Project Proposal: Scalable Analysis of EEG Data for Brain-Computer Interfaces

Course: DS/CMPSC 410

Semester: Fall 2025

Team Name: *[Your Team Name Here]* **Team Members: Hamid Shah, Kuria Mbatia, Nathan Mannings, Ryan Hussey, Xu Wang, Krish Chavan.**

---

1. Introduction and Motivation

Domain: Neuroscience / Brain-Computer Interfaces (BCI).

Problem: Electroencephalography (EEG) is a non-invasive technique that measures brain activity, forming the backbone of many clinical diagnostic tools and emerging BCI technologies. A significant challenge in this field is decoding a person's intended action or cognitive state from complex EEG signals. This project aims to build a scalable pipeline to classify different mental tasks (e.g., imagining moving the left vs. right hand) from a large, multi-subject EEG dataset.

Big Data Approach: A single EEG recording session generates millions of data points across dozens of channels. Scaling this to a study with hundreds of subjects results in datasets that are terabytes in size. Conventional single-machine analysis is too slow for feature extraction and model training on such data. A distributed framework like **PySpark is essential** for parallelizing the signal processing and machine learning workflows required to build a robust and generalizable BCI model.

---

2. Project Objectives

Our primary goals are to:

- Develop a robust PySpark workflow to ingest, clean, and preprocess a large-scale, multi-subject EEG dataset from its raw format into an analysis-ready state.
- Implement and parallelize the extraction of key time-series and spectral features (e.g., Power Spectral Density) from EEG signals for every trial and subject.
- Train and evaluate several machine learning models using PySpark MLlib to classify brain states with high accuracy.
- Analyze the scalability and performance of our pipeline by measuring job completion times with a varying number of cluster nodes.

3. Data Description

- **Data Source:** We will use the **"EEG Motor Movement/Imagery Dataset"** from PhysioNet. This public repository contains over 1,500 one- and two-minute EEG recordings from 109 volunteers.
- **Data Size:** The dataset contains over 300 million data points, totaling approximately 25 GB. It's an ideal size to demonstrate the need for a distributed system.
- **Data Characteristics:** The data is structured, multivariate time-series data from 64 EEG channels. It is temporal and high-dimensional.
- **Data Challenges:** The primary challenges include handling signal artifacts (e.g., from eye blinks), the low signal-to-noise ratio inherent in EEG, and ensuring consistent feature scaling across different subjects.

4. Technical Approach

- **Tools & Frameworks:** The core of our project will be **PySpark**, specifically using DataFrames for data manipulation and **MLlib** for classification models. We will integrate external Python libraries like **MNE-Python** and **SciPy** for specialized signal processing tasks, executed in parallel via Pandas UDFs.
- **Cluster Usage Plan:**
  1. We plan to use **4 CPU nodes** for our main processing and model training jobs.
  2. Jobs will be primarily **batch processing** for feature extraction and iterative model training.
  3. The raw data will be converted to the **Parquet** file format for efficient, distributed I/O and storage.
- **Pipeline Overview:**
  1. **Ingestion & Conversion:** A preliminary script will convert the raw `.edf` files into a partitioned Parquet dataset suitable for Spark.
  2. **Preprocessing & Feature Extraction:** In PySpark, we will filter the signals and then use a `groupBy("subject", "trial").applyInPandas()` workflow to calculate Power Spectral Density (PSD) features for each trial in parallel.
  3. **Model Building:** We will feed the extracted features into PySpark MLlib classifiers like **Logistic Regression** and **Random Forest** to distinguish between different imagined movements.
  4. **Evaluation & Visualization:** Results will be aggregated to evaluate model accuracy, and we will create visualizations showing feature importance and performance scaling curves.

5. Expected Outcomes

- **Deliverables:** A clean, well-documented PySpark codebase, trained classification models, a final report detailing our methodology and findings, and a presentation summarizing the project.
- **Success Metrics:**
  - **Model Performance:** Achieving a classification accuracy significantly above chance level (e.g., >70% for a binary task).
  - **Scalability:** Demonstrating a near-linear decrease in processing time as we increase the number of cluster nodes from 1 to 4.
  - **Scientific Insight:** Visualizations identifying which EEG channels and frequency bands are most predictive of motor imagery.

---

6. Work Plan & Timeline

- **Week 1-2:** Data Exploration & Ingestion. Familiarize ourselves with the dataset. Develop and run the script to convert data to Parquet.
- **Week 3-5:** PySpark Preprocessing Pipeline. Implement parallel filtering and artifact handling.
- **Week 6-8:** Feature Extraction. Develop and scale the PSD feature extraction workflow using Pandas UDFs.
- **Week 9-11:** Model Training & Tuning. Implement, train, and tune MLlib classification models.
- **Week 12-14:** Evaluation & Visualization. Run scalability tests, aggregate results, and prepare visualizations.
- **Week 15:** Final Report & Presentation. Finalize all deliverables.

---

7. Division of Labor

To ensure parallel progress and clear ownership, we will assign primary and secondary roles:

- **Team Lead & Project Management (1 person):** Oversees the timeline, integrates components, and manages documentation.
- **Data Engineering Leads (2 people):** Responsible for the initial data conversion (EDF to Parquet) and building the core data ingestion and preprocessing pipeline in PySpark.
- **Machine Learning Leads (2 people):** Responsible for the feature extraction UDFs, implementing the MLlib models, and running hyperparameter tuning.
- **Analysis & Visualization Lead (1 person):** Responsible for evaluating model performance, running scalability benchmarks, and creating all final plots and visualizations for the report and presentation.

All team members will contribute to the final report and presentation.

---

8. Potential Challenges & Mitigation

- **Challenge:** The initial conversion of the non-native `.edf` file format could be a bottleneck.
  - **Mitigation:** We will dedicate two team members (Data Engineering Leads) to tackle this task early in the project.
- **Challenge:** Managing Python library dependencies (like MNE) across the Spark cluster can be complex.
  - **Mitigation:** We will use a virtual environment and document a clear setup process that can be replicated on each node.
- **Challenge:** A single trial's data might be too large for a worker's memory in a Pandas UDF.
  - **Mitigation:** We will monitor Spark UI for memory issues and can resample the data or increase worker memory if necessary.

---

9. References

1. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.
2. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R. (2004). BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. IEEE Transactions on Biomedical Engineering 51(6):1034-1043.
3. MNE-Python: A robust library for EEG/MEG data analysis. (https://mne.tools/