenge in EEG classification is the variability of signal patterns across patients. Studies have proposed various preprocessing techniques, including data slic-

ing and feature extraction, to enhance model performance. Our project builds upon these techniques by employing data selection and slicing to balance the dataset, ensuring better performance on minority classes like seizures[3].

3. Dataset and Preprocessing

Dataset Details: The dataset consists of six categories EEG recordings with expert-labels such as GPD, GRDA, Seizure, LPD, LRDA, and Others. The dataset is imbalanced and the LPD recording being under-represented among all brain patterns. Below is image sample of EEG

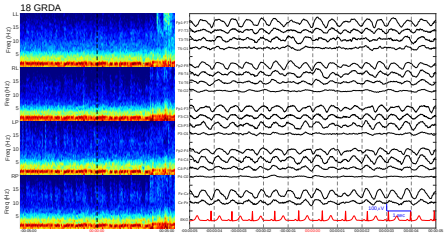

Figure 1. Sample Image

This figure represents the multi-channel EEG and ECG data recordings within a time frame of about 10 seconds, where the x-axis represents time and the y-axis represents amplitude in volts. The different electrode positions group the signals according to the international 10-20 system of electrode placement, for example, Fp1-F3, F3-C3, and Cz-Pz, which represent different brain regions.

Spectrum of frequencies: The power spectral density has been plotted in a graph. It is split at the middle: LL, RL, LP, RP (representations of the different sides/regions in the left brain, right brain, etc.). In the y-axis, representation in units of frequency measured in Hertz and in a range of values from 5-15Hz.

EEG Channels: On the bottom, several EEG channels are shown like Fz-Cz and P4-O2. These EEG channels are each recording the electrical activity coming out of different areas of the brain. The visual information appears as a time series. The oscillations of the amplitude represent neuronal activity over some seconds.

ECG Trace: An electrocardiogram (EKG/ECG) is also present, which records the electrical activity of the heart over time. This trace is included in the top section of the chart, represented by sharp peaks that correspond to cardiac cycles.

Key Observations: The data visualizations present EEG rhythms like alpha or beta oscillations, typically falling within the 5-15 Hz range, and correlate them with the ECG signal. The periodic patterns of the ECG can be compared with fluctuations in the EEG to explore potential cardio-neural interactions.

Data Preprocessing:

Data selection techniques are applied to handle the imbalanced dataset, which results in selecting a specific number of data points for each class. Label encoding was used to convert categorical labels into numerical form. Additionally, all EEG signals were normalized to ensure a uniform scale. Feature extraction was performed to enhance the discriminative power of the model. The following types of features were computed for each channel:

- **Time-Domain Features:** Mean, variance, standard deviation, skewness, kurtosis, entropy, and zero-crossing rate.
- **Frequency-Domain Features:** Power spectral density (PSD) and relative power across specific frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–45 Hz).
- **Wavelet Features:** Statistical measures of wavelet decomposition coefficients, including mean and variance at multiple decomposition levels.

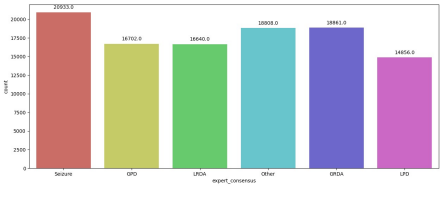

Figure 2. Dataset Distribution (Bar Chart)

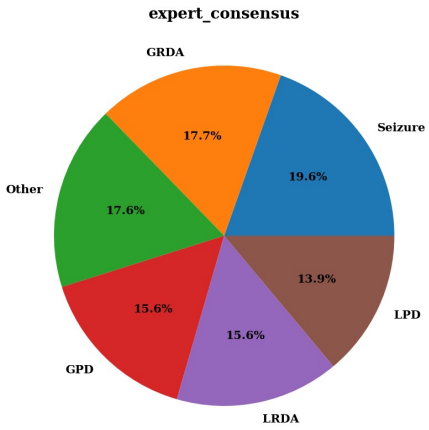

Figure 3. Dataset Distribution (Pie Chart)

4. Methodology and Model Details

Random Forest Model: Random Forest is an ensemble learning method that creates many decision trees during

training, and the final prediction is made based on the majority vote of all trees generated. The following hyperparameters were used in our model:

- Number of Trees: 500
- Maximum Depth: 10
- Minimum Samples per Leaf: 2
- Minimum Samples per Split: 13
- Max Features: 'sqrt' (square root of the total number of features)
- Random State: 42

This model performed very well in removing the noise of the features. However, it was noisy itself because it didn't behave appropriately to overfitting cases and faced discrimination of the minor classes with regard to the given dataset in question.

Support Vector Machine Model: The SVM model was configured using GridSearchCV to optimize hyperparameters, leading to the selection of the following parameters:

- Kernel: Radial Basis Function (RBF)
- Regularization Parameter (C): 100
- Gamma: 1
- Random State: 42

SVM is sensitive to class boundaries, which makes it efficient with overlapping classes and where the differences between similar classes are subtle. Such performance demonstrates the strength of the model in using high-dimensional feature spaces at a significant preprocessing and computationally resource-consuming cost.

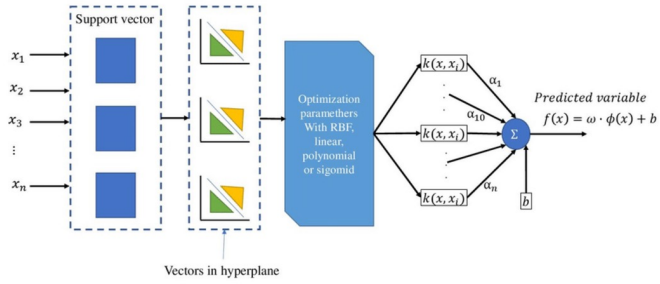

Figure 4. Illustration of SVM Model[4].

Multi-Layer Perceptron Model: The MLPClassifier was implemented with the following parameters:

- **Hidden Layer Sizes:** (128, 64, 32) — Three hidden layers with respective neurons.
- **Activation Function:** ReLU (Rectified Linear Unit) — Used to introduce non-linearity into the model.
- **Solver:** Adam — Optimization algorithm that combines the advantages of RMSProp and SGD with momentum.
- **Alpha:** 0.0001 — Regularization parameter to control overfitting (L2 regularization).
- **Batch Size:** 32 — Mini-batch size for training.
- **Learning Rate:** Adaptive — Adjusts the learning rate based on the network's performance during training.
- **Maximum Iterations:** 300 — The maximum number of iterations for optimization convergence.
- **Random State:** 42 — Ensures reproducibility by controlling the randomness in weight initialization and training data splits.

This configuration enabled the MLP model to learn effectively from the data, handling non-linear and complex patterns while balancing computational efficiency. However, its performance was highly sensitive to hyperparameter tuning, necessitating careful adjustments to achieve optimal results.

4.1. Training and Test

The dataset was split into training (80%) and test sets (20%). This division ensures that the model's accuracy and reliability are properly validated before deployment.

5. Results and Analysis

Evaluation Metrics: The performance of the models was evaluated using metrics such as accuracy, precision, recall, F1-score, and macro averages. Below is a detailed comparison of the three machine learning models used in this study: Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

Random Forest Results: The Random Forest model achieved a test accuracy of 83.42%, with a macro average F1-score of 0.83.

Support Vector Machine Results: The SVM model achieved the highest test accuracy of 90.5%, with a macro average F1-score of 0.90.

Multi-Layer Perceptron Results: The MLP model achieved a test accuracy of 88.83%, with a macro average F1-score of 0.89.

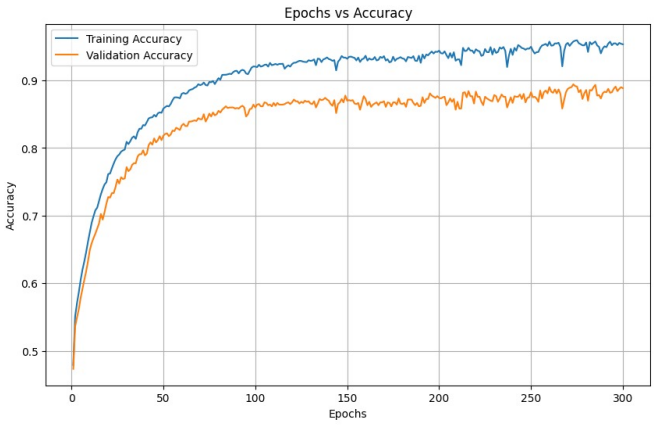

Figure 5. Illustration of training and validation accuracy over epochs For MLP.

Model	Test Accuracy (%)	Macro Avg F1-score
Random Forest	83.42	0.83
SVM	90.50	0.90
MLP	88.83	0.89

Table 1. Performance Comparison of Models

5.1. Critical Analysis of Results

Comparison of Models:

SVM performed the best among other models, achieving the highest accuracy and macro average. Its ability to determine precise decision boundaries was beneficial in distinguishing between closely related classes, such as GRDA and LRDA. MLP performed slightly less than SVM, but it showed better adaptability for more complex data patterns. Its performance marks the potential of neural networks in handling EEG data. Random Forest, while it is robust and interpretable, did not match the performance of SVM or MLP. It was found that tree-based methods were less effective for datasets with overlapping features or imbalanced distributions.

Implications:

The results showed that SVM was the most suitable model for EEG-based classification in this type of study, where datasets are of moderate size and have data imbalance. Its better accuracy makes it a more reliable model for real-world deployment in clinical settings. The MLP model, with its comparable performance, shows promise for future applications, particularly when it is combined with additional data augmentation techniques or larger datasets. Random Forest was slightly less accurate, it offers significant advantages in terms of interpretability and ease of implementation, making it a reliable option for scenarios where explainability is a priority.

Class Imbalance Challenge:

Despite the strong performance of all models, the imbalanced nature of the dataset impacted the classification of minority classes like "GPD" and "LPD". Future work will explore more advanced models like CNN.

6. Conclusion

This paper has analyzed the ability of machine learning algorithms to completely automate the classification of undesirable brain activity patterns in the EEG datasets, with an example on Random Forests, SVMs, and MLPs. Among them, SVM was the best candidate with a maximum accuracy of up to 90.5%, further showing its capability to deal even with complex and overlapping boundaries of classes. The MLP model was promising and was able to adapt the complex architecture to the complex patterns in EEG data. While the Random Forest model had slightly lower accuracy, it was robust and interpretable and could be a good choice in applications where interpretability is a requirement.

Results indicate the necessity of model selection and preprocessing in EEG-based classification. Significant improvements were observed in time, frequency, and wavelet domains with data normalization, label encoding, and feature extraction in performances. However, the dataset distribution was not uniform; it was difficult to classify some of the more minor categories, like "LPD" and "Others," in detail.

The results of this work are very important for the field of neurocritical care because timely and precise detection of harmful brain activity can be a difference-maker. Such automation of EEG pattern classification may expedite diagnosis in epilepsy, stroke, brain trauma, and other severe conditions. While SVM shows the highest accuracy, an extension of future research needs to explore the combination for more interpretability with an adaptation ability of MLPs and SVMs, especially on larger and newer datasets with advanced data augmentation strategies.

6.1. Challenges Encountered

Working with a 26.4 GB dataset posed several challenges due to its size and complexity:

1. **Memory Constraints and Batch Size:** The limited system memory necessitated reducing the batch size during training, which, in turn, increased the overall time complexity of the models. This trade-off made the training process slower and more computationally intensive.
2. **Noise in Specific Classes:** Certain classes, such as LRDA, were particularly difficult to classify due to their noisy and overlapping patterns, reducing the accuracy of feature extraction and classification.
3. **Feature Extraction Bottlenecks:** Feature extraction from such a large dataset was time-consuming, adding to the computational burden and limiting the ability to iterate quickly on model improvements.

In conclusion, this work contributes to the growing field of EEG-based machine learning applications, providing a foundation for further advancements in clinical diagnostics and real-time brain monitoring systems.

6.2. Individual Contributions

- **Kshitij:** Worked on EDA , dataset preprocessing and MLP model.
- **Mridul Goel:** Worked on the SVM model ,literature review and Critical Analysis of Results .
- **Sarthak Singh:** Worked on Result analysis, SVM model and references.
- **Roshan Kumar Mahto:** Focused on working on the Random Forest model , Methodology Model Details and Conclusion.

7. References

1. Kunekar, P., Gupta, M. K., Gaur, P. (2024). Detection of epileptic seizure in EEG signals using machine learning and deep learning techniques. *Journal of Engineering and Applied Science*, 71(1).
2. I. Guler and E. D. Ubeyli, "Multiclass Support Vector Machines for EEG-Signals Classification," IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 2, pp. 117-126, Mar. 2007
3. Chen, Y., Wang, H., Zhang, D., Zhang, L., Tao, L. (2023). Multi-feature fusion learning for Alzheimer's disease prediction using EEG signals in resting state. Frontiers in Neuroscience, 17, 1272834.
4. Hybrid Techniques to Predict Solar Radiation Using Support Vector Machine and Search Optimization Algorithms: A Review - Scientific Figure on ResearchGate.