Supervised Learning Model for Motor Movement Prediction using Brain-Computer Interfaces Phase 2 – Project Progress Report

Angel Barrera
Master's Student in Applied Data Science
East Tennessee State University

March 4, 2025

1 Progress Summary

So far, I have made solid progress in building a machine learning model that classifies motor movements (left fist, right fist, both feet) using EEG data. The goal is to develop a system that can interpret brain signals and predict movement intentions, which could be useful in assistive technologies, brain-computer interfaces (BCI).

I have successfully trained a **Random Forest Classifier** and evaluated its performance. The model achieved an **accuracy of 100% on the training set** and **95.38% on the test set**, suggesting strong generalization while maintaining high performance. The classification report and confusion matrix confirm that the model is capable of accurately distinguishing between different motor movements.

However, I remain somewhat skeptical about the perfect training accuracy, as it might indicate overfitting. A crucial next step is to validate whether this performance holds when introducing new data from different subjects, especially since the dataset includes a large number of participants. Further testing with unseen subjects will help determine if the model truly generalizes well or if adjustments are necessary to improve its robustness.

2 Tasks Completed

2.1 Data Collection

The dataset used in this project comes from **PhysioNet** [1] and was recorded using the **BCI2000** system [3], a well-established brain-computer interface platform. It consists of EEG recordings from **109 subjects**, each performing various motor and motor imagery tasks.

Each recording contains **64-channel EEG signals sampled at 160 Hz**, following the **international 10-10 electrode placement system** [2]. The dataset includes 14 experimental runs per subject, covering different movement conditions:

- **Task 1**: Opening and closing the left or right fist.
- **Task 2**: Imagining opening and closing the left or right fist.
- **Task 3**: Moving both fists or both feet.
- **Task 4**: Imagining moving both fists or both feet.

The EEG recordings are provided in **EDF+ format** with annotation labels (**T0, T1, T2**), which were mapped to movement classes for classification [7]. This dataset is widely used in BCI research, making it a solid foundation for developing a movement prediction model.

2.2 Data Cleaning

To ensure data quality, I handled missing values using **mean imputation**. Applied a **bandpass filter (1-40 Hz)** to remove noise such as muscle activity artifacts and electrical interference. Merged multiple EEG sessions into a unified dataset that includes all movement types.

2.3 Data Encoding

Mapped EEG event annotations **(T0, T1, T2)** to numerical movement labels for classification. Ensured correct labeling based on the dataset documentation.

2.4 Feature Selection and Feature Extraction

To effectively classify motor movements using EEG signals, I selected features that capture meaningful information from the time, frequency, and spatial domains [Figure 4]. These features help reduce the dimensionality of raw EEG data while preserving relevant neural activity patterns.

2.4.1 Time-Domain Features (Statistical Analysis)

Time-domain features characterize how EEG signals fluctuate over time, revealing patterns linked to brain activity changes during movement. The following statistical measures were extracted:

- **Mean** (μ): Represents the average amplitude of the EEG signal, providing a baseline indicator of brain activity.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

- **Variance** (σ^2): Measures signal fluctuation, useful in detecting movement-induced changes.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2}$$

- **Root Mean Square (RMS)**: Quantifies overall signal power.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2} \tag{3}$$

- **Hjorth Parameters**:
  - **Activity**: Measures signal power (variance) [4, 2].

$$\text{Activity} = \frac{1}{N}\sum_{i=1}^{N}(x_i - AM_x)^2 \tag{4}$$

  - **Mobility**: Indicates frequency characteristics by analyzing how rapidly the signal changes [4, 2].

$$\text{Mobility} = \sqrt{\frac{Var(d)}{Var(x)}} \tag{5}$$

  - **Complexity**: Detects unpredictability in the signal, distinguishing real movements from imagined ones [4, 2].

$$\text{Complexity} = \frac{\text{Mobility of first derivative}}{\text{Mobility}} \tag{6}$$

Recent studies [5, 2] have shown that these features play a crucial role in capturing key characteristics of EEG signals and improving classification performance.

After computing these metrics, each EEG channel contributes six time-domain features, leading to a 64×6 feature matrix.

2.4.2 Frequency-Domain Features (Power Spectral Analysis)

EEG activity is inherently frequency-dependent, with motor movements associated with specific frequency bands, such as Mu (8-12 Hz) and Beta (13-30 Hz). To extract frequency-related information, I computed:

- **Power Spectral Density (PSD)**: PSD quantifies how power is distributed across various frequency components in EEG signals, providing critical insights into motor-related neural activity. Instead of relying solely

on raw signal amplitudes, PSD allows the identification of dominant frequency bands, such as Mu (8-12 Hz) and Beta (13-30 Hz), which are essential for movement classification [10].

To compute PSD, the multitaper spectral estimation method is employed. This approach enhances frequency resolution while minimizing variance by averaging multiple orthogonal windowed periodograms. The equation for PSD is:

$$P_{xx}(f) = \frac{1}{N} \sum_{n=0}^{N-1} |X_n(f)|^2 \tag{7}$$

where $P_{xx}(f)$ represents the power spectral density at frequency f, $X_n(f)$ is the Fourier-transformed EEG signal, and N is the total number of samples.

PSD-based features have been successfully integrated into EEG classification models, particularly in brain-computer interface (BCI) applications. They provide a robust way to differentiate between movement states and resting-state brain activity. Studies have demonstrated the effectiveness of PSD features in improving classification accuracy when combined with machine learning classifiers such as Gaussian Process Classifiers (GPC) and Support Vector Machines (SVM) [10].

This extraction step results in two additional frequency-domain features per channel, forming a 64×2 matrix.

- **Fourier Transform (FFT Mean)**: Converts EEG signals from the time domain into the frequency domain, revealing dominant oscillatory patterns. This transformation allows for the extraction of relevant frequency components that are critical in distinguishing movement-related brain activity [2].

$$X(f) = \sum_{n=0}^{N-1} x_n e^{-j2\pi fn/N} \tag{8}$$

The original Fourier Transform is expressed in continuous form as:

$$s(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \tag{9}$$

where $s(f)$ represents the frequency-domain signal, and $s(t)$ is the time-domain signal.

However, EEG signals are **discrete** in nature, meaning they are sampled at specific time intervals. To process EEG signals computationally, we use the **Discrete Fourier Transform (DFT)**, which is represented as:

$$X(f) = \sum_{n=0}^{N-1} x_n e^{-j2\pi fn/N} \tag{10}$$

The **Fast Fourier Transform (FFT)** is an efficient algorithm used to compute the DFT. While a direct computation of the DFT takes $O(N^2)$ operations, the FFT reduces this complexity to $O(N \log N)$, making it significantly faster for large datasets [9].

This extraction step results in two additional frequency-domain features per channel, forming a 64×2 matrix.

2.4.3 Spatial Features (Common Spatial Patterns - CSP)

To extract movement-related information from EEG signals, spatial filtering techniques such as **Common Spatial Patterns (CSP)** are employed. CSP is a supervised feature extraction method designed to maximize variance differences between two classes of EEG signals, thereby enhancing discriminability in motor movement classification [11].

The core objective of CSP is to identify an optimal set of spatial filters W that project EEG signals into a subspace where the variance is maximized for one class while minimized for the other. This is achieved through the following steps:

- **Covariance Computation**: Calculate the sample covariance matrices for EEG trials corresponding to each movement class [11].
- **Generalized Eigenvalue Decomposition**: Solve the following eigenvalue problem:
$$C_1 W = \lambda C_2 W \tag{11}$$
where C_1 and C_2 are the estimated covariance matrices of two different movement classes, W represents the spatial filter, and λ denotes the eigenvalues that determine the discriminability of the transformation [11].
- **Projection and Feature Extraction**: Transform the EEG signals using the learned spatial filters:
$$Z = W^\top X \tag{12}$$
where X is the original EEG signal matrix, and Z is the transformed spatially filtered signal.

While CSP is highly effective in enhancing movement-relevant EEG signals, it is prone to overfitting, particularly when the number of training trials is limited [12]. To address this, **Probabilistic CSP (P-CSP)** introduces a Bayesian framework that incorporates prior knowledge to regularize spatial filters, mitigating overfitting and improving generalization [11].

The P-CSP model extends the standard CSP by modeling EEG signals as:

$$X_k = AZ_k + E_k \tag{13}$$

where:

- X_k is the observed EEG signal for condition k.

- A represents the spatial patterns (mixing matrix).
- Z_k is the underlying component signal for condition k.
- E_k is the additive Gaussian noise term [11].

To find optimal spatial filters under the P-CSP framework, the following variational Bayesian inference approach is used:

$$p(A, Z_k|X_k) \propto p(X_k|A, Z_k)p(A)p(Z_k) \tag{14}$$

where prior distributions are imposed on A and Z_k to enforce sparsity and prevent overfitting [12].

Regularization Strategies: To further improve CSP performance, various regularized versions have been developed:

- Tikhonov Regularization (TR-CSP): Introduces an ℓ_2-norm penalty term to enforce smoothness in the spatial filters:

$$J(W) = \frac{W^\top C_1 W}{W^\top (C_1 + \rho H) W} \tag{15}$$

  where ρ is the regularization parameter, and H is a constraint matrix [11].

- Sparse CSP (S-CSP): Utilizes ℓ_1-norm constraints to enforce sparsity in spatial filters, effectively reducing the number of channels used for classification [11].

2.4.4 Final Feature Matrix

After feature selection and extraction, the dataset consists of:

- **Time-domain features**: 64×6
- **Frequency-domain features**: 64×2
- **Spatial features**: 4 CSP components

These features were then **standardized using Z-score normalization** to ensure uniform scaling before model training.

This feature set provides a robust representation of movement-related EEG activity, ensuring an optimal balance between computational efficiency and classification accuracy.

2.5 Model Building

Chose **Random Forest** as the classifier because it's robust to noise and provides insights into feature importance. Since the dataset was imbalanced, I applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the classes. Split the dataset into **80% training and 20% testing** for model evaluation.

2.6 Model Evaluation

The model's performance was assessed by evaluating both training and testing accuracy to ensure its generalization capability [Figure 7, Figure 8]. The results showed:

- **Training Accuracy**: **100%**, indicating that the model fits the training data perfectly.
- **Testing Accuracy**: **95.38%**, demonstrating strong generalization but warranting further validation on new subjects.

To further analyze the model's effectiveness, the following evaluation metrics were used:

- **Precision, Recall, and F1-score**: These metrics confirmed a well-balanced classification performance, with high recall ensuring minimal false negatives.
- **Confusion Matrix**: Provided insights into misclassification patterns, showing that most predictions were correct, though some movement classes had minor confusion.
- **Feature Importance Analysis**: Common Spatial Patterns (CSP) features played the most significant role, followed by Hjorth Mobility, confirming the importance of spatial and time-domain characteristics in EEG-based movement prediction [Figure 4].

Although the training accuracy suggests the model learned the dataset well [Figure 9], the perfect score raises skepticism about potential overfitting. Testing with unseen EEG recordings from different subjects will be crucial in verifying the model's real-world applicability.

3 Challenges & Solutions

3.1 Dealing with Class Imbalance

Challenge: The dataset had more samples for some movement types than others, leading to biased predictions [Figure 5]. **Solution**: Used **SMOTE** to generate synthetic samples for underrepresented classes, ensuring a more balanced dataset [Figure 6].

3.2 High-Dimensional EEG Data

Challenge: The original dataset had **64 channels per sample**, making it computationally heavy. **Solution**: Applied **feature engineering** (CSP, PSD, Hjorth Parameters) to extract meaningful information and reduce dimensionality.

3.3 Noise & Artifacts in EEG Signals

Challenge: EEG data often contains unwanted noise from **muscle activity, power line interference, and environmental factors** [Figure 2]. **Solution**: Applied **a bandpass filter (1-40 Hz)** to remove unwanted frequencies while preserving motor-related EEG signals.

3.4 Model Generalization

Challenge: Ensuring the model generalizes well to unseen EEG data. **Solution**: Used **cross-validation** and **feature importance analysis** to confirm that extracted features genuinely contribute to movement classification.

4 Project Timeline

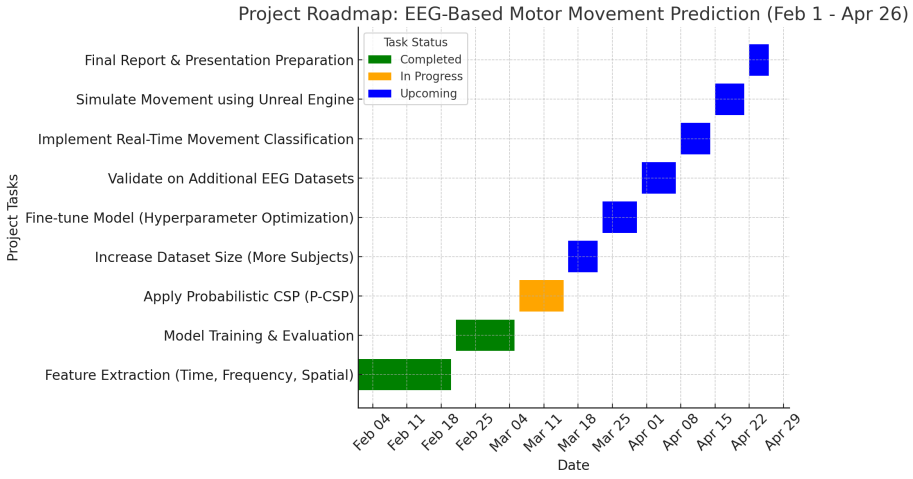

Figure 1: Timeline Graph

Task	Start Date	End Date	Status
Feature Extraction (Time, Frequency, Spatial)	Feb 1	Feb 20	Completed
Model Training & Evaluation	Feb 21	Mar 5	Completed
Apply Probabilistic CSP (P-CSP)	Mar 6	Mar 15	In Progress
Increase Dataset Size (More Subjects)	Mar 16	Mar 22	Upcoming
Fine-tune Model (Hyperparameter Optimization)	Mar 23	Mar 30	Upcoming
Validate on Additional EEG Datasets	Mar 31	Apr 7	Upcoming
Implement Real-Time Movement Classification	Apr 8	Apr 14	Upcoming
Simulate Movement using Unreal Engine	Apr 15	Apr 21	Upcoming
Final Report & Presentation Preparation	Apr 22	Apr 26	Upcoming

Table 1: Project Timeline Overview

The final presentation will be scheduled for **April 26th, 2025**. The final deliverable will include a comprehensive report and a presentation summarizing the findings and outcomes of the study.

References

[1] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000.

[2] W. H. Alawee, B. A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis," *Journal of Electrical Bioimpedance*, vol. 14, no. 1, pp. 66-72, Dec. 2023. doi: 10.2478/joeb-2023-0009. PMID: 38162817; PMCID: PMC10750318.

[3] G. Schalk et al., "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034-1043, 2004.

[4] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306-310, 1970.

[5] C. W. Stevenson et al., "EEG-based classification of motor imagery tasks," *Journal of Neuroscience Methods*, vol. 198, no. 1, pp. 152-158, 2011.

[6] M. Hassan et al., "Brain network analysis: From functional connectivity to small-world organization," *Neuroscience*, vol. 316, pp. 137-155, 2016.

[7] H. Ramoser et al., "Optimal spatial filtering of single-trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441-446, 2000.

[8] D. C. R. Novitasari, S. Suwanto, M. H. Bisri, and A. H. Asyhar, "Classification of EEG Signals using Fast Fourier Transform (FFT) and Adaptive Neuro-Fuzzy Inference System (ANFIS)," *Jurnal Matematika MANTIK*, vol. 5, no. 1, pp. 36-45, May 2019.doi: 10.15642/mantik.2019.5.1.35-44.

[9] D. C. R. Novitasari, S. Suwanto, M. H. Bisri, and A. H. Asyhar, "Classification of EEG signals using Fast Fourier Transform (FFT) and Adaptive Neuro-Fuzzy Inference System (ANFIS)," *Mantik: Jurnal Matematika*, vol. 5, no. 1, pp. 35-44, May 2019. doi: 10.15642/mantik.2019.5.1.35-44.

[10] S. M. Redwan, M. P. Uddin, A. Ulhaq, et al., "Power spectral density-based resting-state EEG classification of first-episode psychosis," *Scientific Reports*, vol. 14, no. 15154, 2024. doi: 10.1038/s41598-024-66110-0. [Online]. Available: https://doi.org/10.1038/s41598-024-66110-0

[11] W. Wu, Z. Chen, X. Gao, Y. Li, E. N. Brown, and S. Gao, "Probabilistic Common Spatial Patterns for Multichannel EEG Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 639–653, Mar. 2015, doi: 10.1109/TPAMI.2014.2330598.

[12] "Probabilistic Common Spatial Patterns for Multichannel EEG Analysis," Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4441303/, Accessed: Mar. 4, 2025.

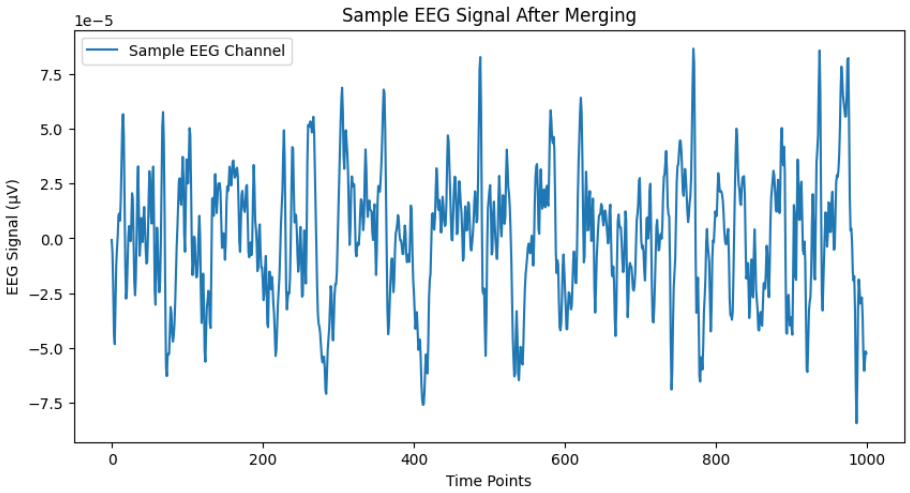

Figure 2: EGG Signals

	Mean	Variance	RMS	Hjorth_Activity	Hjorth_Mobility	\
0	1.507611e-09	1.419974e-09	0.000038	1.419974e-09	0.387710	
1	6.057725e-11	1.547173e-09	0.000039	1.547173e-09	0.376559	
2	-1.791042e-09	1.611370e-09	0.000040	1.611370e-09	0.366584	
3	-2.576242e-09	1.623943e-09	0.000040	1.623943e-09	0.365716	
4	-8.131073e-10	1.514953e-09	0.000039	1.514953e-09	0.361650	

	Hjorth_Complexity	FFT_Mean	PSD_Mean	CSP_1	CSP_2	CSP_3	\
0	0.921347	0.004185	1.783021e-11	-0.250792	0.496230	0.688502	
1	0.881351	0.004319	1.941544e-11	1.318472	-1.036527	-0.788893	
2	0.871932	0.004354	2.020097e-11	-0.445343	-0.762802	-0.522497	
3	0.877637	0.004366	2.027850e-11	-1.753108	-1.842257	-1.307880	
4	0.877082	0.004186	1.891996e-11	-2.128529	-1.927870	-1.570716	

	CSP_4	Movement_Label
0	-2.492621	1
1	-1.380543	1
2	-1.574479	1
3	-0.270498	3
4	0.311338	3

Figure 3: Engineered Features

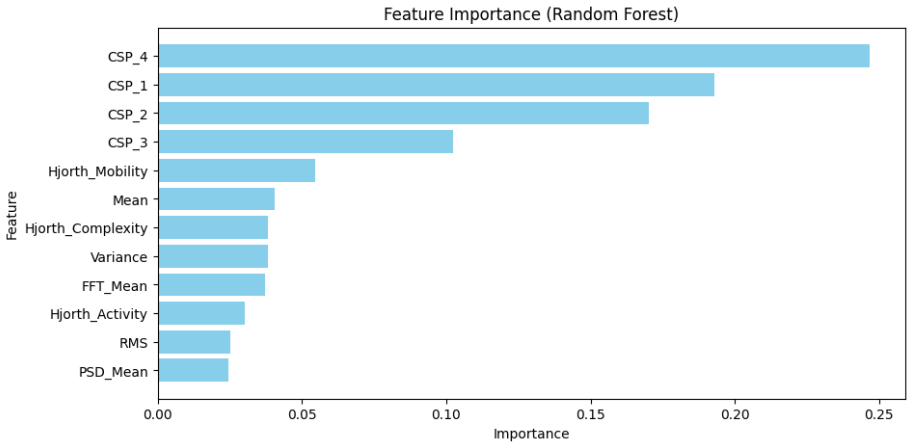

Figure 4: Feature Importance

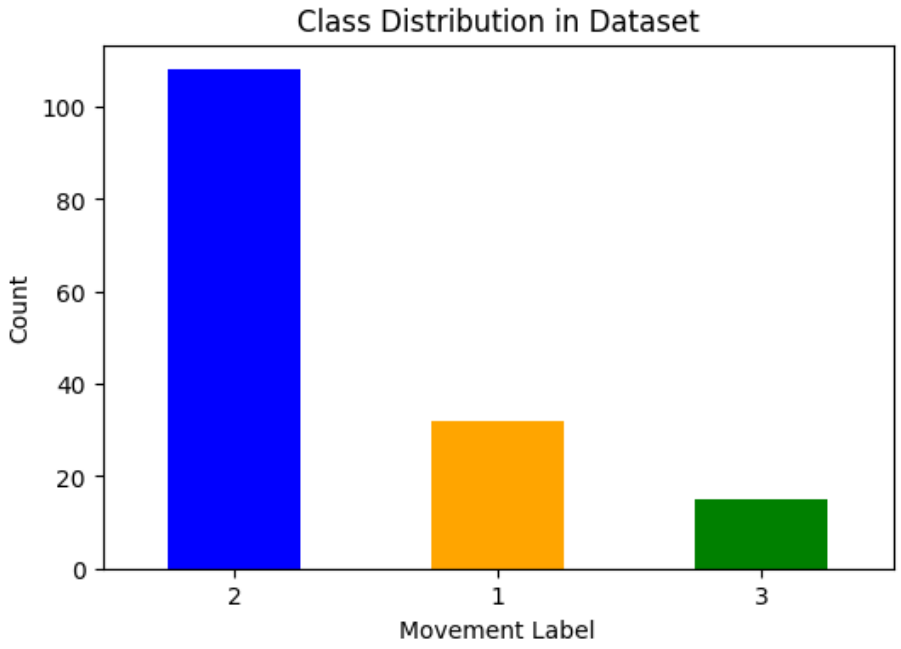

Figure 5: Class Distribution Before SMOTE

Figure 6: Class Distribution After SMOTE

```
Training Accuracy: 1.0000
Testing Accuracy: 0.9538

Classification Report (Test Set):
              precision    recall  f1-score   support

         1.0       1.00      1.00      1.00        87
         2.0       1.00      1.00      1.00        86
         3.0       1.00      1.00      1.00        86

    accuracy                           1.00       259
   macro avg       1.00      1.00      1.00       259
weighted avg       1.00      1.00      1.00       259


Classification Report (Test Set):
              precision    recall  f1-score   support

         1.0       0.88      1.00      0.93        21
         2.0       1.00      0.86      0.93        22
         3.0       1.00      1.00      1.00        22

    accuracy                           0.95        65
   macro avg       0.96      0.95      0.95        65
weighted avg       0.96      0.95      0.95        65
```

Figure 7: Random Forest Model Evaluation

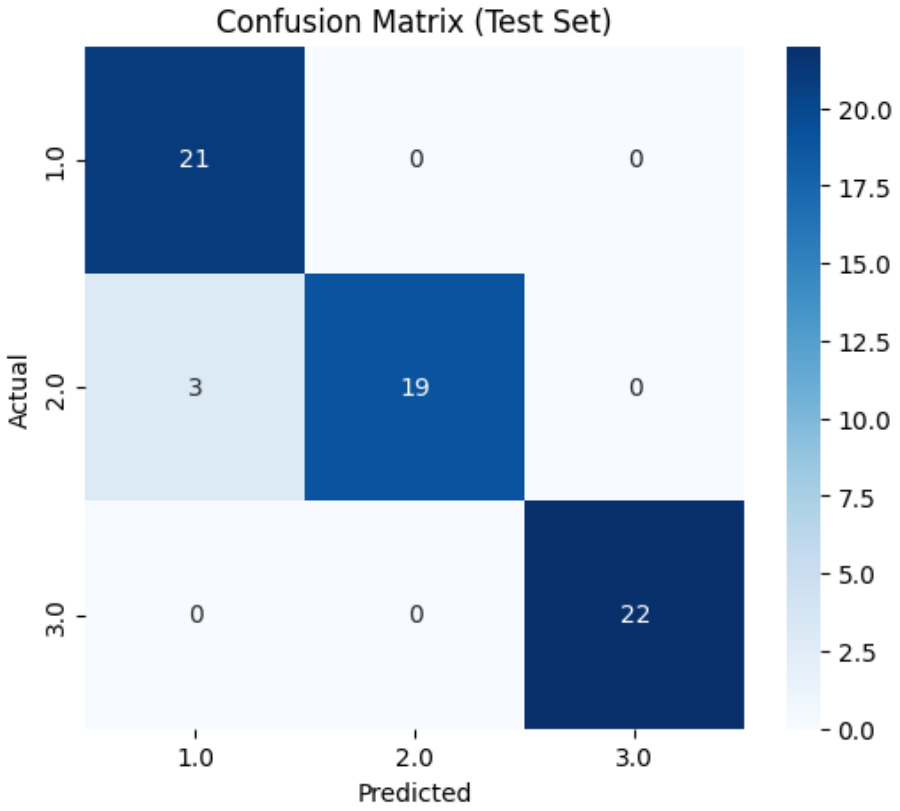

Figure 8: Random Forest Confusion Matrix

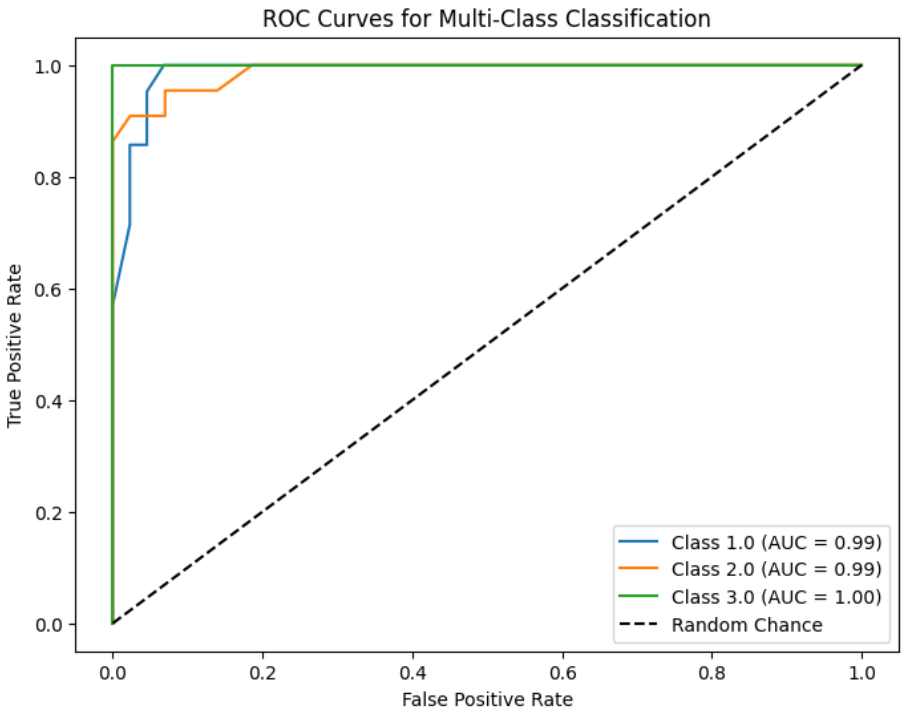

Figure 9: ROC Multi-Class Classification