System Description Paper: A Hybrid Feature Engineering Approach for EEG-BCI Signal Classification

Team: Elites
Date: June 30, 2025

1. Introduction and System Overview

This paper details the architecture and methodology of our system developed for the MTC-AIC3: Egypt National Artificial Intelligence Competition. The objective of the competition was to classify EEG signals from two distinct Brain-Computer Interface (BCI) paradigms: Motor Imagery (MI) and Steady-State Visual Evoked Potentials (SSVEP).

Our final approach is a **Hybrid Feature Engineering System** that leverages two specialized data processing pipelines tailored to the unique characteristics of the MI and SSVEP tasks. The core of our system is built on the discovery that the provided **motion sensor data contained a significantly stronger and more reliable predictive signal than the EEG data itself.** Our final model achieved a score of **0.71** on the public leaderboard.

The system architecture consists of two independent LightGBM classifiers:

1. An **MI Model** trained on a comprehensive set of time-series shape features.
2. An **SSVEP Model** trained on a comprehensive robust set of statistical and frequency-based features.

This task-specific approach proved superior to any single, universal modeling strategy.

2. Preprocessing and Data Handling

Our initial exploratory data analysis (EDA) revealed several critical properties of the dataset that guided our strategy.

2.1. Initial Data Consolidation

The dataset was provided in a deeply nested directory structure, which was inefficient for processing. Our first step was to write a consolidation script to parse the train.csv, validation.csv, and test.csv metadata files and merge the raw time-series data from the disparate EEGdata.csv files into six master CSVs (MI_train.csv, SSVEP_train.csv, etc.). This created a clean, tabular format that served as the foundation for all subsequent feature engineering.

2.2. Core Discovery: EEG vs. Motion Sensors

A crucial finding from our EDA was the nature of the sensor signals:

- **EEG Data:** The 8 EEG channels were found to be extremely noisy, exhibiting a massive DC offset that required centering for any analysis. Even after preprocessing, standard neural signals (such as Event-Related Desynchronization for MI or frequency peaks for SSVEP) were absent or unreliable.
- **Motion Sensor Data:** In stark contrast, the 6 motion sensor channels (AccX, AccY, AccZ, Gyro1, Gyro2, Gyro3) contained a strong, clean, and highly discriminative signal that correlated strongly with the target labels.

Conclusion: We concluded that the competition was primarily a **motion artifact classification problem**, not a pure BCI problem. Our strategy therefore shifted to focus on extracting the maximum amount of information from the motion sensor data.

3. Feature Engineering Methodology

Our investigation proved that a single feature engineering pipeline was suboptimal. We developed two specialized pipelines tailored to the unique demands of the MI and SSVEP tasks.

3.1. MI Task: Comprehensive Time-Series Feature Extraction

For the MI task, we discovered that the "shape" and dynamic character of the motion signals were most predictive.

- **Library:** We used the **tsfresh** library, a powerful tool for automated time-series feature extraction.
- **Parameters:** We employed the ComprehensiveFCParameters, which calculates an extensive set of over 30.000 selected top 1000 features for each time series. This includes statistical moments, frequency components (FFT), autocorrelation, entropy, and various measures of signal complexity.
- **Implementation:** We used a memory-safe, subject-by-subject processing script to handle the large dataset without exhausting system RAM. The features were extracted from the 6 motion sensor channels for each MI trial.

3.2. SSVEP Task: Robust Statistical and Frequency-Based Features

Our analysis showed that the tsfresh library was unstable on the SSVEP data, likely due to pathological data patterns causing some feature calculators to hang. Therefore, we developed a separate, highly robust feature set for this task.

- **Motion Sensor Features:** We also employed ComprehensiveFCParameters and calculated a set of 6 core statistical features (mean, std, min, max, median, skew) for each of the 6 motion sensor channels.

- **EEG Frequency Features:** Despite the noise, we found that the EEG data contained a weak but useful secondary signal. We calculated the average power in four standard frequency bands (**Delta** [1-4Hz], **Theta** [4-8Hz], **Alpha** [8-13Hz], **Beta** [13-30Hz]) for each of the 8 EEG channels.
- **Combined Feature Set:** The final feature vector for each SSVEP trial was a concatenation of the 36 motion sensor statistics and the 32 EEG band power features. This "Kitchen Sink" approach proved to be our most stable and reliable method for the SSVEP task.

4. Modeling and System Architecture

Our final system consists of two independent classifiers, reflecting the specialized feature sets.

- **Model Choice:** For both tasks, we used the **LightGBM (Light Gradient Boosting Machine)** classifier. We found it consistently outperformed other models, including RandomForest, MLP, and more complex deep learning architectures like 1D-CNNs, primarily due to its robustness in handling high-dimensional, noisy feature sets.
- **MI Model:** A LightGBM classifier trained on the tsfresh (ComprehensiveFCParameters) features extracted from the MI training data.
- **SSVEP Model:** A second LightGBM classifier trained on the "Kitchen Sink" (Motion Stats + EEG Band Powers) features extracted from the SSVEP training data.
- **Training Strategy:** We employed a "universal model" approach, training each classifier on the full set of available training data for its respective task. Our validation experiments showed that this was superior to any personalization (e.g., proxy-based) or stacking strategies, which were negatively impacted by the high noise level in the data of individual subjects.

5. Challenges and Solutions

1. **Challenge: Unusable EEG Data:** Our initial BCI-focused models failed due to the extremely noisy nature of the EEG signals.
    - **Solution:** We pivoted our entire strategy after a thorough EDA revealed the clean, predictive signal hidden in the motion sensor data.
2. **Challenge: tsfresh Memory Overload and Instability:** The full tsfresh feature extraction on the entire dataset caused memory crashes and process hangs, particularly on the SSVEP data.
    - **Solution:** We developed two solutions. First, a memory-safe, subject-by-subject processing script to manage RAM usage. Second, when tsfresh proved unstable even with this approach on the SSVEP data, we created the "Asymmetrical

Hybrid" strategy, using the robust "Kitchen Sink" features for the SSVEP task while retaining the more powerful tsfresh features for the stable MI task.

3. **Challenge: Unreliable Local Validation:** We discovered a significant distribution shift between our validation set and the hidden test set, making local validation an untrustworthy predictor of leaderboard performance.
    - **Solution:** We shifted our strategy to trust the public leaderboard as our primary validation tool, using it to confirm that the tsfresh-based approach was indeed superior, despite its poor local validation score.

6. Conclusion

Our final model, which achieved a score of **0.71**, is a testament to a rigorous, iterative, and evidence-based data science process. The key to success was not in applying a single, complex model, but in a deep investigation of the data's fundamental properties. By discovering the importance of the motion sensors, recognizing the different nature of the MI and SSVEP tasks, and engineering two separate, specialized feature pipelines, we were able to build a robust and highly competitive system.