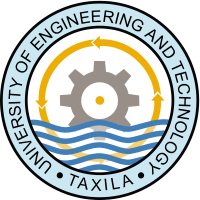
**University of Engineering & Technology, Taxila**



**Machine Learning**

**PROJECT REPORT – EEG Burnout Classification**

**Submitted to: Course instructor: Dr. Waqar Ahmad**

**Lab instructor: Engr. Mona Waseem**

**Submitted by: Ahamed Ayyash (22-CP-88)**

**Mohamed Mirzan (22-CP-86)**

**Tanveer Ahmed Khan (22-CP-76)**

**Semester:  6th**

**Dated: 13.05.2025**

**Department of Computer Engineering**

**EEG Burnout Classification**

**1. Project Overview**

The aim of this project is to develop a machine learning model that can classify the level of burnout (Low, Medium, High) in individuals using raw EEG signal data. The data is pre-processed and segmented, features are extracted, and multiple models are trained to evaluate the best-performing classifier.

**2. Dataset Description**

* **Source**: Raw EEG signal data of approximately 3GB collected from 30 patients.
* **Structure**:
  + Each patient has EEG data for both *eyes open* and *eyes closed* conditions.
  + Each condition is split into separate CSV files.
  + Each CSV records 3 minutes of EEG data, containing approximately 50,000 rows.
  + Up to 3 CSV files per condition per patient.
  + EEG signals are labelled with burnout levels derived from MBI (Maslach Burnout Inventory) scores.

**3. Data Preprocessing**

* **Cleaning**:
  + Removed null values and duplicate rows.
  + Checked for inconsistencies in signal timestamps and values.
* **Labelling**:
  + Each EEG recording is labelled with a burnout score derived from MBI.
* **Segmentation**:
  + A sliding window segmentation approach was used.
  + **Window Size**: 10 seconds.
  + **Reason**: If a fixed window size of 2560 samples was used, the last segment in a CSV might contain less than 2560 samples, leading to data loss. Using a time-based window (10 seconds) ensures consistent segment sizes (18 segments per 3-minute file).

**4. Feature Extraction**

From each 10-second EEG segment, features were extracted from four primary channels: RAW\_AF7, RAW\_AF8, RAW\_TP9, and RAW\_TP10. Features include:

* Mean
* Standard Deviation
* Skewness
* Kurtosis
* Zero Crossing Rate
* Line Length
* Hjorth Activity
* Hjorth Mobility
* Hjorth Complexity

Additional features:

* Encoded Gender (as a constant per file)

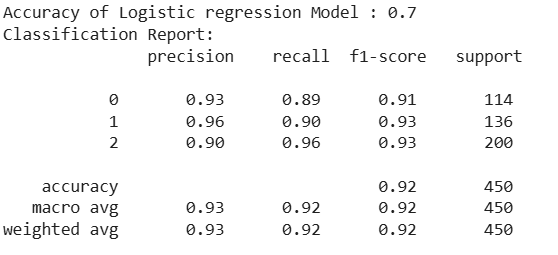
This resulted in 36 features per segment.

**5. Model Training and Evaluation**

Two primary models were trained and evaluated:

**5.1 Logistic Regression with L2 Regularization**

* **Accuracy**: 70%
* **Cross-Validation Accuracy** (5-fold): 69.43%
* **Precision (macro average)**: 93%
* **F1-Score (macro average)**: 92%
* **Recall (macro average)**: 92%



* Despite high precision, recall, and F1-score, the model's overall **accuracy (70%)** and **cross-validation accuracy (69.43%)** were significantly lower, indicating overfitting and poor generalization to unseen data.

**5.2 Random Forest Classifier**

* **Accuracy**: 92.4%
* **Cross-Validation Accuracy** (5-fold): 92.26%
* **Precision (macro average)**: 93%
* **F1-Score (macro average)**: 92%
* **Recall (macro average)**: 92%

A screenshot of a computer

AI-generated content may be incorrect.

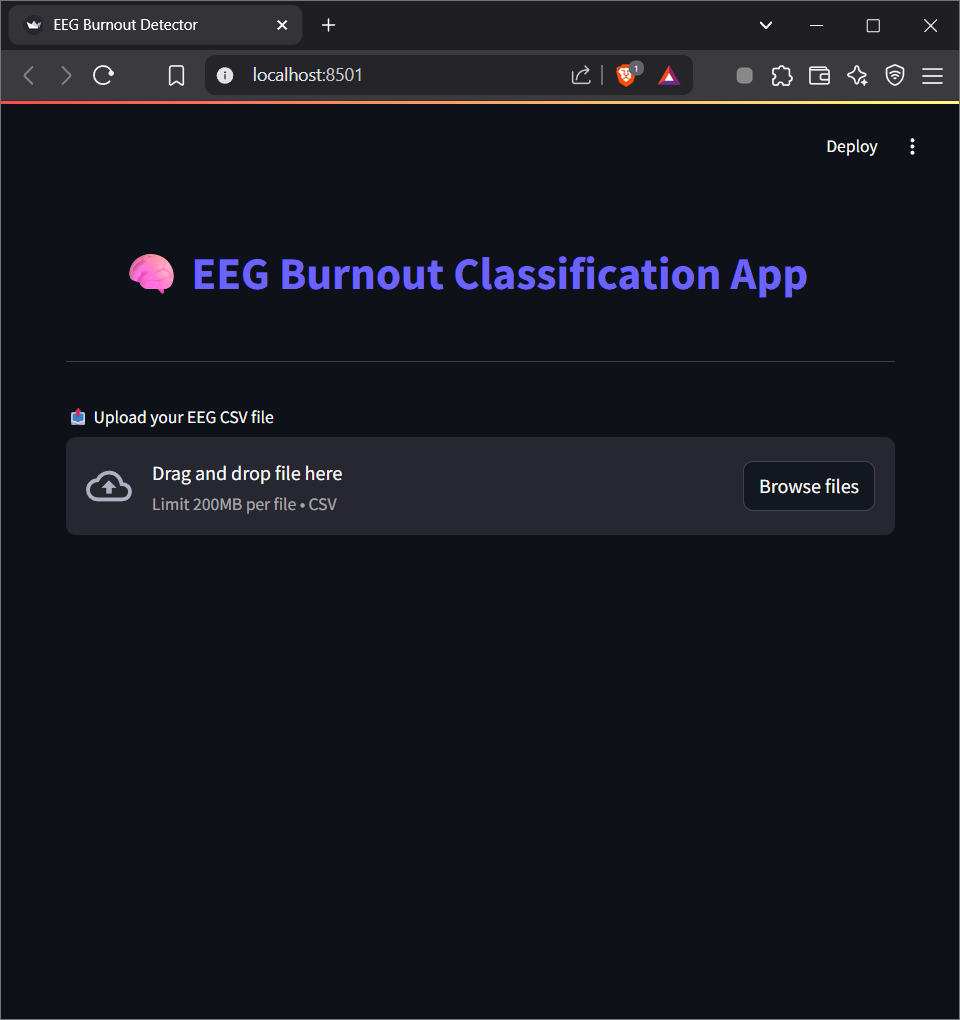
* The Random Forest model demonstrated **superior accuracy (92.4%)** and **cross-validation consistency (92.26%)**, along with strong macro-averaged precision, recall, and F1-scores, making it a more robust and generalizable choice compared to the L2-regularized Logistic Regression model.
* **Initial Model**: Logistic Regression with L2 Regularization
  + Performed well on a **small dataset (~1GB, 30 CSVs)**
  + Accuracy degraded on the **full dataset (~3GB, 180 CSVs)**
  + **Challenge**: Over-regularization or inability to generalize on larger dataset
* **Final Model**: Random Forest Classifier
  + Robust performance on larger dataset
  + Better at handling nonlinearities and variations in EEG signal patterns

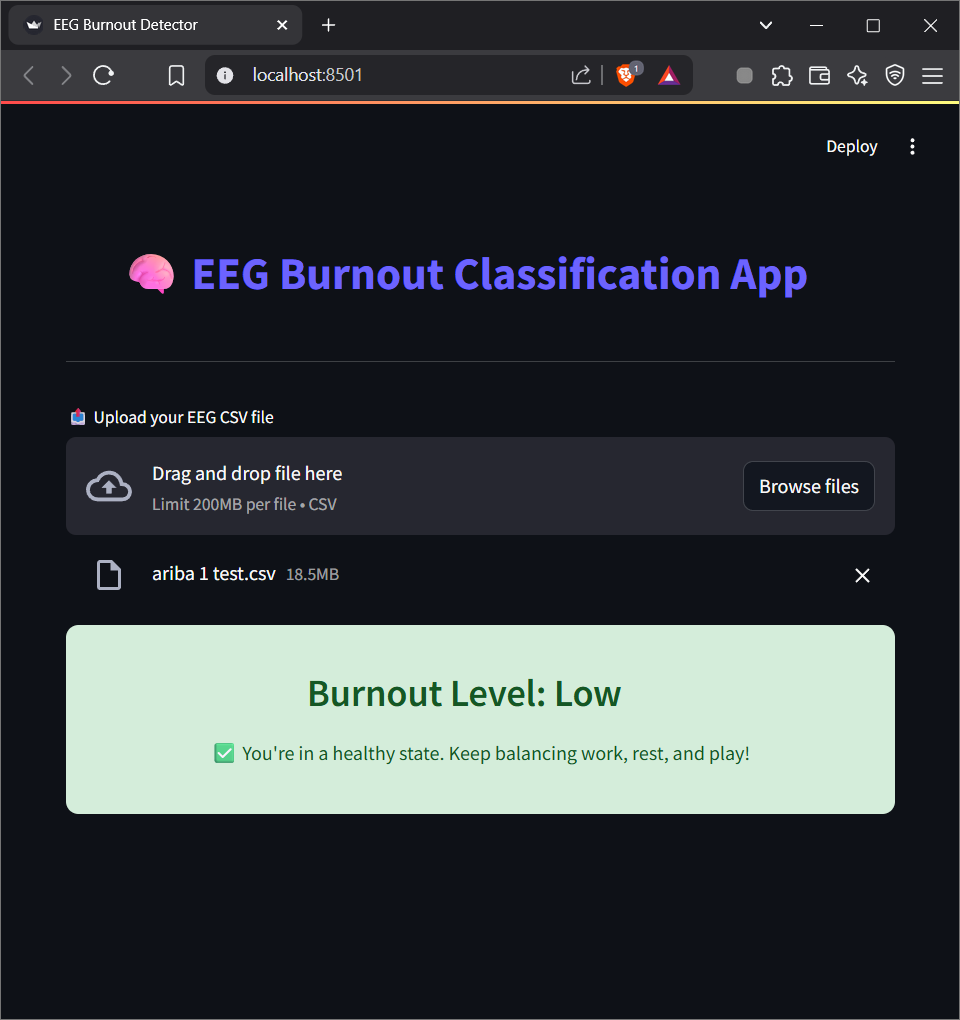
**6. Challenges Faced**

* **Data Loss in Fixed-Size Segmentation**:
  + Using a fixed window of 2560 samples would drop final segment data when <2560 samples remain.
  + Solved by using a 10-second time-based window.
* **Model Generalization**:
  + Logistic Regression (L2) worked for small datasets but failed on larger sets.
  + Random Forest provided better generalization and performance.
* **Data Volume**:
  + Handling 3GB of EEG data required efficient memory management and optimized processing loops.
* **Label Consistency**:
  + Ensured that each EEG file had a consistent burnout label based on the subject's MBI result.

**6. Deployment**

The EEG burnout classification model was deployed through a user-friendly GUI using Streamlit. Upon uploading a raw EEG CSV file through the interface, the system automatically processes the data, extracts relevant features, and predicts the burnout label using the trained machine learning model. Based on the predicted burnout level, the application also provides a brief advisory message tailored to the patient’s mental state. This deployment enables non-technical users, such as clinicians or researchers, to easily interpret EEG results and receive decision-support guidance in a streamlined, accessible manner.





**7. Conclusion**

This project demonstrates the effectiveness of feature-based machine learning in classifying EEG-based burnout levels. The structured approach of segmentation, statistical feature extraction, and model evaluation on real-world, large-scale data underscores the potential of EEG data in mental health diagnostics.

Deployment phase is planned as a next step, focusing on creating a user interface to process and classify uploaded EEG recordings.